Universität Bremen
Fachbereich 11
Human- und Gesundheitswissenschaften

Berücksichtigung von Heterogenität in Meta-Analyse von Randomisierten Kontrollierten Studien

Dissertation zur Erlangung des akademischen Grads
Doktor Public Health

vorgelegt von
Abdel Moniem Mukhtar
Bremen, den 25. März 2008

Erstgutachter: Prof. Dr. Norbert Schmacke

Zweitgutachter: Prof. Dr. Dr. hc. Jürgen Timm

Colloquium am 16.7.2008

Abstract

Heterogeneity in meta-analysis can be caused by chance, methodological or clinical variations between the included primary studies. To identify a clinical variable as a cause of heterogeneity in meta-analysis, one should firstly investigate chance and methodological variations. Meta-analysis with individual patient data (MA-IPD) has a greater potential than that with aggregate patient data (MA-APD) to detect which subgroups of patients get less, more or no benefit from an intervention. In this thesis two systematic reviews and a MA-APD were undertaken. The first review searched systematically for empirical studies on the impact of bias in randomised controlled trials (RCTs) on the results of meta-analysis. 39 studies were identified and 134 empirical comparisons between trials of high and low methodological quality could be extracted and summarized using a random effects model. RCTs with low quality scores, with inadequate method of randomisation, with inadequate concealment of allocation and those without any type of blinding, on average, overestimated the treatment efficacy. However, most of the empirical studies did not consider clinical causes of heterogeneity. This fact may confound the relation between methodological quality and effect size. Simultaneous investigation of diverse sources of heterogeneity is required. The second review searched systematically for empirical comparisons between MA-IPD and MA-APD. 70 comparisons were extracted from 25 empirical studies. Two thirds of the comparisons showed a tendency to overestimate the effect size and to reduce its precision by MA-APD in comparison to MA-IPD. However, the differences between the point estimates of both types of meta-analysis were small in all comparisons but one, and the paired t-test found no significant difference between the two types of meta-analysis. Furthermore, only half of the studies reported the results of a heterogeneity test. The methodological quality of RCTs was assessed in both types of meta-analysis only in a quarter of the studies. Clinical heterogeneity was investigated only in one third of the studies, using both types of meta-analysis, with no consistent results. A published MA-IPD on the efficacy of statins was reanalysed using aggregate patient data. The summary effect sizes of all cause mortality in both meta-analyses were extremely similar. Although the Cochran test of heterogeneity was significant, this was neither reported nor taken into account in MA-IPD. MA-APD combined the RCTs using a random effects model. The methodological variations between the RCTs were not assessed in MA-IPD. MA-APD investigated possible methodological and clinical causes of variation between the RCTs. These were determined a priori and justified with external evidence. Subgroup analysis and meta-regression were used to explore the relation between the selected causes of heterogeneity and mortality. By using a multivariate regression model that included terms for the methodological quality and baseline low-density lipoprotein a significant negative relationship between the inclusion of women in RCTs and the extent of reduction in mortality was detected. This result suggests a lower efficacy of statins in women which deserves further investigation in future RCTs. It can be concluded that the investigation of various sources of heterogeneity in meta-analysis of RCTs is still rare. Careful exploration of chance and methodological variations should precede the investigation of clinical heterogeneity and confounding between the sources of heterogeneity should be taken into consideration.

Inhaltsverzeichnis **Seite**

Abkürzungsverzeichnis

ACE-Hemmer	Hemmer des Angiotensin konvertierenden Enzyms (Angiotensin Converting Enzyme)
APDs	Aggregierte Patientendaten
Basis-LDL-K	Durchschnittliche Konzentration von Low-Density-Lipoprotein am Anfang einer Studie
EbM	Evidenz-basierte Medizin
FEM	Fixed-Effects-Modell
HR	Hazard-Ratio
HTA	Health Technology Assessment
IPDs	Individuelle Patientendaten
ITT-Analyse	Intention-to-Treat-Analyse
KI	Konfidenz-Intervall
MA	Meta-Analyse
MA-APD	Meta-Analyse mit aggregierten Patientendaten
MA-IPD	Meta-Analyse mit individuellen Patientendaten
NCTs	Nicht-randomisierte kontrollierte Studien (Non-randomised Controlled Trials)
OR	Odds-Ratio
Q-Test	Cochran-Test
RCT	Randomisierte kontrollierte Studien (Randomised Controlled Trials)
RCTs-HF	RCTs mit hohem Frauenanteil
RCTs-NF	RCTs mit niedrigem Frauenanteil
RCTs-HL	RCTs mit hoher Basis-LDL-K
RCTs-NL	RCTs mit niedriger Basis-LDL-K
RCTs-HQ	RCTs mit hoher methodischer Qualität
RCTs-NQ	RCTs mit niedriger methodischer Qualität
REM	Random-Effects-Modell
RR	Relatives Risiko
SR	Systematische Review (Systematic Review)
SMD	Standardisierte Mittelwertdifferenz
tRCT	trunkierte randomisierte kontrollierte Studien (truncated Randomised Controlled Trials)

Tabellenverzeichnis **Seite**

Abbildungsverzeichnis

Einleitung

Mit dem stetigen Anstieg der Anzahl randomisierter kontrollierter Studien (RCTs) [Dickersin, 2003] erhöht sich der Bedarf nach systematischen Zusammenfassungen ihrer Ergebnisse mittels systematischer Reviews (SRs) und Meta-Analysen (MAs). Allerdings weisen die Effektgrößen von RCTs zu ein und derselben Fragestellung oft große Unterschiede auf [Engels, 2000]. Diese Unterschiede, die im Rahmen der Evidenz-Synthese „Heterogenität" genannt werden, können die statistische Zusammenfassung der Ergebnisse von RCTs durch eine MA unmöglich machen oder erschweren.

Zunehmend wird MA nicht nur als ein Verfahren zur Evidenz-Synthese gesehen, sondern auch als eine Methode zur Heterogenitäts-Analyse betrachtet [Sutton, 2008]. Damit wird ihre synthetische Funktion um eine analytische Funktion ergänzt. Durch die Heterogenitäts-Analyse können wichtige klinische Unterschiede bezüglich des Interventionseffekts in Subgruppen der Patienten gefunden werden. Dies kann zur Therapieoptimierung durch Maßschneidern der Behandlung nach bestimmten Merkmalen der Patienten führen.

Heterogenität in MAs kann auf zufallsbedingte, methodische oder klinische Variation zwischen den RCTs zurückgeführt werden. Die klinische Heterogenität, die für die Patientenversorgung relevant ist, kann im Rahmen von MAs erst untersucht werden, wenn zufallsbedingte und methodische Unterschiede zwischen den RCTs berücksichtigt wurden.

Die vorliegende Arbeit beschäftigt sich mit der Berücksichtigung von Heterogenität in MAs von RCTs. Im ersten Kapitel wird eine kritische Einführung in MA dargestellt, die einen vom Verfasser entwickelten Leitfaden zur Durchführung von MAs einschließt. Danach werden vorhandene Methoden zur Berücksichtigung von Heterogenität bei der Synthese präsentiert, einschließlich aktueller Ansätze zur Schätzung der Inter-Studien-Varianz. Subgruppen-Analysen und Meta-Regression werden im Rahmen der Vorstellung vorhandener Methoden zur Untersuchung von Heterogenitätsursachen ausführlich diskutiert. Zusammenfassend werden Einschränkungen bei der Untersuchung von klinischer Heterogenität anhand bisheriger RCTs beschrieben und Lösungsvorschläge skizziert. Ein umfassender Ansatz zur Bewertung der methodischen Qualität von RCTs, der auch die Komponenten der Fragestellung miteinbezieht, wird ausführlich beschrieben. Im zweiten Kapitel wird durch eine vom Verfasser durchgefühhrte SR der Einfluss verschiedener Biasarten in RCTs auf die Ergebnisse von MAs untersucht und quantifiziert. Im dritten Kapitel wird eine weitere SR der Vergleiche zwischen MAs mit individuellen und MAs mit aggregierten Patientendaten konzipiert und durchgeführt. Anhand einer Fallstudie zu Statinen wird im vierten Kapitel die

Berücksichtigung von zufallsbedingter und methodischer Heterogenität bei der Unterschung von klinischer Heterogenität demonstriert. Abschließend werden die beiden SRs und die Fallstudie, die im Rahmen der Dissertation durchgeführt wurden, zusammengefasst.

1. Forschungsstand

1.1. Kritische Einführung in Meta-Analyse

Bereits vor einem Jahrhundert wurde Meta-Analyse (MA) verwendet als ein Verfahren zur Evidenzsynthese für präventive Gesundheitstechnologien und vor über 50 Jahren für therapeutische Interventionen [Egger, 2005]. Der Begriff „Meta-Analysis“ wurde zur Bezeichnung dieses Syntheseverfahrens allerdings erst vor drei Jahrzehnten vorgeschlagen [Glass, 1976]. MA fungiert nicht nur als ein standardisierbares Instrumentarium der Evidenzsynthese und der Heterogenitäts-Analyse, sondern auch als ein kritischer Ansatz der Diagnose von Fehl-, Über- und Unterforschung (s. Abschnitt 1.1.3.5). Allerdings besteht ein enormer Bedarf, bestimmte methodische Aspekte der MA weiter zu entwickeln [Parmigiani, 2002]. Die Entwicklung valider Strategien für die Suche nach Primärstudien [Sanders, 2005; Haynes, 2005], die als Bausteine der MA zu betrachten sind, sowie für die Prävention, Detektion und Modellierung des Publication-Bias [Bennett, 2004; Dickersin, 2003; Terrin, 2003; Macaskill, 2001] dient der Verringerung von Selection-Bias in MA. Beachtenswerte methodische und wissenschaftspolitische Fortschritte zur Verringerung von Selection-Bias in MA wurden erzielt. Sie werden in den Abschnitten 1.1.4.3 und 1.1.9.3 dargestellt. Noch nicht ausreichend entwickelt sind allerdings die bisherigen Modelle zur Berücksichtigung statistischer Heterogenität bei der Synthese von Primärstudien und die vorhandenen Methoden zur Untersuchung diverser Ursachen für die Variation zwischen den Primärstudien.

1.1.1. Definition von Meta-Analyse

Das “Dictionary of Epidemiology” von Last definierte “Systematische Review” (SR) als “*The application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic. Meta-analysis may be, but is not necessarily, used as part of this process.*” und MA als “*The statistical synthesis of the data from separate but similar, i.e. comparable studies, leading to a quantitative summary of the pooled results.*” [Last, 2001]. Im Rahmen dieser Arbeit wird eine MA als eine SR mit einer gewichteten, statistischen Kombination aller oder einiger vorherbestimmter Subgruppen der in der SR eingeschlossenen Primärstudien behandelt, wobei die Heterogenität zwischen den Primärstudien angemessene Berücksichtigung finden muss. Die Schwerpunkte dieser Definition liegen einerseits auf der Systematisierung der Suchen nach und der Auswahl von

Primärstudien, um Selection-Bias bei der Bestimmung und Eignung der Evidenzgrundlage zu vermeiden, und andererseits auf der Berücksichtigung der zufallsbedingten, methodischen, und klinischen Heterogenität eingeschlossener Primärstudien, um Fehler systematischer und zufälliger Art bei der Synthese und Analyse der Evidenzgrundlage zu reduzieren. MA bezieht sich auf die systematische Recherche nach Primärstudien zu einer präzisen Fragestellung, auf die explizite Bewertung von deren Qualität und auf die gewichtete, statistische Zusammenfassung ihrer Ergebnisse. MA kann als Sekundärstudie betrachtet werden, die den Forschungsstand über eine Fragestellung zusammenfasst und analysiert, wobei die in der MA eingeschlossenen Primärstudien als die Synthese- und Analyseeinheiten dienen. MA und quantitative SR gelten als Synonyme [Delgado-Rodríguez, 2001]. Die Ergebnisse der Primärstudien werden im Rahmen einer MA als Effektgrößen genannt und deren gewichtete Kombination wird als zusammengefasste Effektgröße bezeichnet.

Bemerkenswert ist die Tatsache, dass die „ National Library of Medicine“ in den USA MA als Publikationstyp zwar seit 1993 führte, aber bislang keinen Publikationstyp „Systematic Review“ eingeführt hat. Dies wurde mit dem Mangel an einheitlicher Definition von SR begründet [Schulman, 2005]. Ein aktuelles Konsensverfahren, das SRs durchführende Institutionen in Großbritannien und internationale Anbieter Public-Health-relevanter elektronischer Datenbanken einbezog, führte nur zu mäßiger Übereinstimmung über die Definition und die Qualitätsmerkmale von SRs [Sander, 2006].

1.1.2. Historische Meilensteine für Meta-Analyse

Erste Ansätze zur Kombination von Beobachtungen aus verschiedenen Primärstudien wurden im 18. und 19. Jahrhundert von Astronomen und Mathematikern wie George Biddell Airy, Johann Carl Friedrich Gauß und Pierre-Simon Laplace entwickelt [O'Rourke, 2006]. Der Statistiker Karl Pearson gilt als der erste Wissenschaftler, der Ergebnisse aus verschiedenen klinischen Studien statistisch kombiniert hat. 1904 veröffentlichte er im „British Medical Journal“ eine quantitative Zusammenfassung von 11 Primärstudien zum Zusammenhang zwischen der Impfung gegen Typhus und der Mortalität bzw. Infektion bei britischen Soldaten. Er kalkulierte den Mittelwert der Korrelationskoeffizienten der Primärstudien. Drei Jahre später veröffentlichte der Epidemiologe Josef Goldberger in „Hygienic Laboratory“ den Mittelwert der Ergebnisse von nach Kriterien selektierten 26 beobachtenden Primärstudien zur Häufigkeit von Harnwegsinfektionen bei Patienten mit typhoidem Fieber [Winkelstein, 1998]. 1932 stellte der Statistiker Ronald Aylmer Fisher eine Methode zur Kombinierung von p-Werten aus agrarwirtschaftlichen Studien dar [Chalmers, 2002] und gab eine Anregung zu einem der ersten Artikel zum Publication-Bias [Sterling, 1959; nach: O'Rourke, 2006]. Fünf

Jahre später präsentierten die Statistiker Frank Yates und William Gemmell Cochran die Invers-Varianz-Methode zur gewichteten Kombinierung von Effektgrößen aus agrarwirtschaftlichen Primärstudien [Hunt, 1997]. Die erste MA über eine therapeutische Intervention wurde von Henry K. Beecher 1955 in „The Journal of American Medical Association“ veröffentlicht. Dabei wurde die Wirksamkeit von Placebo bei 35% der Patienten verschiedener Konditionen wie postoperative Wundschmerzen, Husten und Angina Pectoris gefunden [Egger, 2005].

Die methodische Weiterentwicklung der MA wurde in den 1970er Jahren von Sozialwissenschaftlern, insbesondere im Bereich der Psychologie und Bildungsforschung, getragen. Der Begriff MA wurde zuerst durch den Psychologen Gene V. Glass 1976 eingeführt, als "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" definiert und als „analysis of analysis“ bezeichnet [Glass 1976, Seite 3]. Drei Jahre später thematisierte der Arzt und Epidemiologe Archie L. Cochrane den Mangel an kritischen Reviews aller relevanten randomisierten kontrollierten Studien (Randomised Clinical Trials= RCTs) und 20 Jahre, nachdem er 1972 das für „Evidenz-basierte Medizin“ (EbM) als bahnbrechend geltende Buch: „Effectiveness and efficiency: random reflections on health services“ veröffentlicht hatte, wurde die nach ihm benannte, weltgrößte Kollaboration von Reviewern gegründet: die Cochrane Collaboration [Chalmers, 2002; Guillemin, 2006]. Die Biostatistikerin Rebecca DerSimonian entwickelte 1986 ein Random-Effects-Modell mit einem einfachen und nicht-iterativen Schätzer für die Inter-Studien-Varianz [DerSimonian, 1986]. Seit 1987 setzte sich der Biostatistiker Richard Peto für die Synthese nur im Fixed-Effects-Modell ein [O’Rourke, 2006].

1.1.3. Potenziale von Meta-Analyse

1.1.3.1. Synthese von Primärstudien

Mit der wachsenden Zahl der Primärstudien in der medizinischen Forschung erhöht sich der Bedarf an Methoden zur Synthese ihrer Ergebnisse. Es wird geschätzt, dass Fachleute in Gesundheitsberufen ca. 17-20 Originalartikel pro Tag lesen müssten, um ihren Wissensstand aktuell zu halten [Davidoff, 1995]. Dies kann durch standardisierte Zusammenfassungen in Form von SRs oder MAs erreicht werden. Jährlich werden über zwei Millionen Aufsätze in mehr als 20 000 biomedizinischen Zeitschriften publiziert [Mulrow,

1994]. Dickersin geht davon aus, dass seit der einen Meilenstein setzenden RCT zum Antibiotikum Streptomycin für pulmonale Tuberkulose, 1948, etwa eine Million nicht-randomisierte kontrollierte Studien (Non-randomised Controlled Trials = NCTs) und RCTs durchgeführt, aber nur die Hälfte davon publiziert wurden [Dickersin, 2003]. In der seit 1966 etablierten Literaturdatenbank MEDLINE wurden im Februar 2002 über 4 600 Zeitschriften mit etwa 11 Millionen Eintragungen indexiert, wobei etwa 120 neue Zeitschriften jedes Jahr hinzukommen; 30% der Einträge stammen aus den USA [Scheinfeld, 2003; Kotzin, 2005]. Die eher europäische Publikationen verzeichnende Datenbank EMBASE, die seit 1974 besteht, bibliographierte im Mai 2002 mehr als 4000 Zeitschriften mit etwa 9 Millionen Einträgen [Scheinfeld, 2003]. Das „Cochrane Central Register of Controlled Clinical Trials“ (CENTRAL) beinhaltete im zweiten Quartal 2007 495 002 Berichte zu NCTs und RCTs. Im Juni 2007 wurden in MEDLINE 234 681 Einträge zu RCTs bibliographiert.

Über 40 000 laufende klinische Studien (NCTs und RCTs) sind jeweils in den Studienregistern „ClinicalTrials.gov“ (Stand: April 2007) [Laine, 2007] und „CenterWatch“ (Stand: Mai 2003) eingetragen [Dickersin, 2003]. Bei der Betrachtung der jährlichen Rate von klinischen Studien und von SRs wurde geschätzt, dass es bis zum Jahr 2015 dauern wird, bis die vorhandenen klinischen Studien durch etwa 10 000 SRs zusammengefasst sein werden [Mallett, 2003].

1.1.3.2. Erhöhung der statistischen Power

Eine klinische Studie kann ein nicht signifikantes Ergebnis über den Unterschied zwischen einer Interventions- und einer Kontrollgruppe liefern, wenn in Wahrheit jedoch ein Unterschied besteht. Dieses falsch negative Ergebnis wird in der Biometrie als Fehler zweiter Art oder Beta-Fehler bezeichnet. Seine Wahrscheinlichkeit wird bei der Fallzahlkalkulation konventionell auf 10% oder 20% festgelegt. Die statistische Power einer Studie ist die Wahrscheinlichkeit, keinen Beta-Fehler zu begehen (1-β), d.h. Unterschiede zu erkennen, wenn sie bestehen. Dieser Fehler tritt oft dann auf, wenn die Fallzahl in klinischen Studien gering ist, was in der medizinischen Forschung nicht selten vorkommt. Eine SR zweiarmiger, nicht signifikanter RCTs über Wirbelsäulenchirurgie zeigte, dass bei den 37 identifizierten Primärstudien lediglich 17% über Fallzahlkalkulation berichteten und dass eine durchschnittliche Prävalenz von über 20%igen Beta-Fehlern bei 82% der Primärstudien bestand [Bailey, 2004]. Weiterhin wurde in einer SR von 117 RCTs zur Behandlung von Frakturen bei älteren Menschen eine 90% Rate von über 20%igen Beta-Fehlern verzeichnet [Lochner, 2001]. Ähnliche Ergebnisse zur Epidemiologie der Beta-Fehler bei nicht

signifikanten RCTs wurden in der Dermatologie [Williams, 1993], Notfallmedizin [Brown, 1987] und Psychiatrie [Edlund, 1985] gefunden. Eine Untersuchung von 1 941 RCTs zu Behandlungen von Schizophrenie ergab, dass lediglich 3% von ihnen über eine für die Detektion klinisch relevanter Effektgrößen ausreichende Fallzahl verfügten [Thornley, 1998].

MA kann die statistische Power erhöhen, insbesondere bei kleinen „underpowered" Primärstudien und bei Primärstudien mit seltenen erwünschten oder unerwünschten Interventionseffekten.

1.1.3.3. Verbesserung von externer Validität

Primärstudien sollen nicht nur die wahre Effektgröße einer Intervention möglichst richtig einschätzen (interne Validität), sondern auch die Schätzung an repräsentativen Stichproben der Patienten durchführen. Es ist nicht mehr umstritten, dass RCTs ein Studiendesign mit hoher interner Validität darstellen. Da die meisten bisherigen RCTs aus verschiedenen Gründen eher selektierte Patientengruppen einschließen, bestehen weiter Bedenken, ob ihre Ergebnisse auf mehr als ein eng umgrenztes Kollektiv der Betroffenen übertragbar ist (externe Validität). Allerdings ist festzustellen, dass interne Validität die Voraussetzung für die externe Validität ist, d.h., dass die Ergebnisse einer verzerrten Primärstudie keine Anwendbarkeit aufweisen.

Die Generalisierbarkeit der Ergebnisse von Primärstudien auf nicht in sie eingeschlossene Patientengruppen, Interventionsregime und Rahmenbedingungen sowie die Übertragbarkeit ihrer Ergebnisse auf einzelne Patienten sind ein oft zitierter Kritikpunkt an EbM und deren Hauptinstrumenten: SR und MA. Beobachtende Primärstudien zeigten diskordante Ergebnisse bezüglich der Wirksamkeit von Interventionen bei Patienten, die nicht in RCTs eingeschlossen wurden. Während eine beobachtende Primärstudie zu Beta-Blockern bei akutem Myokardinfarkt diese Medikamente auch bei oft von RCTs ausgeschlossenen Patienten, z.B. Patienten mit Herzinsuffizienz oder pulmonaler Krankheit, wirksam fanden [Gottlieb, 1998], zeigte eine beobachtende Primärstudie zu Hemmern der Angiotensin konvertierenden Enzyme (ACE-Hemmer) keine Mortalitätsreduzierung bei Herzinsuffizienz-Patienten mit diastolischer Dysfunktion, die oft nicht in RCTs eingeschlossen waren [McAlister, 1999].

Da die beste verfügbare interne Validität durch RCTs erreicht werden kann und da RCTs in unterschiedlichen Subpopulationen, mit variablen Interventionsregimen und unter diversen

Bedingungen durchgeführt werden können, kann eine Zusammenfassung ihrer Ergebnisse mittels SRs und MAs die externe Validität verbessern.

1.1.3.4. Berücksichtigung und Untersuchung von Heterogenität

Variationen des Effekts einer Intervention sind innerhalb und zwischen den Primärstudien zu finden. Während die Variabilität innerhalb einer RCT oft auf Streuungen bei der Stichprobenziehung züruckgeführt werden kann, gehen die Unterschiede zwishen den Primärstudien nicht selten auf methodische und klinische Heterogenität züruck (s. Abschnitt 1.2).

Die Berücksichtigung von Heterogenität in SRs und MAs scheint bisher in der Forschung nicht etabliert zu sein und nach Medizingebiet zu variieren. Eine SR, die 38 MAs von klinischen Studien zur Infektion mit Helicobacter pylori einschloss, fand heraus, dass etwa die Hälfte der MAs das Ergebnis eines Heterogenitäts-Tests berichtete. 11 MAs gaben das Signifikanzniveau genauer an, davon fassten über die Hälfte, trotz statistischer Heterogenität, die Ergebnisse meta-analytisch zusammen. 40% der 38 MAs berichteten die Auswahl des Synthese-Modells und 26% begründeten diese [Huang, 2004]. Eine SR, die 256 SRs mit einem oder mehreren unerwünschten Ereignissen als Primärendpunkt einschloss, zeigte, dass 90% der SRs Heterogenität evaluierten, etwa 40% keine MA durchführten, von denen etwa ein Drittel die narrative Synthese begründeten und dies am häufigsten mit Heterogenität rechtfertigten [Golder, 2006a].

Die Untersuchung des Interventionseffekts in Subgruppen der Patienten dient an erster Stelle der Überprüfung der Konsistenz des Effekts in diesen Subgruppen. Falls bezüglich des Interventionseffekts quantitative (d.h. Unterschiede in dem Ausmaß, aber nicht in der Richtung des Effekts) oder qualitative (d.h. Unterschiede im Ausmaß und der Richtung des Effekts) Variationen zwischen den Subgruppen gefunden werden, können sie von hoher klinischer Relevanz sein [Glasziou, 1998]. So können Patientensubgruppen identifiziert werden, die mehr, weniger oder keinen Nutzen aus der Intervention ziehen. Dies kann zu einer maßgeschneiderten Patientenversorgung führen.

1.1.3.5. Monitoring des Forschungsbedarfs

Es wurde davon ausgegangen, dass die Mehrheit der Ergebnisse von Primärstudien falsch [Ioannidis, 2005] oder mit großen Unsicherheiten verbunden ist [Djulbegovic, 2005]. Glasziou und Kollegen schlagen vor, dass bei einem Odds-Ratio (OR) von über 10 von einem „realen" Interventionseffekt auszugehen ist, auch wenn Confounders existieren [Glasziou, 2007]. In solchen Fällen kann eine Evaluation der Wirksamkeit durch RCTs unnötig sein. Allerdings ist die Liste von Interventionen, die sehr große Effektgrößen aufweisen, nicht lang, z.B. Fallschirm für freien Fall [Smith, 2003], orale Rehydratation für Säuglingsdiarrhoe [Potts, 2006], Insulin für Diabetiker, Bluttransfusion für hämorrhagischen Schock, Sulphanilimide für puerperale Sepsis, Tracheostomie für tracheale Obstruktion, Äther für Anästhesie oder Fundoplikation für Pyrosis [Glasziou, 2007]. Die meisten Gesundheitsinterventionen weisen dagegen mittlere bis geringe Effektgrößen auf.

Die Evaluation von Interventionen mit weltweit kaum koordinierten experimentellen Primärstudien, ohne Zwischenbilanzen anhand von SRs bzw. MAs zu ziehen, wurde als wissenschaftliche Verfehlung bezeichnet und als ethisches Dilemma diskutiert [Chalmers, 2005; Cooper, 2005]. Schon vor 18 Jahren zeigte eine kumulierte MA, dass die Gabe von Kortikosteroiden an Mütter mit erwarteten Frühentbindungen die Mortalität und Morbidität der Säuglinge signifikant reduziert. Dieses Ergebnis wäre mit den damals veröffentlichten Primärstudien bereits 10 Jahre vorher festzustellen gewesen, hätte man eine MA durchgeführt [Crowley, 1990]. Weiterhin zeigte ein Vergleich der Ergebnisse von kumulativen MAs der RCTs und den in Standardlehrbüchern und einschlägigen unsystematischen Reviews gegebenen Empfehlungen der Fachexperten zu Interventionen für die Behandlung von Myokardinfarkt, dass letztere zu manchen Interventionen 10 Jahre hinter dem Wissensstand lagen und dass mehrere Experten Interventionen ohne Effekt auf die Gesamtmortalität oder sogar mit potenziellem Schaden für die Patienten weiterempfahlen [Antman, 1992]. Eine kumulative MA, die 33 zwischen 1959 und 1988 publizierte klinische Studien zum Thrombolytikum Streptokinase für akuten Myokardinfarkt identifizierte, fand heraus, dass das zusammengefasste OR zur Gesamtmortalität schon 1973, als die MA lediglich 8 klinische Studien mit weniger als 8% der in allen Studien eingeschlossenen Patienten kumulierte, sich bei markanten 0,74 stabilisierte [Lau, 1992]. Eine kumulative MA, die 64 zwischen 1987 und 2002 veröffentlichte RCTs zum Proteasehemmer Aprotinin für kardiale Chirurgie einschloss, zeigte, dass das zusammengefasste OR zur Reduzierung präoperativer Transfusion schon 1992, als die MA lediglich 12 RCTs mit weniger als 30% der in allen klinischen Studien eingeschlossenen Patienten kumulierte, sich bei unverkennbaren 0,25 verankerte [Fergusson, 2005]. Es ergaben sich keine qualitativen Änderungen der

Punktschätzer und des Konfidenz-Intervalls aus den Sensitivitäts-Analysen bezüglich mehrerer potenzieller Ursachen für methodische und klinische Heterogenität [Fergusson, 2005].

Monitoring durch SRs und MAs versetzt alle Betroffenen, vor allem Patienten, in die Lage, rechtzeitig robuste Evidenz sowohl für bestehenden Nutzen als auch für fehlenden Nutzen bei diesbezüglich konsistenten Effekten zu ermitteln und damit Überforschung zu vermeiden. Eine sequenzielle MA, die zwischen 1999 und 2003 fünf veröffentlichte RCTs zur hochfrequenten oszillatorischen Ventilation versus konventioneller mechanischer Ventilation für Frühgeburt lokalisierte, beobachtete, dass das OR schon nach Einschluss der ersten (13% aller Patienten) bzw. der zweiten RCT (27% aller Patienten) die statistische Entscheidungsgrenze für fehlende Reduzierung der Gesamtmortalität oder chronischer Pneumonie um 15% bzw. 10% überschritt. Dieses Ergebnis blieb robust bei Sensitivitäts-Analysen zu Interventionsmerkmalen [Bollen, 2006].

Des Weiteren ermöglicht die Begleitung von Primärstudien mit SRs und MAs die Identifizierung von Evidenzbreschen und untererforschten Fragestellungen. Eine Analyse aller Ende 2005 in der „Cochrane Library“ indexierten SRs fand heraus, dass lediglich 3,2% der SRs keinen weiteren Forschungsbedarf konstatieren und 21,2% auf eine laufende oder geplante Primärstudie hinweisen [Clarke, 2007a]. Zudem können SRs und MAs zur Aufdeckung von Duplikatpublikation (s. Abschnitt 1.1.4.3) und von Betrug bei Primärstudien beitragen. Während einer SR zur Epiduralanalgesie bei der Geburt wurde eine Primärstudie als ein klarer Fall von Plagiat entdeckt [Chalmers, 2006].

Bei wenigen Zeitschriften (British Medical Journal, Lancet) und Sponsoren (UK „Medical Research Council“, „Wellcome Trust“) ist eine Tendenz zur Reduzierung von alleinstehenden Primärstudien und zur Forderung der Einbettung neuer Primärstudien in den vorherigen Wissensstand zu beobachten [Chalmers, 2005]. Obwohl die erste Version der „Consolidated Standards of Reporting Clinical Trials“ (CONSORT-Statement) dies auch empfahl [Begg, 1997], wurde es in der überarbeiteten Version getilgt [Altman, 2001]. Diese Entwicklung entbehrt jeglicher Legitimation und ist als Rückschritt zu beurteilen.

Auf dem jeweils aktuellen Stand der Forschung basierende SRs und MAs befähigen regulatorische Instanzen, „Institutional Review Boards“, Sponsoren, prospektive Prüfärzte und Studienteilnehmer sowie Leistungserstatter, Leistungserbringer und Patienten, Evidenz-

basierte Entscheidungen zur Vornahme, Gestaltung bzw. Auswertung neuer klinischer Studien zu fällen.

1.1.3.6. Konzeptualisierung von neuen Primärstudien

Rechtfertigung, Design und Management neuer Primärstudien sollen anhand von SRs bzw. MAs erfolgen. Aus ethischer Sicht darf eine Interventionsstudie, z.B. eine RCT, erst durchgeführt werden, wenn bei bisheriger Evidenzlage bezüglich der Intervention weiter Unsicherheit besteht. Ob die Unsicherheit durch eine SR bzw. MA, einen Expertenkonsens oder individuelle Präferenzen von Ärzten und Patienten zu bestimmen ist, bleibt weiterhin umstritten [Lilford, 1995; Lilford, 2001; Sackett, 2001; Gifford, 2001]. Außerdem versäumen die Prüfärzte, bei Nicht-Inanspruchnahme der Ergebnisse von SRs bzw. MAs von Stärken und Schwächen vorheriger Primärstudien zu lernen. Evidenz-basiertes Studiendesign soll die Validität der klinischen Studien anreichern und die Interessen der Probanden stärken. Effektgrößen und Studienaustritte vorheriger RCTs können beispielsweise für die Fallzahlkalkulation und die Strategien zur Bindung der Patienten wichtige Informationen liefern. Zur Basierung der Fallzahlkalkulation für neue RCTs auf aktuellen MAs wurde ein Bayesianischer Ansatz vorgeschlagen [Sutton, 2007].

Seit 1998 verlangen die Ethikkommissionen in Dänemark, dass Prüfärzte bei der Antragstellung für die Genehmigung neuer klinischer Studien eine aktuelle und umfassende SR durchführen, bzw. vorlegen und diese während des Studienverlaufs aktualisieren und bei der „Beantwortung“ der Fragestellung durch rezente Evidenz die Studie frühzeitig beenden [Goldbeck-Wood, 1998]. Anträge für die Finanzierung neuer klinischer Studien bei dem UK „Medical Research Council“ und dem „Wellcome Trust“ sollen eine SR beinhalten, die den Bedarf für die Studie legitimiert und die Studienergebnisse interpretiert [Chalmers, 2005]. Eine Befragung von 24 Prüfärzten zeigte, dass nur 42% von ihnen relevante SRs beim Studiendesign einbezogen [Cooper, 2005]. Drei Untersuchungen aller RCTs (n= 77), die im Mai 1997, 2001 und 2005 in fünf Fachzeitschriften (Annals of Internal Medicine, British Medical Journal, Journal of the American Medical Association, Lancet und the New England Journal of Medicine) veröffentlicht wurden, fanden, dass die berichteten Ergebnisse von 72% der RCTs keine Einbettung im Kontext vorheriger RCTs zeigten, wobei keine Verbesserung über die Jahre beobachtet wurde [Clarke, 2007b; Clarke, 2002; Clarke, 1998a]. Eine MA, die 64 RCTs zum Proteasehemmer Aprotinin für die Reduzierung präoperativer Transfusion einschloss, zeigte, dass 3% der Studien vorherige SRs zitierten und der Median der Zitationen von ein Jahr zuvor publizierten Studien 20% betrug [Fergusson, 2005].

Eine SR bzw. MA ist kostengünstiger als eine große RCT. Ein HTA-Bericht von dem UK „National Institute for Health and Clinical Excellence“ kostet etwa 78.000 USD, eine klinische Studie des US „National Institutes of Health“ kostet im Durchschnitt 12.000.000 USD [Glasziou, 2006].

1.1.4. Einschränkungen von Meta-Analyse

1.1.4.1. Confounding bei der Untersuchung von Heterogenität

Obwohl Parallelen zwischen MA und den von Austin Bradford Hill ausgearbeiteten Kausalitätskriterien illuminiert wurden [Matt, 1997; Weed, 2000], soll MA - auch MA von RCTs - als beobachtende Studie gesehen werden, die anfällig ist für Confounding bei der Untersuchung von Heterogenitätsursachen zwischen den Primärstudien. Ein Zusammenhang zwischen den Effektgrößen von Primärstudien, die in eine MA eingeschlossen sind, und einer klinischen Kovariablen (z.B. Serum-Low-Lipid-Lipoprotein) kann durch eine starke Assoziation der klinischen mit einer methodischen Kovariablen (z.B. Allocation Concealment) verzerrt werden. Weiterhin kann eine Assoziation zwischen den Effektgrößen und einer klinischen Kovariablen durch eine weitere klinische Kovariable (z.B. Dosis eines Statins) verzerrt werden. Da MA eine beobachtende Studie ist und da Heterogenität zwischen den eingeschlossenen Primärstudien nicht selten multifaktorielle Ursachen aufweist, sollen zufallsbedingte, methodische und klinische Variationen der Primärstudien gleichzeitig untersucht werden. Confounding durch nicht untersuchte oder unbekannte Heterogenitäts-Determinanten ist in MA möglich, findet bislang aber kaum Beachtung (s. Abschnitte 2.5 und 3.5). Ein im Rahmen einer MA gefundener Zusammenhang zwischen dem Interventionseffekt und einer methodischen oder einer klinischen Kovariablen soll in der Regel in weiteren Studien, vorzugsweise in RCTs, überprüft werden [Thompson, 1999].

1.1.4.2. Niedrige statistische Power bei der Untersuchung von Heterogenität

Sowohl bei MAs mit individuellen Patientendaten (MA-IPDs) als auch bei MAs mit aggregierten Patientendaten (MA-APDs) soll Heterogenität zwischen den Primärstudien berücksichtigt werden. Die Untersuchung von zufallsbedingter, methodischer und klinischer

Heterogenität zwischen den Primärstudien hängt von der Zahl und der Größe der in die MA eingeschlossenen Primärstudien ab. Empirische Arbeiten wiesen darauf hin, dass mehr als 50% der MAs weniger als 10 Primärstudien einschließen [Engels, 2000; Sterne, 2000]. Demzufolge wird die Untersuchung von Heterogenitätsursachen auf Primärstudien-Ebene oft durch die niedrige statistische Power der MA eingeschränkt (s. Abschnitt 1.2.5). Während MA-IPD dank hoher statistischer Power einen großen Vorteil bei der Untersuchung von Patientenheterogenität anbietet, gelingt es MA-APD mit niedriger Anzahl von Primärstudien aufgrund hoher Beta-Fehler nicht, Patientenvariationen bezüglich des Interventionseffekts zu finden (s. Abschnitt 1.2.3).

1.1.4.3. Verzerrung durch Publication-Bias

1.1.4.3.1. Begriffsbestimmung

Publication-Bias ist eine Haupteinschränkung der wissenschaftlichen Forschung, nicht nur der SR oder der MA. Sie tritt auf, wenn publizierte Primärstudien sich in Bezug auf ihre Ergebnisse systematisch von unpublizierten Studien unterscheiden, d.h. die Ergebnisse von Primärstudien beeinflussen deren Wahrscheinlichkeit, publiziert zu werden und damit für SR und MA identifizierbar und zugänglich zu werden. Damit werden die in einer SR/ MA eingeschlossenen Studien nicht für die Grundgesamtheit durchgeführter Studien repräsentativ sein [Rothstein, 2005]. Obwohl Publication-Bias in der psychosozialen Forschung seit 1956 [Smith, 1956; nach: Thornton, 2000] und im biomedizinischen, wissenschaftlichen Diskurs seit 1959 [Sterling, 1959; nach: Song, 2000] behandelt wird, brachte erst die Unterdrückung von Daten zum Antirheumatikum Rofecoxib [Tanne, 2006; Jüni, 2004; Konstam, 2001] und zum selektiven Serotonin-Wiederaufnahmehemmer Paroxetine [Whittington, 2004] das Problem verstärkt auch in die öffentliche Diskussion [Rothstein, 2005].

Der Ausschluss von unpublizierten Daten führt zur Reduzierung der statistischen Power und kann in verzerrten Schätzungen der MA resultieren. Es wurde beobachtet, dass Primärstudien mit stärkeren Effektgrößen oder statistisch signifikanten Ergebnissen öfter als diejenigen mit kleineren Effektgrößen oder nicht signifikanten Ergebnissen zur Veröffentlichung eingereicht und akzeptiert werden. Daher kann eine MA veröffentlichter Primärstudien zu einer überoptimistischen Schlussfolgerung führen [Sutton, 1998]. Dieser Bias trifft kleine Studien häufiger als große. Ungeachtet ihrer Ergebnisse werden große

Studien, aufgrund des damit verbundenen hohen Aufwands öfter als kleine Studien veröffentlicht [Sterne, 2001].

Selektive Publikation betrifft nicht nur Primärstudien als Ganzes, manchmal auch Endpunkte oder Subgruppen innerhalb veröffentlichter Studien. Publication-Bias ist ein Bestandteil von einem größeren Problem, sogenannter Dissemination-Bias, das zur Unterdrückung von Ergebnissen aus Primärstudien führt [Song, 2000]. Unten aufgeführt stehen verwandte Arten von Publication-Bias, die oft in der Literatur [Sutton, 1998; Rothstein, 2005] erwähnt werden:

1. Grey-Literature-Bias: Dieser tritt auf, wenn die Stärke und die Richtung der Ergebnisse der Primärstudien, die durch professionelle Verleger veröffentlicht werden, sich von den Ergebnissen schwer zugänglicher Primärstudien, sogenannte Grauliteratur z.B. Dissertationen, Konferenzberichte, Diskussionspapiere, Industrie- oder Behördenberichte, unterscheiden [Song, 2000; Hopewell, 2007a].
2. Language-Bias: Dieser tritt auf, wenn die Publikationssprache von der Stärke und der Richtung der Ergebnisse von Primärstudien abhängt [Gregoire, 1995; Jüni, 2002; Moher, 2003].
3. Time-Lag-Bias: Es dauert länger, Primärstudien mit nicht signifikanten Ergebnissen oder mit signifikanten Ergebnissen zugunsten der Kontrollintervention zu veröffentlichen als Primärstudien mit signifikanten Ergebnissen zugunsten experimenteller Intervention [Ioannidis, 1998; Hopewell, 2007b].
4. Outcome-Reporting-Bias: Bei Primärstudien mit multiplen Endpunkten werden lediglich signifikante Endpunkte zugunsten der experimentellen Intervention veröffentlicht [Chan, 2004a; Chan, 2004b].
5. Duplicate-Reporting-Bias: Dieselben Autoren veröffentlichen dieselben Ergebnisse einer Primärstudie in unterschiedlichen Fachzeitschriften oder verschiedene Autoren einer (Multizentren- oder Multinational-)Primärstudie publizieren dieselben Ergebnisse [Gøtzsche, 1989a; Tramer, 1997].
6. Retrieval-Bias: Dieser tritt auf, wenn die Verzeichnung in elektronischen Datenbanken von den Merkmalen primärer Studien abhängt (Database-Bias), wenn Inkonsistenz bei der Verschlagwortung ähnlicher Primärstudien in Datenbanken besteht (Coding-Bias) oder wenn die Zitationsrate von den Merkmalen primärer Studien abhängt (Citation-Bias).

Der Einschluss von Daten in SRs und MAs nach deren Publikationsstatus (Ausschluss von unpublizierten oder grauen Primärstudien), deren Publikationssprache (Ausschluss von nicht englischsprachigen Primärstudien), deren Publikationszeitpunkt (verspätete Publikation von

Primärstudien mit nicht signifikanten Ergebnissen oder mit signifikanten Ergebnissen zugunsten der Kontrollintervention), deren Wiederabrufstatus (Ausschluss von uneinheitlich oder nicht in Datenbanken indexierten oder selten zitierten Primärstudien) oder nach deren Endpunktstatus (Ausschluss von nicht signifikanten Endpunkten oder von signifikanten Endpunkten zugunsten der Kontrollintervention) kann also die Ergebnisse der Synthese durch Selection-Bias verzerren.

1.1.4.3.2. Methoden zur Identifizierung und Adjustierung

Vorhandene Methoden zur Identifizierung und Adjustierung für Publication-Bias stützen sich auf die Annahme des sogenannten „small study effect“. Letzterer bezieht sich auf Beobachtungen, dass kleine und große Primärstudien unterschiedliche Effektschätzer und Präzision aufweisen [Sterne, 2000]. Der Funnel-Plot ist eine graphische Darstellung der Primärstudien einer MA, wobei deren Studienpräzision oder deren Studiengrößen auf der horizontalen Achse gegenüber deren Effektgrößen auf der vertikalen Achse stehen. Beim Fehlen von Bias soll ein symmetrischer Plot im umgekehrten Trichter-Format entstehen, mit größerer Streuung der Effektgrößen kleinerer Primärstudien im unteren Plotteil und mit geringerer Streuung der Effektgrößen größerer Primärstudien im oberen Plotteil. Wenn der Funnel Plot asymmetrisch ist, kann das auf einen Publication-Bias aufgrund des Fehlens kleiner Primärstudien mit nicht signifikanten Ergebnissen hindeuten. Allerdings kann ein asymmetrischer Funnel Plot auch auf Zufall, verwendete Effektmaße, verwendete Präzisionsmaße oder Heterogenität der Effektgrößen, zurückzuführen sein, was bei der Interpretation des Plots miteinzubeziehen ist [Song, 2000]. Sterne und Egger empfahlen die Verwendung einer Relativgröße in der logarithmischen Skala, insbesondere das OR, als Effektmaß und Standardfehler als Präzisionsmaß [Sterne, 2001]. Da die Interpretation von Plots mit Subjektivität verbunden ist, soll der Funnel-Plot nur zur Exploration angewendet werden [Egger, 1997].

Die adjustierte Rangkorrelationsmethode [Begg, 1994] und die Methode der linearen Regression [Egger, 1997] sind statistische Methoden zum Testen für die Asymmetrie in Funnel Plots. Mehrere Simulationsstudien evaluierten, auch wenn sie mit unterschiedlichen Einschränkungen behaftet waren und zu variierenden Ergebnissen führten, diese Testverfahren und zeigten eine Überlegenheitstendenz des Egger-Tests im Vergleich zum Begg-Test in Bezug auf statistische Power [Begg, 1994; Sterne, 2000; Macaskill, 2001; Schwarzer, 2002].

Die „Trim-and-Fill“ Methode von Duval und Tweedie dient der Berechnung einer für Publication-Bias adjustierten zusammengefassten Effektgröße anhand einer Symmetrieherstellung des Funnel-Plots durch die Entfernung von Primärstudien auf einer Seite des Plots und die Imputation derselben Zahl von Primärstudien auf der anderen Seite des Plots [Duval, 2000]. Eine Simulationsstudie zeigte angemessene Eigenschaften dieses Modellierungsansatzes [Terrin, 2003]. Eine andere Simulationsstudie [Williamson, 2007] fand heraus, dass die Methode von Copas und Jackson zur Adjustierung für den Publication-Bias und Outcome-Reporting-Bias nützlich war [Copas, 2004].

Die Verwendung bisheriger Methoden zur Identifizierung, Quantifizierung und Adjustierung für Publication-Bias gilt nicht als angemessen bei:

- hoher Heterogenität (Heterogenitäts-Maß „I^2“ $\geq$ 50%)
- wenigen Primärstudien (n < 10)
- alle Primärstudien haben ähnliche Präzision (das Ratio extremer Varianzen ≤ 2)
- alle Primärstudien sind nicht signifikant [Sterne, 2000; Ioannidis, 2007a]

Starke Hinweise für Outcome-Reporting-Bias bestehen, wenn:

- kein Primärendpunkt definiert wurde und der berichtete Endpunkt signifikant ist
- ursachenspezifische Mortalität, aber nicht Gesamtmortalität berichtet wurde
- nur ein Teil von oft zusammen durchgeführten Untersuchungen, wie dem systolischen und diastolischen Blutdruck, berichtet wurde [Chan, 2004a; Chan, 2004b; Williamson, 2007]

1.1.4.3.3. Ansätze zur Prävention

Die prospektive Registrierung von Protokollen, verbunden mit dem freien Zugang zu den Ergebnissen primärer Studien, wurde seit über zwei Jahrzehnten als die beste Lösung für Publication-Bias vorgeschlagen [Simes, 1986; Rothstein, 2005]. Inzwischen gibt es politische Unterstützung staatlicher und überstaatlicher Organe (z.B. US „National Institutes of Health“, US „Food and Drug Administration“, UK „Medical Research Council“ „European Science Foundation“, „World Health Organisation“), verbindliche Unterstützung von Herausgebern mancher Fachzeitschriften (z.B. „International Committee of Medical Journal Editors“, „BioMed Central“) und Herstellern mancher pharmazeutischer Produkte (z.B. „Glaxo Wellcome“, „Schering Health Care“) sowie detaillierte Umsetzungsmodule für diesen Ansatz. Allerdings fehlen bisher ein ausreichendes öffentliches und professionelles Bewusstsein ebenso wie legaler Zwang und die Kooperation der Mehrheit der pharmazeutischen

Hersteller [Dickersin, 2003; DeAngelis, 2004]. Bestehende Register (z.B. „Current Controlled Clinical Trials“, „ClinicalTrials.gov“, „CenterWatch“) sind nicht ausreichend umfassend und mehrere Registrierungssysteme (z.B. „International Standard Randomised Controlled Trial Number“) können die Eindeutigkeit der registrierten klinischen Studien gefährden.

1.1.4.3.4. Empirische Untersuchungen

Der Zusammenhang zwischen Ergebnissen (Effektgröße, Präzision, statistische Signifikanz) und Merkmalen (z.B. Studiendesign, methodische Qualität, Publikationsstatus) von Primärstudien ist der Gegenstand der sich seit Mitte der 1990er Jahre entfaltenden methodischen Disziplin „Meta-Epidemiologie“. Obwohl das US „Cochrane Center“ seit 1948 Handsuchen von über 2200 Fachzeitschriften durchführt, zeigte diese kostenintensive Arbeit, dass seit der Gründung von MEDLINE 1966 mindestens ein Drittel der RCTs dort nicht verzeichnet wurden [Hopewell, 2002].

Eine SR mit 19 empirischen Studien zeigte, dass nur 63,1% von 30 394 Abstracts zu klinischen Studien im Volltext publiziert wurden, wobei signifikante Ergebnisse einen signifikanten Zusammenhang mit Volltextveröffentlichung zeigten [Sherer, 2007]. Eine SR zu nichtsteroidalen Antirheumatika zeigte, dass nur eine von 37 zwecks der Arzneimittel-zulassung für die US „Food and Drug Administration“ eingereichten RCTs publiziert wurde [MacLean, 2003]. Eine 1983 publizierte SR, die 14 klinische Studien zu Antiarrhythmika der Klasse I für akuten Myokardinfarkt einschloss, versäumte es, eine bereits 1980 abgeschlossene, aber erst 1993 publizierte klinische Studie, zu identifizieren. Der Einschluss von damals unpublizierten klinischen Studien in die SR hätte eine erhöhte Gesamtmortalität bei diesen Medikamenten ein Jahrzehnt früher festgestellt. So hätten schätzungsweise 50 000 bis 75 000 Patienten pro Jahr in den USA in den 1980er Jahren nicht das Leben durch die Medikamenteneinnahme verlieren müssen [Furberg, 1983; Teo, 1993].

Eine SR, die fünf empirische Studien zum Grey-Literature-Bias einschloss, fand heraus, dass publizierte klinische Studien im Durchschnitt eine größere Fallzahl und größere Interventionseffekte aufwiesen als klinische Studien aus der grauen Literatur [Hopewell, 2007a].

Eine SR mit 405 nicht zur Akupunktur publizierten klinischen Studien fand heraus, dass 95% der in bestimmten Ländern (China, Japan, Russland/UdSSR und Taiwan) publizierten klinischen Studien „positive“ Ergebnisse zeigten, während dies nur bei 75% der in England

publizierten klinischen Studien auch der Fall war [Vickers, 1998]. Eine SR von 303 MAs zeigte, dass nicht englischsprachige klinische Studien im Vergleich zu englischsprachigen klinischen Studien kleinere Studiengröße, niedrigere methodische Qualität sowie einen höheren Interventionsnutzen aufwiesen und öfter signifikant waren. Allerdings führte der Ausschluss von nicht englischsprachigen klinischen Studien zu einer kleinen Änderung der zusammengefassten Effektgrößen [Jüni, 2002]. Eine SR mit 130 SRs zeigte, dass englischsprachigen und nicht englischsprachigen RCTs einschließende MAs im Vergleich zu nur englischsprachige RCTs einschließenden MAs eine größere Zahl von RCTs und eine bessere methodische Qualität aufwiesen. Allerdings führte der Ausschluss von nicht englischsprachigen RCTs zu den der sogenannten „Schulmedizin“ zuzuschreibenden Interventionen aus den MAs zu keiner signifikanten Änderung der zusammengefassten Effektgröße. Eine ähnliche Sensitivitäts-Analyse bei MAs in der sogenannten „komplementären und alternativen Medizin“ resultierte in signifikanter Reduzierung der zusammengefassten Effektgröße um durchschnittliche 63% bei Ausschluss der nicht englischsprachigen RCTs [Moher, 2003].

Eine SR, die zwei empirische Studien mit 196 klinischen Studien zum Time-Lag-Bias einschloss, fand heraus, dass nur die Hälfte der klinischen Studien als Volltext publiziert wurde und dass die Vollpublikation bei Primärstudien mit nicht signifikanten Ergebnissen oder mit signifikanten Ergebnissen zugunsten der Kontrollintervention 6-8 Jahre und bei Primärstudien mit signifikanten Ergebnissen zugunsten experimenteller Intervention 4-5 Jahre dauerte [Hopewell, 2007b].

Für 122 RCTs wurden die Studienprotokolle, die einer dänischen Ethikkommission vorgelegt wurden, mit ihren Publikationen verglichen. Dieser Vergleich ergab, dass durchschnittlich (Median) 50% der Nutzenendpunkte und 65% der Schadenendpunkte pro RCT, die im Studienprotokoll angekündigt wurden, in der Publikation unvollständig berichtet wurden. Statistisch signifikante Endpunkte bezüglich des Interventionsnutzens und der -risiken wurden signifikant öfter vollständig berichtet als nicht signifikante Endpunkte. In 62% der RCTs wurde sogar der Primärendpunkt geändert, eingeführt oder ausgelassen [Chan, 2004a]. In einer weiteren empirischen Studie wurde für jede von 48 RCTs das Studienprotokoll, das einem kanadischen staatlichen Gesundheitsforschungssponsor vorgelegt wurde, mit ihrer Publikation verglichen. Im Durchschnitt (Median) wurden 31% der Nutzenendpunkte und 59% der Schadenendpunkte pro RCT unvollständig berichtet. Wiederum wurde die statistische Signifikanz des Endpunkts als signifikanter Prädiktor vollständiger Berichterstattung identifiziert. In 40% der RCTs wurden Diskrepanzen zwischen dem Protokoll und der Publikation bezüglich des Primärendpunkts gefunden [Chan, 2004b].

Das laufende ORBIT-Projekt („Outcome Reporting Bias in Trials“) versucht, durch Interviews mit Prüfärzten eine Methode zur Detektion dieses Bias zu evaluieren [Williamson, 2007].

Huth bezeichnete die multiple Publikation einer Studie als „Salami Science“ [Huth, 1986]. In einer MA der klinischen Studien zu nichtsteroidalen Antirheumatika waren 18% der 244 eingeschlossenen Publikationen Duplikate, wobei 73% der Duplikate „versteckt“ waren [Gøtzsche, 1989a], d.h. maskiert durch veränderte Autoren, Sprachen oder Datensätze und ohne Querverweis. Eine SR zum Antiemetikum Ondansetron fand, dass es sich bei 17% der 84 eingeschlossenen Publikationen von RCTs um „versteckte“ Duplikate handelte, und dass der Einschluss der Duplikate in die MA zur Überschätzung der zusammengefassten Effektgröße der Intervention um 23% führte [Tramer, 1997].

Da fehlende Primärstudien nach der Annahme des „small study effect“ kleine Studien mit niedrigem Gewicht sind, ist von keinem großen Impact des Publication-Bias auf MA auszugehen. Eine empirische Untersuchung von 48 Cochrane MAs zeigte, dass Publication-Bias die statistische Folgerung in weniger als 10% der MAs änderte [Sutton, 2000].

Weniger als ein Viertel von in MEDLINE im November 2004 publizierten SRs untersuchte das Potenzial für Publication-Bias [Moher, 2007]. Eine SR der meta-epidemiologischen Studien zum Impact von Publication-Bias auf die Ergebnisse von MA erkannte keinen Publication-Bias bei diesen methodischen Studien [Dubben, 2005]. Ebenso fand eine empirische Untersuchung keinen Publication-Bias bei MAs mit individuellen Patientendaten in der Onkologie [Tierney, 2000]. In einer Untersuchung von Terrin und Kollegen wurden nur 52,5% der Funnel-Plots, die 41 medizinischen Forschern vorgelegt wurden, korrekt bezüglich des Vorhandenseins oder Nichtvorhandenseins von Asymmetrie evaluiert [Terrin, 2005].

Nach Ausschluss der klinischen Studien ohne prospektive Registrierung fand eine MA keinen Nutzen mehr für Kombinations-Chemotherapie zum fortgeschrittenen ovarialen Krebs [Simes, 1986]. Die retrospektive Identifikation von unpublizierten klinischen Studien durch die Befragung von Geburtshelfern und Kinderärzten stellte sich als nicht erfolgreich heraus [Hetherington, 1989].

1.1.4.4. Verzerrung durch Interessenkonflikt

Sehr wenige SRs berichteten, dass sie durch eine gewinnorientierte Institution finanziert wurden [Moher, 2007]. Eine SR identifizierte acht Paare von durch die Cochrane

Collaboration und durch die pharmazeutische Industrie finanzierten MAs, die dieselbe Intervention und Erkrankung untersuchten, aber unterschiedliche Autoren aufwiesen und innerhalb von zwei Jahren publiziert wurden. Sie fand heraus, dass die vorbehaltlose Empfehlung der experimentellen Arzneimittel häufiger bei industriefinanzierten MAs als bei Cochrane MAs vorkam, obwohl die MA-Paare ähnliche Effektschätzungen aufwiesen. Zudem schnitten die Cochrane MAs bei nach einer validierten Skala bemessener, methodischer Qualität besser ab als die industriefinanzierten MAs [Jørgensen, 2006]. Eine weitere SR fand bei 71 MAs, die Antihypertensiva bei nicht schwangeren Erwachsensen evaluierten, ebenfalls heraus, dass die Empfehlung der experimentellen Intervention häufiger bei industriefinanzierten MAs als bei Cochrane MAs war, obwohl beide ähnliche Effektschätzungen und methodische Qualität aufwiesen [Yank, 2005]. Während eine von Merck finanzierte MA kein erhöhtes kardiovaskuläres Risiko für Rofecoxib fand [Konstam, 2001], zeigte eine industrieunabhängige MA, die auch für die Autoren der Merck-MA zugängliche Primärstudien verwendete, erhöhte kardiovaskuläre Risiken [Jüni, 2004]. Die Empfehlung der experimentellen Arzneimittel war 5-fach häufiger bei durch „For-Profit-Organisationen“ als durch „Non-Profit-Organisationen“ finanzierten RCTs [Als-Nielsen, 2003b]. Keine von 56 industriefinanzierten RCTs zu NSAIDs berichtete ein für das Firmenprodukt ungünstiges Ergebnis [Rochon, 1994].

Eine SR, die 30 empirische Studien einschloss, fand heraus, dass von der pharmazeutischen Industrie finanzierte klinische Studien und MAs vierfach eher Ergebnisse zugunsten der Sponsorenintervention zeigten als von anderen finanzierte klinische Studien und MAs. Als mögliche Gründe für diese Verzerrung wurden unangemessene Kontrollintervention und Publication-Bias, nicht jedoch niedrige methodische Qualität der industriefinanzierten Studien, beobachtet [Lexchin, 2003]. Eine weitere SR, die 37 empirische Studien mit Überschneidungen zur Lexchins Review einschloss, fand ebenfalls einen statistisch signifikanten Zusammenhang zwischen Industriefinanzierung und Pro-Sponsor-Ergebnissen [OR= 3.6 (95%-KI: 2.6 – 4.6)]. Außerdem wurde ein Zusammenhang zwischen Industriefinanzierung und Einschränkungen der Publikation und Datenbeteiligung beobachtet [Bekelman, 2003].

1.1.5. Einfluss von Meta-Analyse

Der Einsatz von SRs und MAs ist nicht nur auf die Gesundheitswissenschaften beschränkt, sondern erstreckt sich von der Astrologie bis zur Zoologie [Petticrew, 2001]. Entscheidungen im klinischen Alltag sowie im Gesundheitssystem stützen sich zunehmend auf die Totalität

best verfügbarer Evidenz. SRs und MAs nehmen vermehrt eine Rolle bei der Entscheidungsfindung sowohl in der Praxis der Gesundheitsversorgung (z.B. klinische Leitlinien) als auch in der Gesundheitspolitik (z.B. Health Technology Assessment „HTA“) ein. Staatliche Investitionen zur Generierung von SRs als Entscheidungsgrundlagen sind beträchtlich [Atkins, 2005]. SRs und MAs von RCTs guter Qualität stehen bei den meisten Evidenzhierarchien an erster Stelle. Eine Analyse von 433 HTA-Berichten aus 9 Ländern (1989-2002) ergab, dass unsystematische und systematische Reviews der am häufigsten verwendete Studientyp zur Generierung von HTA-Berichten waren [Draborg, 2005].

Eine empirische Arbeit, die 2 646 Studienberichte umfasste, fand heraus, dass MA sowohl 1991 ($p < 0,05$) als auch 2001 ($p < 0,001$) eine höhere Zitierungsrate als alle anderen Studiendesigns aufwies [Patsopoulos, 2005]. Dies galt für die Zitierungsrate in den ersten zwei Jahren nach der Publikation und für längere Zeiträume. Die Adjustierung nach dem Publikationsjahr, dem sogenannten „High Journal Impact Factor“ und dem Quellstaat hob die Signifikanz dieser Ergebnisse nicht auf. MAs mit mehr als zehn Zitationen zwei Jahre nach der Publikation machten 32,4% aller MAs im Jahr 1991 und 43,6% aller MAs 2001 aus. An zweiter Stelle in der Stichprobe lagen RCTs, gefolgt von Kohortenstudien, Fall-Kontrollstudien, Fallberichten, unsystematischen Reviews und Entscheidungsanalysen bzw. Kosten-Wirksamkeits-Analysen [Patsopoulos, 2005]. Mehrere empirische Untersuchungen zeigten einen starken Impact von abgeschlossenen RCTs auf die Versorgungspraxis [Mamdani, 2001; Tu, 1998; Boissel, 1989] und eine empirische Untersuchung fand eine erhöhte Rate der Apherese (Blutreinigungsverfahren) in der Praxis, während 3 RCTs dazu liefen [Clark, 2003].

Allerdings fand eine SR zum ärztlichen Verhalten bei der Informationssuche, die 19 beobachtende Studien einschloss, dass nach Selbstangaben die von Ärzten am häufigsten konsultierten Quellen Lehrbücher waren, gefolgt von Kollegenratschlägen, die oft nicht SR-gestützt waren [Dawes, 2003]. Ähnliche Ergebnisse wurden auch bei Pflegekräften beobachtet [Olade, 2004; Kajermo, 2001]. Empirische Untersuchungen zu selektierten Entscheidungen auf Gesundheitssystemebene in Kanada [Lavis, 2002] und bei der Weltgesundheitsorganisation [Oxman, 2007] zeigten, dass SRs dazu nicht häufig benutzt wurden.

1.1.6. Epidemiologie von Meta-Analyse

In den Gesundheitswissenschaften wird die Zahl der SRs und MAs unterschätzt, da bis vor kurzem von vielen Forschern Schwierigkeiten bei ihrer Lokalisierung beklagt wurden [Hunt, 1997; Shojania, 2001]. Anhand der Handsuchen von 161 [Montori, 2005a] bzw. 55 [Wilczynski, 2007] Fachzeitschriften validierte das „Hedges Team" Suchstrategien mit hoher Sensitivität zur Identifizierung von SRs in MEDLINE bzw. EMBASE. Durch eine mehr spezifische als sensitive elektronische Suche in MEDLINE zwischen 1980 und 2000 wurden 3025 MAs identifiziert. Dabei wurde ein linearer Anstieg der Anzahl publizierter MAs festgestellt [Lee, 2001]. Im Zeitraum 1973-1998 wurde eine signifikante Erhöhung der Zahl von MAs in MEDLINE und EMBASE beobachtet [Ceballos, 2000].

Die 1993 etablierte „Cochrane Collaboration" gilt als das weltweit größte Netzwerk für die Herstellung von SRs mit etwa 15 000 Mitarbeitern aus 100 Staaten [Clarke, 2007c]. Im zweiten Quartal 2007 beinhaltete die „Cochrane Library" 4 801 und die „Database of Abstracts of Reviews of Effects" 6 113 abgeschlossene Reviews und Protokolle für Reviews, wobei eine stetig wachsende Anzahl von SRs über die letzten Jahre zu beobachten ist [Grimshaw, 2004]. Es wurde geschätzt, dass andere Einrichtungen und Forscher drei- bis fünfmal so viele SRs durchführen wie die Cochrane Collaboration [Montori, 2003]. Eine aktuelle Querschnittstudie geht von etwa 2 500 englischsprachigen SRs aus, die jährlich in MEDLINE publiziert werden. Etwa ein Fünftel von denen sind Cochrane SRs und über die Hälfte kombinieren die Ergebnisse eingeschlossener Primärstudien statistisch, d.h. meta-analytisch [Moher, 2007]. Allerdings ist der Anteil von SRs in der medizinischen Literatur weiterhin gering. Er beträgt nach den Arbeiten des „Hedges Teams" etwa 1,5% [Montori, 2005a], was bei der Synthese primärer Forschung den Erwartungen entspricht. Zu bemängeln ist der zu beobachtende Umstand, dass die bisherigen SRs ein Spektrum von Erkrankungen abdecken, das nicht der anhand der DALYs („Disability Adjusted Life Years") geschätzten globalen Krankheitslast entspricht. Unter den 10 Konditionen mit dem höchsten globalen Public Health Impact wurden Mangelernährung, Verwundungen und Infektionserkrankungen durch SRs am wenigsten untersucht [Swingler, 2003]. Vorläufige Ergebnisse einer Befragung der Mitglieder der „Deutschen Diabetes Gesellschaft" ergaben dagegen, dass die bestehenden Cochrane SRs zu Interventionen gegen Diabetes Mellitus Typ 2 der von ihr empfundenen Versorgungsrelevanz und dem Evidenzbedarf entsprachen [Kamprath, 2007].

Der Gegenstand der Mehrheit von SRs sind Primärstudien zu therapeutischen Interventionen und insbesondere zu pharmakologischen Interventionen. SRs zur Diagnostik, Prognose und Epidemiologie sind weitaus seltener [Moher, 2007]. Die meisten SRs [Moher, 2007] und RCTs [Chan, 2005] werden in spezialisierten Fachzeitschriften präsentiert.

Weiterhin ist eine weitaus geringere Zahl von SRs und MAs zu unerwünschten Interventionsereignissen zu finden. Eine SR der SRs zu therapeutischen Interventionen zeigt, dass die Sicherheit therapeutischer Interventionen zwischen 1986 und 2000 nur bei unveränderten 3-5% der SRs Primärendpunkt war und bei ansteigenden 10-27% der SRs als Sekundärendpunkt fungiert [Ernst, 2001]. Als Haupteinschränkungen bei der Abfragung unerwünschter Ereignisse in SRs [Golder, 2006b], in klinischen Studien [Derry, 2001] sowie in beobachtenden Studien [Wieland, 2005] wurden heterogene Terminologien und Unterverschlagwortung identifiziert.

1.1.7. Funktionen von Meta-Analyse

MA ist eine standardisierte Methode zur gewichteten Zusammenfassung der Ergebnisse vorhandener Primärstudien über eine Fragestellung. Dadurch kann eine zusammengefasste Effektgröße, bezogen auf alle Primärstudien, geschätzt werden. Im Rahmen dieser Arbeit wird dies die „synthetische Funktion“ von MA genannt. MA bietet jedoch auch die Möglichkeit, die methodischen Variationen zwischen den Primärstudien und die Interventionseffekte in Patientensubgruppen zu untersuchen. Dies wird als die „analytische Funktion“ von MA bezeichnet. Heterogenität spielt bei beiden Funktionen von MA eine Schlüsselrolle. In ihrer analytischen Funktion untersucht MA die Ursachen für die Heterogenität zwischen den Primärstudien. Auf der einen Seite erleichtert der Einschluss von relativ homogenen Primärstudien die synthetische Funktion von MA; auf der anderen Seite kann große Heterogenität die Synthese erschweren oder sogar unzulässig machen.

1.1.8. Kritische Bewertung von Meta-Analyse

Da unter einflussreichen Gesundheitswissenschaftlern weiter eine Einstellung gegen als Sekundärforschung betrachtete SRs und MAs besteht [Alderson, 2003] und da Reviewer den hohen Anspruch erheben, die Evidenz in ihrer Totalität zusammenzufassen und deren Validität zu analysieren, sollen SRs und MAs, was sie von Primärstudien verlangen, nämlich

Relevanz der Fragestellung, Repräsentativität der Teilnehmer (Patienten versus Primärstudien) und hohe methodische Qualität zu gewährleisten, selbst anstreben. Dies ist besonders zu betonen angesichts des relativ hohen Impacts von MA auf die Entscheidungsfindung in der Versorgungspraxis (Evidenz-basierte Medizin) und im Versorgungssystem (Health Technology Assessment) auf der einen Seite und ihrer bislang unterentwickelten, teilweise fehlenden sowie nicht harmonisierten Standardisierung und Qualitätssicherung, im Vergleich zu Genehmigungsverfahren und Leitlinien zur „Guten Klinischen Praxis“ bei RCTs, auf der anderen Seite [Schulman, 2005; Grimes, 2005; Sander, 2006].

Obwohl es seit Mulrow, der zum ersten Mal die mangelnde Qualitätsbewertung von SRs thematisierte und empirisch untersuchte [Mulrow, 1987], und seit der Entwicklung des ersten bekannten Instruments zur Bewertung der methodischen Qualität von SRs bereits zwei Jahrzehnte vergangen sind [Sacks, 1987], wurden diese Instrumente bislang selten angewendet. Leider wurden die meisten SRs bisher in spezialisierten Fachzeitschriften veröffentlicht [Moher, 2007], wo die Einhaltung von Leitlinien zur Qualität und Berichterstattung oft nicht zur Herausgabepolitik gehört. Shea und Kollegen identifizierten 24 Instrumente zur Bewertung methodischer Qualität von SRs bzw. MAs, von denen die meisten allerdings eine große Zahl von Items aufwiesen und sich keiner Validierung unterzogen [Shea, 2001]. Allerdings stellt das „Quality of Reporting of Meta-analyses“ (QUOROM)-Statement eine mit Konsens entwickelte Leitlinie zur Berichterstattung der MAs dar. Es schließt 18 Items und ein Flussdiagramm ein [Moher, 1999a], wird zurzeit überarbeitet und in PRISMA („Preferred Reporting Items for Systematic Reviews and Meta-Analyses“) umbenannt. Das „Overview Quality Assessment Questionnaire“ (OQAQ,) ist eine validierte Checklist mit 10 Items zur Bewertung der Qualität von SRs [Oxman, 1988; Oxman, 1991]. Das „Assessment of Multiple Systematic Reviews“ (AMSTAR) erstellte eine validierte Checkliste, die aus 11 Items besteht, welche durch eine objektive Selektion und Zusammenfügung von Items aus der OQAQ, der Checkliste von Sacks [Sacks, 1987] und drei weiteren Komponenten hergeleitet wurden [Shea, 2007].

Die „European Medicines Agency“ verfasste strenge Leitlinien, nach denen MAs für die Zulassungsverfahren für Arzneimittel herangezogen werden können, und erkannte an, dass MAs eingesetzt werden können zur Erreichung präziser Schätzung des Effekts in der Gesamtpopulation, zur Überprüfung des Effekts in prä-spezifizierten Subgruppen der Patienten, zur Untersuchung heterogener Effektschätzer der Primärstudien und zur Power-Erhöhung für die Evaluation von zusätzlichen Wirksamkeitsendpunkten, seltenen

unerwünschten Ereignissen in der Gesamtpopulation oder Sicherheitsendpunkten in Subgruppen der Patienten [EMEA, 2001].

Wie in Abschnitt 1.1.4.4 ausgeführt wurde, endeten SRs und MAs zu denselben Fragestellungen aus vermuteten Interessenkonflikten in unterschiedlichen Schlussfolgerungen. Widersprüchliche Konklusionen von SRs waren allerdings auch auf deren methodische Unterschiede zurückzuführen. Eine SR identifizierte zehn SRs zu Acetylzystein für die Prävention von Kontrastmittel-assoziierter Nephropathie mit unterschiedlicher methodischer Qualität und verschiedenen Empfehlungen [Biondi-Zoccai, 2006]. Eine SR der SRs zu chronischen Kreuzschmerzen fand widersprüchliche Fazite bei 11 von 13 Interventionen, für die zwei oder mehr SRs existierten [Furlan, 2001]. Ähnliches war zu beobachten bei drei SRs zur epiduralen Steroidinjektion für Ischias [Hopayian, 2001], bei zwei MAs zu Antibiotika gegen akutes Husten [Lindbaek, 1999], bei 23 MAs zur chirurgischen Thromboprophylaxe [Petticrew, 1997] und bei vier MAs zu Aminoglykosiden (Antibiotika) [Prins, 1996].

Im Gegensatz zu Online publizierten SRs, wie den Cochrane SRs, litten die in Fachzeitschriften ohne Online-Zusatzmaterial veröffentlichten SRs oft unter Unterberichterstattung. Eine empirische Studie zur Interventionsbeschreibung in SRs, die in der Fachzeitschrift „EBM“ erschienen war, fand heraus, dass in weniger als 15% der SRs die Beschreibung als ausreichend zu klassifizieren war [Glasziou, 2007].

1.1.9. Durchführung von Meta-Analysen

Die internationale „Cochrane Collaboration“ [Higgins, 2006], die „Agency for Healthcare Research and Quality“ (AHRQ) der USA [West, 2002], das britische „Centre for Reviews and Dissemination“ (CRD) [Khan, 2001] und die „Potsdam Consultation on Meta-Analysis“ [Cook, 1995] stellen strukturierte Leitlinien zur Planung und Durchführung von SRs und MAs bereit. In Anlehnung an diese Leitlinien, weitere Quellen und eigene Überlegungen ergeben sich die in den folgenden Abschnitten dargestellten Anforderungen an die Methodik von SRs und MAs.

1.1.9.1. Entwicklung der Fragestellung

Eine suchtaugliche Fragestellung setzt die präzise Definition der sogenannten PICO-D-Komponenten voraus (s. Tab. 1). Durch die Festlegung der PICO-D-Komponenten werden die Ein- und Ausschlusskriterien für Primärstudien bestimmt. Beispiele für eine breite und eine engere Fragestellung sind in Tab. 1 dargestellt. Vorkenntnisse über vorhandene Primärstudien können zur bewussten oder unbewussten Manipulation der Ein- und Ausschlusskriterien für deren Auswahl führen. Haupteinschränkungen bei Fragestellungen zu unerwünschten Ereignissen sind, dass sie oft nicht vor der Durchführung von SR bzw. MA bekannt und in Primärstudien nicht einheitlich definiert und klassifiziert sind. Daher sind die Endpunkte bei Fragestellungen zu Interventionsrisiken breit zu definieren.

Tab. 1 Das PICO-D-Schema zur Entwicklung von Fragestellungen an zwei Beispielen

Patienten (**P**atients)	Intervention (**I**ntervention)	Komparator (**C**omparison)	Endpunkt (**O**utcome)	Design (**D**esign)
Herzinsuffizienz	ACE-Hemmer	Beta-Blocker	LV-Funktion	RCT
Hypertonus	Losartan	Atenolol	Gesamtmortalität	RCT

1.1.9.2. Anfertigung des Protokolls

Das Protokoll der SR bzw. MA soll nach einer Festlegung des Bedarfs für eine SR, einer Recherche des Hintergrunds und der Entwicklung der Fragestellung erfolgen, d.h. es soll vor der Durchführung des Suchplans entwickelt werden. Die Reviewgruppe schließt im Idealfall Experten aus der Gesundheitssubdisziplin, der klinischen Epidemiologie/ EbM/ HTA, der medizinischen Biostatistik und der medizinischen Dokumentation mit ein. Die folgende Aufzählung enthält die wesentlichen Aspekte, die ein Protokoll enthalten und so ausführlich wie möglich festlegen soll.

- Haupt- und Nebenfragestellungen, einschließlich Primär- und Sekundärendpunkte
- Suchplan
- Kriterien und Verfahren zur Selektion von Primärstudien
- Verfahren zur Datenextraktion
- Kriterien und Verfahren zur Bewertung der Qualität von Primärstudien
- Synthese- und Auswertungsplan
- Mitglieder der Reviewgruppe und Kontaktperson

Während bei fast jeder Cochrane SR ein Protokoll angefertigt und online veröffentlicht und zur Diskussion gestellt wurde, findet dies bei anderen SRs nur bei einer Minderheit statt [Moher, 2007]. Allerdings fand ein Abgleich von 47 Cochrane SRs mit ihren Protokollen eine Hauptabweichung vom Protokoll bei 91% der SRs. Dies geschah am häufigsten im Methodikteil, wo das Biaspotenzial am größten ist [Silagy, 2002]. Die Publikation von Protokollen soll zur Vermeidung des sogenannten “HARKing: Hypothesizing After the Results are Known“ beitragen [Kerr, 1998, S. 196].

Die Forderung nach Registrierung von SRs und MAs ist umstritten. Empirische Untersuchungen deuten darauf hin, dass SRs und MAs weniger von Publication-Bias betroffen sind als RCTs [Tierney, 2000]. Eine Analyse aller 300 SRs, die im November 2004 in MEDLINE verzeichnet waren, fand keine einzige SR, die sich einer Registrierung unterzog [Moher, 2007].

1.1.9.3. Suchen nach Primärstudien

Die Qualität des Suchplans nach Primärstudien bestimmt die Anfälligkeit einer SR bzw. einer MA für Selection-Bias. Es soll immer mehr als eine elektronische Datenbank durchsucht werden, wobei die Suchstrategien den Aufbau und die Verschlagwortung verschiedener Datenbanken (z.B. MEDLINE, EMBASE) berücksichtigen sollen. Das „Cochrane Central Register of Controlled Clinical Trials“ (CENTRAL) stellt die umfassendste Datenbank für RCTs und NCTs dar, da die Cochrane Collaboration elektronische Recherche mit Handsuchen von über 1 700 Fachzeitschriften kombiniert [Dickersin, 2002]. Komponenten der Fragestellung bestimmen, ob mehr spezialisierte Datenbanken mit einzubeziehen sind (z.B. ASSIA „Applied Social Sciences Index and Abstracts“, CINAHL „Cumulative Index for Nursing and Allied Health Literature“, ERIC „Educational Resources Information Center“, HMIC „Health Management Information Consortium“, PsycInfo „Databank of the American Psychological Association“). Das Hedges Team entwickelte und validierte Suchstrategien für RCTs in MEDLINE [Haynes, 2005]. Es kann auf evaluierte Suchansätze für unerwünschte Interventionsereignisse in MEDLINE [Badgett, 1999], in EMBASE [Loke, 2002] und sowohl in MEDLINE als auch in EMBASE [Golder, 2006c] zurückgegriffen werden. 51 „Cochrane Review Groups“ bieten detaillierte und krankheitsspezifische Suchalogarithmen [http://www.cochrane.org/contact/entities.htm]. Mehrere Suchalogarithmen können angewendet und verglichen werden.

Es sollen mehr sensitive als spezifische Suchen angestrebt werden, wobei zu beachten ist, dass Schlagwörter in Datenbanken (z.B. MESH „Medical Subject Headings“ in MEDLINE und EMTREE in EMBASE) relativ spezifisch sind. Obwohl die „Cochrane Highly Sensitive Search Strategy“ (HSSS) einen relativ komplexen Ansatz darstellt, weist sie eine hohe Sensitivität auf, englischsprachige RCTs und NRTs zu identifizieren, die in Volltext publiziert und in Datenbanken verzeichnet sind [Hopewell, 2007a]. In der Regel soll jedoch auch nach nicht in englischer Sprache publizierten Primärstudien gesucht werden.

Vorherige SRs, MAs, HTAs und die Literaturverzeichnisse eingeschlossener Primärstudien sowie Kontakte mit Contentexperten und Herstellern gelten als relativ effiziente Quellen. Eine SR, die 34 empirische Vergleiche von elektronischen versus Handsuchen identifizierte, empfahl aufwändige Handrecherchen für die Suche nach RCTs, die nicht in Volltext, in nicht englischsprachigen oder nicht in Datenbanken indexierten Fachzeitschriften publiziert sind [Hopewell, 2007a]. Die Suche nach unpublizierten Primärstudien in Dissertationen, Konferenzberichten, Diskussionspapieren (sogenannte Grauliteratur) ist kostenintensiv und soll mit dem zu erwartenden Nutzen abgewogen werden. Die Recherche in dafür spezialisierten Datenbanken, z.B. in „System for Information on Grey Literature“ (SIGLE), kann effizient sein. Bei niedrigem Datenumfang bzw. schlechter Qualität publizierter Studien kann Grauliteratur nützlich sein. Aufgrund nicht ausreichender Berichterstattung in Abstracts und Präsentationen, insbesondere zur methodischen Qualität, sollen Zusatzangaben möglichst von den kontaktierbaren Prüfärzten erbeten werden. Es ist geboten, Graustudien als solche zu markieren und mittels Sensitivitäts-Analyse ihren Einfluss auf die Gesamtergebnisse der SR bzw. der MA zu untersuchen [Dundar, 2006]. Suchen nach laufenden Primärstudien, insbesondere in Registern wie „Current Controlled Clinical Trials“, „ClinicalTrials.gov“ und „CenterWatch", sollen in der Regel durchgeführt werden; dennoch können Interimergebnisse nur unter Vorbehalt einbezogen werden [Song, 2004].

Bei jeder SR oder MA sollen alle Suchstrategien dokumentiert und beschrieben werden, einschließlich der Suchquellen, der Suchjahre, der Suchbegriffe und der Sucheinschränkungen. Dies gilt auch für Adaptierungen, nicht vorbestimmte Änderungen und Aktualisierungen der Suchstrategie.

1.1.9.4. Selektion von Primärstudien

Primärstudien werden nach den vordefinierten Ein- und Ausschlusskriterien selektiert. Alle Titel und Zusammenfassungen der Publikationen sollen gesichtet werden, im Zweifel soll der

Volltext einbezogen werden. Die Studienauswahl soll von mindestens zwei Reviewern, die unabhängig voneinander arbeiten, erfolgen. Unterschiede zwischen den Reviewern sollen durch Konsensverfahren, Votum eines dritten Reviewers oder Sensitivitäts-Analysen abgehandelt werden. Es ist geboten, über das Maß für die Inter-Rater-Reliabilität, z.B. Kappa-Koeffizient, zu berichten. Zahl und Gründe der Ausschlüsse von Primärstudien sollen dokumentiert, über sie soll berichtet und eine Auflistung ausgeschlossener Primärstudien zugänglich gemacht werden. Die Verblindung von Reviewern bei der Auswahl von Primärstudien ist umstritten. Es bestehen Hinweise darauf, dass der Aufwand, der damit verbunden ist, dem Nutzen nicht gerecht wird [Berlin, 1997; Moher, 1999b].

1.1.9.5. Extraktion von Daten

A priori soll ein Bogen zur Datenextraktion entwickelt und gegebenenfalls an einigen Primärstudien erprobt und modifiziert werden. Ein von Externen erprobter Extraktionsbogen, z.B. „Cochrane Review Groups“, ist zu bevorzugen. Wie bei der Studienselektion sollen die Daten von mindestens zwei Reviewern, die unabhängig voneinander arbeiten, extrahiert werden. Unterschiede zwischen den Reviewern sollen durch ein Konsensverfahren oder das Kontaktieren der Autoren von Primärstudien abgehandelt werden. Hier ist es auch geboten, über das Maß für die Inter-Rater-Reliabilität, z.B. Kappa-Koeffizient, zu berichten. Die Verblindung von Reviewern bei der Auswahl von Primärstudien ist umstritten. Es bestehen Hinweise darauf, dass der Aufwand, der damit verbunden ist, dem Nutzen nicht gerecht wird [Berlin, 1997; Moher, 1999b]. Fehler bei der Extraktion oder der Kalkulation von Effektmaßen sollen seitens der Reviewer vermieden werden und durch Nutzer von SRs und MAs möglichst überprüfbar bleiben. Eine SR, die 27 MAs mit standardisierter Mittelwertdifferenz (SMD) als Effektgröße einschloss, fand heraus, dass bei Wiederberechnung der SMD in 63% der MAs mindestens eine von zwei zufällig aus jeder MA selektierten kontrollierten Studien Fehler aufwies, so dass in 37% der MAs eine Abweichung von $\geq 0{,}1$ beim Punktschätzer oder dem Konfidenz-Intervall in mindestens einer Studie festgestellt wurde. Die Reextraktion der SMD aus allen klinischen Studien in den 10 MAs mit einer Abweichung von $\geq 0{,}1$ in mindestens einer Studie und die Wiederholung der MAs nach der Methode der Originalautoren ergab, dass 7 MAs fehlerhaft waren, von denen eine MA die statistische Signifikanz verlor, eine sie gewann und eine weitere nachträglich retraktiert wurde [Gøtzsche, 2007].

Häufige Einschränkungen der Datenextraktion stellen fehlende Angaben oder Fehler in den Angaben der Primärstudien dar, was zur Verzerrung der Ergebnisse von SR bzw. MA führen

kann. Die statistische Power und die Validität der MA sind dann gefährdet. Eine SR zu Antidepressiva fand heraus, dass nur 9 aus 69 RCTs Standardabweichungen für Endpunkte berichteten [Streiner, 1998]. Eine SR zu selektiven Serotonin-Wiederaufnahmehemmern fand heraus, dass nur 20 von 53 RCTs über Standardabweichungen für Endpunkte berichteten [Song, 1993]. Eine SR zu Methoden des Umgangs mit fehlenden Angaben zur Varianz in Primärstudien bei der Durchführung von MA empfahl die Verwendung algebraischer Rekalkulation, Autorenkontakt, multipler Imputation und Sensitivitäts-Analyse [Wiebe, 2006]. In der Regel sind Primärstudien, die nur als Abstracts veröffentlicht wurden, in SRs einzubeziehen. Allerdings fand ein Abgleich der Abstracts und Vollpublikationen von 37 RCTs in der klinischen Onkologie große Diskrepanzen zwischen den beiden, was in den meisten klinischen Studien auf vorläufige Ergebnisse und mangelnde Angaben zur methodischen Qualität in Abstracts zurückzuführen war [Hopewell, 2006]. Eine bibliographische Untersuchung in MEDLINE fand bei 1,2% der RCTs mit der jeweiligen Publikation verlinkte Errata. Dieser Errataanteil war höher als der bei anderen Publikationstypen. Zudem wurden, in einer Stichprobe von 100 RCTs, 5% der Errata als die Ergebnisse von MAs beeinflussend eingestuft [Royle, 2004]. Publikationen über alte RCTs begnügen sich häufig mit den p-Werten oder sogar mit der Erreichung oder Nichterreichung eines konventionellen Signifikanzniveaus. Die Imputation von Daten aus Graphiken, insbesondere aus klein gedruckten Kaplan-Meier-Kurven bei Überlebensanalysen, sollen durch mehr als einen Ansatz erfolgen und mittels Sensitivitäts-Analysen geprüft werden. Daten aus Text, Tabellen und Graphiken sollen auf Konsistenz geprüft werden.

Das konsensbasierte CONSORT-Statement wurde 1996 bekannt gemacht [Begg, 1996], 2001 überarbeitet [Altman, 2001] und von 175 Fachzeitschriften, Redaktionsgruppen (“Council of Science Editors”, “World Association of Medical Editors”, “International Committee of Medical Journal Editors”) sowie Sponsoren (“Canadian Institutes of Health Research”) unterstützt. Es beinhaltet 22 Items und ein Flussdiagramm und bietet einen Leitfaden für verbesserte Berichterstattung von Parallelgruppen-RCTs [Moher, 2001]. Das CONSORT-Statement wurde durch weitere 10 Items zu Interventionsrisiken ergänzt [Ioannidis, 2004], für Cluster-RCTs [Campbell, 2004] und für Noninferiorität-/ Equivalenz-RCTs [Piaggio, 2006] erweitert. Allerdings fehlen bislang ähnliche Leitlinien für Crossover- und Factorial-RCTs.

1.1.9.6. Qualitätsbewertung von Primärstudien

Die methodische Qualität von RCTs ist ein komplexes Konstrukt, das nicht nur die Angemessenheit des Studiendesigns und der Studien-Analyse einschließlich der Risiken für verschiedene Biasarten beinhaltet, sondern auch die Angemessenheit der Patientenkollektive, der experimentellen Intervention, der Kontrollintervention und der Endpunkte, d.h. aller Komponenten des PICO-Schemas. Der Einschluss von Primärstudien mit niedriger methodischer Qualität in einer MA, auch wenn keine bessere Evidenz besteht, kann zur Entwertung ihrer Schlussfolgerungen führen (garbage in, garbage out). Dieses kann sich in einer Verzerrung der Punktschätzer, der oberen und unteren Grenzen des Konfidenz-Intervalls der zusammengefassten Effektgrößen widerspiegeln.

Eine Fülle von einzelnen Dimensionen, Checklisten und Scores zur Bewertung der methodischen Qualität von Primärstudien liegt vor, aber es besteht kein Konsens über ein einziges, einheitliches Instrument [Moher, 1995; Sutton, 1998]. Da die Subjektivität bei der Abschätzung methodischer Qualität nicht vermeidbar, wohl aber reduzierbar ist, sollen möglichst validierte Instrumente und transparente Kriterien zur Bewertung der Biasrisiken durch mindestens zwei Reviewer, die unabhängig voneinander arbeiten, an Primärstudien angewendet werden. Unterschiede zwischen den Reviewern sollen durch Konsensverfahren, Votum eines dritten Reviewers oder Sensitvitäts-Analysen abgehandelt werden. Es ist geboten, über das Maß für die Inter-Rater-Reliabilität, z.B. Kappa-Koeffizient, zu berichten. Die methodische Qualität der Primärstudien soll bei der Synthese der Evidenz, der Interpretation der Ergebnisse und der Formulierung der Schlussfolgerungen berücksichtigt werden. Die Verblindung von Reviewern bei der Auswahl von Primärstudien ist umstritten und es bestehen Hinweise darauf, dass der Aufwand, der damit verbunden ist, dem Nutzen nicht entspricht [Berlin, 1997; Moher, 1999b]. Die Bewertung der Qualität von RCTs in MA wird ausführlich in Abschnitt 1.3 behandelt.

1.1.9.7. Synthese und Analyse von Primärstudien

Es sollen allgemeine Merkmale der Primärstudien beschrieben und eine deskriptive Statistik dazu dargestellt werden. Die Berücksichtigung von Heterogenität bei der Evidenz-Synthese und die Untersuchung von Heterogenitätsursachen zwischen den Primärstudien sind zentrale Aufgaben von SRs bzw. MAs, die in Abschnitt 1.2 ausführlich behandelt werden. Die Entscheidung, extrahierte Evidenz qualitativ durch narrative Beschreibung oder quantitativ

durch meta-analytische Kombination zusammenzufassen, hängt mit der Heterogenität zwischen den Primärstudien zusammen.

1.1.9.8. Berichterstattung, Dissemination und Aktualisierung

Nicht nur der Bericht abgeschlossener SRs bzw. MAs soll disseminiert werden, sondern auch das davor angefertigte Protokoll. Alle vorher aufgeführten Schritte der SR/ MA sollen detailliert berichtet und übersichtlich dargestellt werden, insbesondere durch Evidenztabellen, ausführliche Anhänge und Online-Zusatzmaterialien. Stärken und Einschränkungen der SR/ MA sollen aufgezeigt werden. Finanzierung der SR/ MA und Interessenkonflikte der Reviewer sollen transparent deklariert werden. Leitlinien zur Formulierung von Empfehlungen unter transparenter Berücksichtigung der Stärke und der Qualität vorhandener Evidenz - wie die von der „GRADE Working Group“ [Atkins, 2004; Atkins, 2005] oder der „U.S. Preventive Services Task Force“ [Guirguis-Blake, 2007; Barton, 2007] - sollen befolgt, Überinterpretation der Ergebnisse, insbesondere von Subgruppen-Analysen, vermieden [Pocock, 2002] und Einschränkungen der Evidenz diskutiert [Ioannidis, 2007b] werden. Konsequenzen für die Praxis und die weitere Forschung sollen skizziert werden. Die Verbreitung der SR/ MA-Ergebnisse soll auch erfolgen, wenn keine Primärstudie gefunden werden kann [Alderson, 2000; Petticrew, 2003]. Schätzungen zufolge bezeichneten mehr als die Hälfte der abgeschlossenen Cochrane SRs die Evidenzlage einer Intervention als eingeschränkt oder dünn [Glasziou, persönliche Kommunikation; nach: Laupacis, 2007]. 20-sekündige, 2-minütige und 2-stündige Versionen des SR/ MA-Berichts wurden für unterschiedliche Zeitressourcen professioneller Konsumenten vorgeschlagen [Laupacis, 2007]. Das DISCERN-Projekt bietet einen Leitfaden für die gebotene Zusammenfassung der Ergebnisse im patientenfreundlichen Format [www.discern.org.uk].

Weil SR/ MA ein Instrument zur Zusammenfassung eines sich oft verändernden Forschungsstandes ist, sind sie auch aktualisierungsbedürftig. Bislang existiert keine allgemeine Regel für die Aktualisierung von SRs, da ihre Aktualisierung von der Entwicklungsgeschwindigkeit der Primärstudien abhängt; letztere variiert nach der jeweiligen Fragestellung bzw. der jeweiligen Fachgebiet. Abgelaufene SRs/ MAs können für Entscheidungen in der klinischen Praxis und im Gesundheitssystem irreführend sein, Updates ohne Zusatznutzen sind als Ressourcenverschwendung zu betrachten [Moher, 2006]. Der Grundsatz der „Cochrane Collaboration“ verpflichtet ihre Reviewer zur zweijährigen Aktualisierung ihrer Reviews bei Bedarf oder zur Kommentierung bei einer Aufschiebung [Higgins, 2006]. Nur 17,7% aller 300 SRs, die im November 2004 in MEDLINE

verzeichnet waren, bezeichneten sich als Updates für vorherige SRs. Dies galt mit Abstand öfter für die Cochrane SRs (37,6%) als für die Nicht-Cochrane SRs (2,3%) [Moher, 2007]. 13,5% der 481 Cochrane SRs, die 1998 in Ausgabe 4 eingeschlossen waren, wiesen mindestens eine zusätzliche Primärstudie auf. Nur in 8% dieser Updates war eine Änderung der statistischen Signifikanz festzustellen [Higgins, 1999]. 70% der 362 im Jahr 2002 in Ausgabe 2 der „Cochrane Library“ eingeschlossenen SRs waren Updates. In nur 9% der SRs erfolgte nach der Aktualisierung eine „Hauptänderung in der Schlussfolgerung“ [French, 2005]. Eine weitere Untersuchung fand Updates für 44% bzw. 58% der SRs aus zwei Organisationen innerhalb von drei Jahren, wobei auch hier wenige Updates in Änderung der Schlussfolgerung resultierten [Chapman, 2002]. Eine SR zur aktualisierungrechtfertigenden Veränderung der Evidenzlage, die 100 SRs von RCTs einschloss, zeigte, dass explizite quantitative oder qualitative Signale für eine Aktualisierung bei 57% der SRs zu beobachten waren und dass der Median für die Überlebenszeit der SRs ohne Signale für Aktualisierung 5,5 Jahre betrug. Zudem war ein Signal für Aktualisierung bei 23%, 15% und 7% der SRs innerhalb von jeweils zwei Jahren, einem Jahr und zum Publikationszeitpunkt zu beobachten. Kardiovaskuläre Fragestellung und statistische Heterogenität waren signifikant mit niedrigen Überlebenszeiten der SR assoziiert [Shojania, 2007]. Aus 17 von der US „Agency for Healthcare Research and Quality“ entwickelten klinischen Leitlinien wurden Dreiviertel als aktualisierungsbedürftig klassifiziert und 50% nach 5,8 bzw. 90% nach 3,6 Jahren als veraltet bezeichnet. Die Effizienz und die Qualitätssicherung von Updates, die durch Reviewgruppen mit rotierender Mitgliedschaft anzufertigen sind, wurden relativ gut bewertet [Shekelle, 2001].

1.2. Berücksichtigung und Untersuchung von Heterogenität in Meta-Analyse

1.2.1. Heterogenitätsursachen in Meta-Analyse

Heterogenität in einer MA bezieht sich auf die Unterschiede zwischen den Effektgrößen der Primärstudien, die in die MA eingeschlossen sind. Im Rahmen dieser Dissertation wurde die Heterogenität zwischen den Primärstudien in MA auf drei Quellen zurückgeführt:

- zufallsbedingte Heterogenität, bei der die Variabilität der Effektgrößen von Primärstudien lediglich auf zufällige Unterschiede bei der Stichprobenziehung der Studienteilnehmer in den einzelnen Primärstudien zurückzuführen ist.

- methodische Heterogenität, bei der die Variabilität der Effektgrößen von Primärstudien auf Unterschiede bei der methodischen Qualität von Primärstudien zurückzuführen ist.
- klinische Heterogenität, bei der die Variabilität der Effektgrößen von Primärstudien auf Unterschiede bei den Merkmalen der Patienten oder der Interventionen zurückzuführen ist.

Methoden zum Testen und zum Schätzen von zufallsbedingter Heterogenität in MA werden in Abschnitt 1.2.4.1 ausführlich dargestellt und diskutiert. Ansätze zur Bewertung der methodischen Qualität von RCTs werden detailliert und kritisch in Abschnitt 1.3 behandelt. Methoden zur Untersuchung von methodischer und klinischer Heterogenität in MA und deren Einschränkungen werden in Abschnitt 1.2.5 erläutert. Es folgen Ausführungen über die klinische Heterogenität. Dann werden MAs mit individuellen Patientendaten MAs mit aggregierten Patientendaten gegenübergestellt.

1.2.2. Klinische Heterogenität in Meta-Analyse

Die mit der demographischen Alterung und der Technologisierung gesundheitlicher Versorgung verbundenen stetigen Anstiege von Multimorbidität und Polypharmazie stellen die Ergebnisse solcher klinischen Studien in Frage, die Patienten mit wenig Komorbiditäten und Komedikationen einschließen. Die Übertragbarkeit des aus RCTs gewonnenen durchschnittlichen Interventionseffekts auf unterschiedliche Subpopulationen kristallisiert sich oft als Herausforderung für die Gesundheitsversorger, z.B. Ärzteschaft, Pflegepersonal, heraus und birgt nicht selten für viele Patienten schwerwiegende Folgen [Heath, 2007]. Häufiger treten Entscheidungsdilemmata bei der polypharmazeutischen Behandlung von multimorbiden Patienten-kollektiven mit multikausalen chronischen Erkrankungen auf [Rosser, 1999]. Zudem kumuliert sich die Evidenz, dass manche Gesundheitsinterventionen keinen universalen Effekt auf Patienten aufweisen [Collins, 1990; Fibrinolytic Therapy Trialists', 1994; Gueyffier, 1999]. Mehrere Primärstudien, überwiegend in der Grundlagenforschung, stützten die Hypothese zu differenziellen Interventionseffekten bei Älteren [McLean, 2004], bei Kindern [Caldwell, 2004], bei Frauen [Anderson, 2005] und bei ethnischen Minderheiten [Bjornsson, 2003].

Eine zentrale Einschränkung bisheriger medizinischer Forschung bei der Untersuchung von klinischer Heterogenität ist die Tatsache, dass manche Subpopulationen aus verschiedenen Gründen von RCTs häufig ausgeschlossen bleiben. Dies trifft vor allem zu auf Kinder, Ältere,

Frauen und ethnische Minderheiten. Demzufolge bleibt die Evidenzlage bezüglich des Interventionseffekts in der Pädiatrie, Geriatrie, bei Frauen und bestimmten Ethnien dünn (s. Tab. 2, 3 und 4). Zunehmende politische und regulatorische Schritte zur vermehrten Beteiligung von Älteren [ICH, 1993], Kindern [European Commission, 2002], Frauen [IOM, 2001] und von ethnischen Minderheiten [ICH, 1998a] in der Gesundheitsforschung sind seit den 1990er Jahren zu beobachten.

Tab. 2 Ausschluss von älteren Patienten aus RCTs [Eigene Darstellung]

Autor (Jahr)	**Gegenstand (Datenquelle)**	**Ergebnisse**
Bartlett (2003)	RCTs zur Statinbehandlung (Medline, 1990-2001)	47 RCTs insgesamt 31 RCTs berichteten Altersobergrenze 11 RCTs berichteten Endpunkte nach Alterskategorien
Heiat (2002)	RCTs zur Behandlung von Herzinsuffizienz (Medline, 1985-1999)	59 RCTs insgesamt 13 RCTs berichteten über Altersverteilung 61,4 Jahre war der Durchschnitt (*SD* = 6,4) 17 RCTs berichteten Altersobergrenze

Tab. 3 Ausschluss von Frauen aus RCTs [Eigene Darstellung]

Autoren (Jahr)	**Gegenstand (Datenquelle)**	**Ergebnisse**
Bartlett (2003)	RCTs zur Statinbehandlung (Medline, 1990-2001)	47 RCTs insgesamt 8 RCTs schlossen Frauen aus 18,6% Anteil der Frauen pro RCT (Median) 14 RCTs berichteten über Wirksamkeit bei Frauen 2 RCTs berichteten über unerwünschte Ereignisse bei Frauen
Heiat (2002)	RCTs zur Behandlung von Herzinsuffizienz (Medline, 1985-1999)	59 RCTs insgesamt 57 RCTs berichteten über Geschlechterverteilung 21% aller eingeschlossenen Personen waren Frauen 4 RCTs schlossen Frauen aus

Tab. 4 Ausschluss von ethnischen Minderheiten aus RCTs [Eigene Darstellung]

Autoren (Jahr)	**Gegenstand (Datenquelle)**	**Ergebnisse**
Bartlett (2003)	RCTs zur Statinbehandlung (Medline, 1990-2001)	47 RCTs insgesamt 8 RCTs berichteten den Anteil ethnischer Minderheiten in der Stichprobe
Heiat (2002)	RCTs zur Behandlung von Herzinsuffizienz (Medline, 1985-1999)	59 RCTs insgesamt 12 RCTs berichteten über Ethnienverteilung 15% der in diese RCTs eingeschlossenen Personen waren Nicht-Weiße 0 RCTs schlossen ethnische Minderheiten aus

Nationale und internationale Organisationen wie die US „Food and Drug Administation“ (FDA), das UK „National Institute for Clinical Excellency“ (NICE), die „European Medicines Agency“ (EMEA) und die „International Conference on Harmonisation “ (ICH) haben Ansätze zu mehr Berücksichtigung von Patientenheterogenität bei der Grundlagenforschung ,z.B. Pharmakokinetik, -dynamik, -genomik, und der klinischen Forschung, z.B. RCTs, entwickelt und gefördert [Temple, 2002; Allmark, 2004]. Die Variationen eines Interventionseffekts zwischen den Patienten, die auf soziodemographische Merkmale, Komorbidität und Komedikation zurückzuführen sind, können mit heterogenen und noch nicht ausreichend entzifferten pharmakokinetischen, -dynamischen und -genomischen Wirkungsweisen assoziiert sein [Siest, 2007; Yong, 2006; Weber, 2001; Nakagawa, 2000]. Allerdings soll die klinische Heterogenität in aller Regel nicht nur in beobachtenden Primärstudien, die für Counfounding und Bias eher anfällig sind, untersucht werden, sondern in RCTs mit guter methodischer Qualität und aureichender statistischer Power.

Das Streben nach Evidenzbasierung mancher Leistungserbringer und Patienten angesichts komplexer oder seltener Risikokonstellationen führte schon vor zwei Jahrzehnten zum sogenannten „N-of-1 Trial“ als einem radikalen Studiendesign zur Berücksichtigung klinischer Heterogenität und Individualisierung von Therapien [Guyatt, 1986]. Bei einem „N-of-1 Trial“ werden mehrere Interventionen verblindet an einem einzigen Patienten randomisiert, d.h. es handelt sich dabei um multiple „cross-over“ Vergleiche. Die geringe Akzeptanz des Experiments seitens der Prüfarzte und der Patienten sowie die Eignung des Designs nur für Interventionen mit einer kurzfristig eintretenden und gering bleibenden Wirkung schränken die Durchführbarkeit solcher klinischen Studien stark ein. Zudem gilt die Extrapolierbarkeit der Ergebnisse aus diesem Studiendesign als stark eingeschränkt [Mahon, 1996; Madhok, 2005]. Dem stehen sogenannte pragmatische klinische Studien gegenüber, die seit 1967 als ein Ansatz zur Maximierung externer Validität und Alltagswirksamkeit propagiert wurden [Lellouch, 1967; nach: Armitage, 1998]. Durch die Weitfassung von Einschlusskriterien, durch die Erhebung einer begrenzten Zahl von einfach erfassten Variablen und durch die limitierte Kontrolle der Studienbedingungen und des Studienverlaufs versuchen pragmatische klinische Studien die Alltagswirksamkeit zu schätzen. Während durch ein solches Studiendesign hohe externe Validität zu erwarten ist, gilt seine interne Validität oft als gefährdet [McMahon, 2002; Godwin, 2003]. Prospektive MA bzw. multizentrische Mega-RCTs (d.h. die Fallzahl liegt bei > 1 000 bis > 10 000), die oft durch multinationale Rahmenbedingungen und Regulationen mit Hindernissen konfrontiert sind [Berlin, 1999], zielen darauf ab, methodischen Variationen durch ein möglichst einheitliches

Planungs-, Durchführungs- und Auswertungsprotokoll vorzubeugen. Gelingt dies, können Patientenvariationen bezüglich des Interventionseffekts ohne Confounding durch methodische Heterogenität untersucht werden. Beide Arten von Studien werden aber selten durchgeführt, in der Regel nur bei Gesundheitstechnologien höchster Public Health Relevanz, meistens nach mehreren kleinen RCTs mit nicht ausreichender und/oder diskordanter Evidenz. Zudem sind prospektive MAs und Mega-RCTs kosten- und managementintensiv. Zu den bisher wenigen prospektiven MAs zählten eine prospektive MA zur Wirksamkeit von Fluorouracil und Folinic-Säure in Dickdarmkrebs [IMPACT, 1995] und eine zum Nutzen von cholesterinsenkenden Statinen [Baigent, 2005].

Bei bisheriger Fallzahlkalkulation im Rahmen von RCTs wird die klinische Heterogenität selten miteinbezogen. Eine Fallzahl für eine RCT unter Berücksichtigung von Verteilung, α-Fehler, β-Fehler, Bereichen der Null- und Alternativ-Hypothesen für mehrere Subpopulationen zu berechnen, gilt als nicht trivial. Statifizierte Randomisierung erfolgt oft für eine begrenzte Zahl von Patientencharakteristika. Manche Effektmodifikatoren bleiben oft bei der Stratifizierung unbeachtet. Für Subgruppen-Analysen bei RCTs fehlt oft die notwendige statistische Power für die Erkennung vermuteter klinischer Heterogenität. Die Erhöhung der statistischen Power durch MA mit individuellen Patientendaten erleichtert die Untersuchung klinischer Heterogenitätsquellen.

1.2.3. Meta-Analyse mit individuellen versus mit aggregierten Patientendaten

Aggregierte Patientendaten (APDs) stellen die Datengrundlage für die Mehrheit der MAs dar. MAs mit aggregierten Patientendaten (MA-APDs) können aus den Publikationen von Primärstudien extrahiert oder von Prüfärzten angefordert werden. Die mit Abstand selteneren MAs mit individuellen Patientendaten (MA-IPDs) können entstehen, wenn eine Kollaboration für eine Datenbeteiligung zwischen den Studienverantwortlichen entsteht. Eine solche Kollaboration beinhaltet die zentrale Sammlung, Validierung und Re-Analyse von „rohen“ Daten aus allen RCTs zu einer Intervention.

Die Expertise und die Vernetzung aller an einer MA-IPD Beteiligten verbessern die Entwicklung der Fragestellung und die Interpretation der Ergebnisse auf einer multidisziplinären Basis und erleichtern die Identifikation unpublizierter Primärstudien, was zur Reduzierung von Publication-Bias führen kann [Stewart, 2002]. Allerdings ist es nicht selten, dass die Datenbesitzer bzw. Datenhalter auch von publizierten RCTs die Beteiligung

an einer MA-IPD verweigern oder einfach die Wiederabrufung der Daten versagen (Retrieval-Bias) [Schmid, 2003]. Eine Übersicht, die 44 MA-IPDs von RCTs einschloss, fand heraus, dass in etwa einem Viertel der MAs weniger als 80% der gesamten individuellen Patientendaten (IPDs) erhältlich waren [Simmonds, 2005]. Der Einschluss nur von RCTs, die sich mit IPDs beteiligen, kann zur Verzerrung der Ergebnisse der MA-IPD führen. Um diese Verzerrung zu verringern und alle Primärstudien in einer MA einzuschließen, wurde eine Kombination von IPDs und APDs empfohlen [Riley, 2007; Sutton, 2008]. Eine SR identifizierte 33 MAs, die eine solche Kombination vornahmen, und zeigte, dass IPDs lediglich für weniger als zwei Drittel der Primärstudien vorhanden waren [Riley, 2007].

Die bei den publizierten Auswertungen ausgeschlossenen Patienten innerhalb einer RCT können anhand der Originaldaten nach dem Intention-to-Treat-Prinizip im Rahmen von MA-IPDs eingegliedert werden. Dies führt zur Verringerung des sogenannten „Patient Exclusion-Bias“. Tatsächlich führen RCTs hoher methodischer Qualität die Datenaggregation nach dem Intention-to-Treat-Prinizip durch. MA-IPDs erlauben die Erforschung der Nachhaltigkeit des Interventionseffekts durch nach der Veröffentlichung erhobene Follow-Up-Daten [Clarke, 1998b]. Die Ergebnisse von RCTs sollen in der Regel erst nach ihrer Beendigung oder dem geplanten Abbruch publiziert werden. Falls eine aktive Nachverfolgung im Rahmen der Studie oder eine passive Erfassung, z.B. durch eine Verknüpfung mit einem Mortalitätsregister, der Patienten nach der Veröffentlichung der RCT stattfindet, bleibt die Frage nach der Erhaltung kontrollierter Studienbedingungen für die zusätzlich gewonnenen Daten (Performance-Bias) im Raum. Zudem kann die Bekanntmachung der RCT-Ergebnisse zu differenziellem Studienarmwechsel von Teilnehmern (Cross-Over-Bias), zu vermehrten Studienaustritten (Withdrawal-Bias) und zu variierender Erhebung der Endpunkte bei unverblindeter Bewertung (Detection-Bias) motivieren [Ioannidis, 1999].

Andererseits dehnt die Akquisition von Originaldaten auf patientenebene die Datenverifizierung und -auswertung aus. MA-IPDs bieten die Möglichkeit, Daten innerhalb von RCTs auf Plausibilität und Vollständigkeit zu prüfen, was Fehler bzw. Betrug aufdecken und fehlender Berichterstattung bzw. Unterschlagung vorbeugen kann (Verringerung des Erfassungsfehlers, Betrug und Reporting-Bias). Unterschiede in Bezug auf Definition, Kodierung oder Skala von APDs verhindern oft deren Kombination. Die Normierung von Variablen zwischen den RCTs in der Planungsphase der MA anhand von IPDs bahnt den Weg für deren Synthese über mehrere Primärstudien hinweg [Lyman, 2005]. Bei standardgemäßer Berichterstattung, z.B. nach dem CONSORT-Statement [Begg, 1997; Altman, 2001], lässt sich die methodische Qualität der RCTs im Rahmen von MA-APDs überprüfen. Die Vergleichbarkeit der Interventionsgruppen bezüglich prognostischer

Faktoren und die Einhaltung der Studienprozeduren können in größerem Umfang durch individuelle Patientendaten (IPDs) untersucht werden. Der Hauptvorteil von MA-IPDs gegenüber MA-APDs liegt darin, anhand von IPDs differenzielle Behandlungseffekte in Subgruppen der Patienten zu untersuchen und Time-to-Event-Auswertungen durchzuführen.

Allerdings sind MA-IPDs weitaus zeit- und personalintensiver als MA-APDs. Die Erzielung und Pflege einer effektiven Kollaboration unter einer größeren Zahl von Prüfzentren und Prüfärzten gilt als nicht trivial. Dies beinhaltet die Entwicklung von Richtlinien zu Datenbeteiligung und Autorenschaft, die Einrichtung eines Datenkoordinationszentrums und eines Schreibkomitees sowie die Organisation von Beteiligtenkommunikation und -treffen. Das Datenmanagement und die Aufbereitung zentral gesammelter „Rohdaten“ fordern die Standardisierung der Datenformate, der Messeinheiten und der Variablen sowie die Übersetzung medizinischer Terminologien bei multinationalen Primärstudien. Weiterhin sollen Datenabstimmung mit den Publikationen und Plausibilitätsprüfungen sowie das Management von fehlenden Werten erfolgen. Schließlich sollen Datentabellen zu Studien- und Patientenmerkmalen, zu klinischen Endpunkten, zu Follow-Up-Visits und Surrogatparametern sowie zu Medikationen für die Auswertung bereit stehen [Schmid, 2003]. Eine MA-IPD zu ACE-Hemmern gegen progressive Nierenerkrankung, die 1946 Patienten aus 11 RCTs einschloss, hat vier Jahre gedauert und 348 000 USD (179 USD pro Patient) gekostet [Schmid, 2003]. Eine weitere MA zum Risiko für ovarialen Krebs bei der Verwendung von Oralkontrazeptiva verursachte die 5-fachen Kosten für die MA-IPD im Vergleich zur MA-APD [Steinberg, 1997]. Ob der zusätzliche Nutzen durch MA-IPDs die Kosten rechtfertigt, ist bislang nicht ausreichend untersucht worden.

1.2.4. Berücksichtigung von Heterogenität bei der Synthese

Primärstudien zu einer Fragestellung ergeben oft unterschiedliche Punktschätzer und unterschiedliche Varianzen für die Punktschätzer. MA als eine Sekundärstudie zur Synthese von Primärstudien weist zwei Quellen statistischer Heterogenität auf: (i) Variabilität innerhalb der Primärstudien (Intra-Studien-Varianz „ σ_i^2 “) und (ii) Variabilität zwischen den Primärstudien (Inter-Studien-Varianz „τ^2 “). Die statistische Heterogenität in einer MA, die durch einen Heterogenitäts-Test zu prüfen und mit einem Heterogenitäts-Maß zu schätzen ist, kann auf methodische, klinische und/ oder zufallsbedingte Variabilität zurückgeführt werden. Bei der Synthese im Fixed-Effects-Modell (FEM) wird die statistische Heterogenität zwischen den Primärstudien lediglich auf Streuungen bei der Stichprobenziehung der Teilnehmer innerhalb jeder Primärstudie (Intra-Studien-Varianz) zurückgeführt. Damit wird die Inter-Studien-Varianz in der MA auf Null gesetzt. MAs im Random-Effects-Modell (REM)

berücksichtigen beide Quellen der statistischen Heterogenität und weisen unterschiedliche Schätzverfahren für die Inter-Studien-Varianz auf.

1.2.4.1. Testen und Schätzen von Heterogenität in Meta-Analyse

1.2.4.1.1. Heterogenitäts-Tests

Es gibt mehrere Testverfahren für die statistische Heterogenität zwischen den in einer MA eingeschlossenen Primärstudien unter Berücksichtigung der Intra-Studien-Varianzen: z.B. der Cochran-Test, der Woolf-Test und der Breslow-Day-Test [Sutton, 1998]. Eine Simulationsstudie, die fünf Heterogenitäts-Tests für MAs verglich, zeigte, dass der Cochran-Test (Q-Test) bezüglich der Vermeidung von Fehlern erster und zweiter Art sowie im Hinblick auf rechnerische Einfachheit am besten abschnitt [Takkouche, 1999]. Der Q-Test wird mit der Summe der quadratischen Abweichungen jeder Primärstudie von der zusammengefassten Effektgröße berechnet, wobei jede Primärstudie mit der Reziproke ihrer Intra-Studien-Varianz gewichtet wird. Der Q-Test ist näherungsweise Chi-Quadrat-verteilt und gilt als der Heterogenitäts-Test, der am meisten in MA verwendet wird [Sutton, 1998; Higgins, 2002b].

Zahlreiche Simulationsstudien zeigten eine niedrige statistische Power für den Q-Test bei MAs mit einer kleinen Anzahl von Primärstudien, was in der Praxis der MAs die häufigste Situation darstellt [Huedo-Medina, 2006; Mittlböck, 2006; Hardy, 1998; Paul, 1992]. Aufgrund der häufig niedrigen Power des Q-Tests wurde von zahlreichen Meta-Analysten eine Erhöhung des konventionellen Signifikanzniveaus für diesen Test auf mindestens 10% empfohlen [Fleiss, 1986; Boissel, 1989; Dickersin, 1992; Sutton, 1998, Petitti, 2001; Jackson, 2006]. Die „Cochrane Eyes and Vision Group“ empfiehlt, bei einem p-Wert des Q-Tests <0,05 keine MA durchzuführen, bei einem p-Wert zwischen 0,05 und 0,10 das FEM und REM zu benutzen und bei einem p-Wert > 0,10 das FEM zu verwenden [Higgins, 2002b]. In der Praxis wenig beachtet ist die Tatsache, dass ein nicht signifikanter Heterogenitäts-Test für eine MA mit einer kleinen Anzahl von Primärstudien entweder auf den Mangel an statistischer Power des Q-Tests zurückgeführt werden kann oder auf die Homogenität der Ergebnisse von Primärstudien, die den Bedarf an einer höheren Anzahl von Primärstudien verringert hat [Villar, 2001].

Es wird als problematisch betrachtet, dass bei einer hohen Anzahl von Primärstudien in einer MA die Nullhypothese der Homogenität zwischen den Primärstudien auch bei sehr geringer Heterogenität, die aus klinischer Perspektive vernachlässigbar ist, mittels des Q-Tests verworfen werden kann [Hardy, 1998; Higgins, 2003]. Da bei der Mehrheit der MAs nur wenige Primärstudien eingeschlossen werden, ist diese Konstellation eher selten vorzufinden.

Simulationsstudien liegen nahe, dass die statistische Power für den Q-Test sich verringert bei absteigender Summe der Gewichte von Primärstudien und bei ungleicher Verteilung der Gewichte zwischen den Primärstudien, insbesondere, wenn eine Primärstudie einen hohen Anteil an der Summe der Gewichte einnimmt [Hardy, 1998]. Weitere Simulationsstudien fanden heraus, dass die Power des Q-Tests sich proportional zur Inter-Studien-Varianz und umgekehrt proportional zur Intra-Studien-Varianz verhält. Sie hängt mit dem Quotienten zwischen Inter- und Intra-Studien-Varianzen zusammen [Mittlböck, 2006].

Eine mathematische „Analysis“ bestätigt die Ergebnisse der Simulationsstudien, dass der Q-Test oft niedrige statistische Power aufweist, und schlägt eine relativ einfache Rechenformel für die Power des Tests vor; allerdings nur für MAs mit Primärstudien, die die gleiche Intra-Studien-Varianz aufweisen [Jackson, 2006], was in der Praxis kaum vorkommt. In Analogie zur „Analysis of Variance“ wurde eine komplexe Rechenformel für die Power des Tests vorgeschlagen [Hedges, 2001, Hedges, 2004], was wegen fehlender Gewichtung der Primärstudien kritisiert wurde [Schmidt, 2007; National Research Council, 1992].

Eine F-Verteilung wurde für den Q-Test bei ungleichen Varianzen der Interventionsgruppen (Heteroskedastizität) innerhalb einer oder mehrerer Primärstudien vorgeschlagen [Kulinskaya, 2004]. Allerdings kann dieser modifizierte Q-Test nur für absolute Differenzen zwischen den Interventionsgruppen als Endpunktmaß angewendet werden und erfordert eine Information über die Varianz für jede Interventionsgruppe in jeder Primärstudie, die in Publikationen oft nicht gegeben werden.

1.2.4.1.2. Heterogenitäts-Maße

Da Heterogenität als ein Kontinuum zu betrachten ist, soll die Entscheidung, keine quantitative Zusammenfassung oder statistische Synthese im FEM oder im REM zu unternehmen, sich nicht nur an dem Testen der Existenz von Heterogenität in MAs, z.B. durch den Q-Test, sondern auch an der Schätzung ihres Ausmaßes orientieren. Zur

Quantifizierung von Heterogenität in MA wurden mehrere Maße vorgeschlagen [Takkouche, 1999; Higgins, 2002a; Higgins, 2003].

Takkouche, Cadarso-Suarez und Spiegelman schlagen zwei Heterogenitäts-Maße vor: (i) R_I ist Anteil der Inter-Studien-Varianz an der Gesamtvarianz (Intra-Studien-Varianz plus Inter-Studien-Varianz) und (ii) CV_B ist die quadratische Wurzel der Inter-Studien-Varianz, dividiert durch die zusammengefasste Effektgröße [Takkouche, 1999]. Higgins und Thompson präsentieren drei Heterogenitäts-Maße: (i) *R* ist die Ratio des Standardfehlers der zusammengefassten Effektgröße im REM zum Standardfehler der zusammengefassten Effektgröße im FEM, (ii) *H* ist die quadratische Wurzel der Q-Test-Statistik, dividiert durch ihre Freiheitsgrade, und (iii) I^2 ist eine Transformation des *H*, die die Heterogenität beschreibt, die nicht dem Zufall aus der Stichprobenziehung zuzuschreiben ist [Higgins, 2002a]. I^2 und *H* wurden von ihren Entwicklern als bevorzugte Heterogenitäts-Maße deklariert [Higgins, 2002a]. Das R_I und das I^2 sind ähnliche Heterogenitäts-Maße, wobei I^2 ein besseres Schätzverfahren für die Intra-Studien-Varianz aufweist als R_I [Mittlböck, 2006; Higgins, 2002a; Takkouche, 1999].

I^2 gilt als das am meisten verwendete Heterogenitäts-Maß. Es ermöglicht eine einfache Interpretation und steht für die Routineanwendung in der Software der Cochrane Collaboration (RevMan) zur Verfügung [Higgins, 2006]. I^2 wird mit folgender Formel ermittelt: $(Q - df) / Q$, wobei *df* die Zahl der Freiheitsgrade des Cochran-Tests *Q* darstellt. I^2 wird für negative Werte trunkiert und kann zwischen 0% und 100% liegen. Niedriger, moderater und hoher Heterogenität werden I^2-Werte von 25%, 50% und 75% zugeschrieben [Higgins, 2003].

Die Entwickler von I^2, *H* und *R* behaupten, dass diese Maße auf alle Skalen und Endpunktmaße anwendbar sind und unabhängig von der Anzahl der Primärstudien funktionieren. Damit seien sie für Vergleiche der Heterogenität zwischen unterschiedlichen MAs geeignet [Higgins, 2002a]. Daher kann die Berechnung des Heterogenitäts-Maßes (I^2) in der Gesamtpopulation und in den Subgruppen von Primärstudien oder Patienten in die Evaluation potenzieller Kovariate einfließen [Higgins, 2003]. Allerdings zeigten Simulationsstudien, dass die Power von I^2 doch von der Anzahl der Primärstudien abhängt [Mittlböck, 2006] und sogar mit der Power des Q-Tests vergleichbar ist [Huedo-Medina, 2006]. Im Gegensatz dazu verhielt sich die Power eines modifizierten *H* (H^2_M) bei variierender Anzahl primärer Studien stabil [Mittlböck, 2006]. H^2_M wird mit folgender Formel ermittelt: $(Q - df) / df$, wobei *df* die Zahl der Freiheitsgrade des Cochran-Tests *Q* darstellt. H^2_M

kann als Anzahl der Fälle, in denen die Inter-Studien-Varianz die Intra-Studien-Varianz übertrifft, interpretiert werden. Ein $I^2 \geq 50$ entspricht einem $H^2_M \geq 1$[Mittlböck, 2006].

1.2.4.1.3. Unsicherheit der Heterogenitäts-Maße

Ein Konfidenz-Intervall für I^2 kann die Unsicherheit bei ihrer Schätzung darstellen. Monte-Carlo-Simulationen, die standardisierte Mittelwertdifferenz als Endpunktmaß verwenden, zeigen, dass I^2 bei einer kleinen Anzahl von Primärstudien (<20) aufgrund der niedrigen statistischen Power ein weites Konfidenz-Intervall aufweist [Huedo-Medina, 2006]. Die Berichterstattung über das I^2 ist bislang, auch in renommierten Zeitschriften wie dem British Medical Journal, selten [Ioannidis, 2007c]. Eine empirische Studie berechnete die Konfidenz-Intervalle für I^2 bei 1011 MAs aus der Cochrane Library. Sie zeigt, dass bei 83% der MAs mit $I^2 \leq 25$ (niedrige Heterogenität) die obere Grenze des 95%-Konfidenz-Intervalls von $I^2 \geq 50$ betrug (hohe Heterogenität). Zudem fand sie heraus, dass bei 67% der MAs mit $I^2 \geq 50$ (hohe Heterogenität) die untere Grenze des 95%- Konfidenz-Intervalls von I^2 bei ≤ 25 (niedrige Heterogenität) lag. Zudem übertraf die obere Grenze des 95%- Konfidenz-Intervalls von I^2 bei allen MAs mit $I^2 = 0$ den 33%-Wert. Ähnliche Ergebnisse wurden für 50 MAs zu Zusammenhängen zwischen Genen und Erkrankungen gefunden [Ioannidis, 2007c]. Allerdings wurde in dieser Studie zur Berechnung der Konfidenz-Intervalle für I^2 nicht ein Bayesianisches Verfahren benutzt, wie von den Entwicklern des I^2 vorgeschlagen wurde, sondern es wurde ein auf einer nicht zentralen Chi-Quadrat-Verteilung basierender Ansatz verwendet [Ioannidis, 2007c; Higgins, 2002a]. Es ist aber davon auszugehen, dass die Wahl des Schätzverfahrens für das Konfidenz-Intervall keinen entscheidenden Einfluss auf die Ergebnisse hat.

1.2.4.2. Synthese-Modelle für Meta-Analyse

1.2.4.2.1. Fixed-Effects-Modell

Die Evidenz-Synthese von Primärstudien im FEM nimmt an, dass alle Primärstudien aus einer Verteilung stammen und eine einzige Effektgröße schätzen und dass die Unterschiede zwischen den Schätzern lediglich auf Streuungen der Stichprobenziehung von Studien-teilnehmern zurückzuführen sind. Diese Annahme ist sehr umstritten [Altman, 1995; Petitti, 1994; Cooper, 1994; Hedges, 1985]. Jedoch ist die Tendenz zu beobachten, diese Annahme

als eine starke Annahme einzustufen [Sutton, 1998; National Research Council, 1992]. Eine formale Überprüfung dieser Annahme mittels Heterogenitäts-Tests und -Maßen leidet unter der niedrigen statistischen Power des Q-Tests und der darauf basierenden Heterogenitäts-Maße (s. Abschnitt 1.2.4.1). Demzufolge soll eine MA im FEM in aller Regel mit einer im Random-Effects-Modell (REM) ergänzt werden und bei diskrepanten Ergebnissen der beiden Modelle zugunsten des REM entschieden werden.

Es bestehen mehrere Methoden zur Synthese im FEM. Die Invers-Varianz-Methode, die in den 1930er Jahren vorgeschlagen wurde, gilt als der am meisten verwendete Ansatz zur Synthese im FEM und im REM [Sutton, 1998]. Dabei wird die Summe der gewichteten Effektgrößen durch die Summe der Gewichte dividiert. Die Gewichtung der Effektgröße jeder Primärstudie erfolgt im FEM durch die Multiplikation der Effektgröße mit der Reziproke ihrer Intra-Studien-Varianz [Cooper, 1994].

Bei der in der biomedizinischen Forschung häufig vorkommende Verwendung von Relativem-Risiko (RR) oder OR, als Effektmaße in MAs, wurde die Synthese mit logarithmisch transformierten Maßen empfohlen [Fleiss, 1993]. Die Absicht ist, das kleine Intervall für relative Risikoreduktion (von 0 bis 1) bei diesen Maßen durch die Transformation zu vergrößern. Bei seltenem Auftreten des Endpunktereignisses in einer Primärstudie kann eine Zelle in der Kreuztabelle mit Nullereignis besetzt werden, was die Berechnung der Intra-Studien-Varianz für diese Primärstudie und damit die Anwendung der Invers-Varianz-Methode für die Synthese mit anderen Primärstudien unmöglich macht. Durch das Addieren von 0,5 zu jeder Zelle kann das Problem umgangen werden, ohne das Ergebnis unzulässig zu verfälschen [Emerson, 1994].

Die Mantel-Haenszel-Methode stellt einen weiteren Syntheseansatz im FEM dar, der insbesondere bei MA-IPDs nicht selten verwendet wurde [Dickersin, 1992]. Es existieren verschiedene Formeln für die Varianz der nach der Mantel-Haenszel-Methode zusammengefassten Effektgröße [Sutton, 1998]. Bei Zellen mit Nullereignis in einer oder mehreren Primärstudien kann die Mantel-Haenszel-Methode ohne eine Korrektur mit 0,5 nicht verwendet werden. Durch das Addieren von 0,5 zu jeder Zelle kann dieses Problem umgangen werden, ohne das Ergebnis unzulässig zu verfälschen [Emerson, 1994].

Bei MAs mit einer kleinen Anzahl von großen Primärstudien wird die Invers-Varianz-Methode empfohlen und bei MAs mit großer Anzahl von kleinen Primärstudien wird die Mantel-Haenszel-Methode bevorzugt [Fleiss, 1981; nach: Sutton, 1998].

Die Peto-Methode kann als eine modifizierte Mantel-Haenszel-Methode angesehen werden [Yusuf, 1985]. Anders als die Mantel-Haenszel- und die Invers-Varianz-Methode kann sie auch bei Zellen mit Nullereignis angewendet werden. Simulationsstudien fanden heraus, dass bei einer Ereignisrate von < 1% das zusammengefasste OR nach der Peto-Methode mit weniger Fehlern erster und zweiter Art verbunden ist als das zusammengefasste OR nach der Invers-Varianz-Methode und nach der Mantel-Haenszel-Methode mit 0,5-Zellen-Korrektur [Bradburn, 2007]. Daher kann die Peto-Methode bei MAs für seltene Ereignisse verwendet werden, z.B. für seltene schwerwiegende unerwünschte Arzneimittelwirkungen [Kearney, 2006]. Allerdings unterschätzt die Peto-Methode die zusammengefasste Effektgröße bei sehr unterschiedlicher Studienteilnehmerzahl in den Interventionsgruppen, bei sehr großem oder sehr kleinem RR bzw. OR [Fleiss, 1993]. Solche Konstellationen oder Ergebnisse sind in RCTs jedoch eher selten zu finden.

Maximum-Likelihood- und Restricted-Maximum-Likelihood-Methoden für das FEM stellen iterative Methoden dar, die komplex sind und äußerst selten verwendet wurden [Emerson, 1994]. Zudem zeigten Simulationsstudien identische Ergebnisse bei seltenen Ereignissen und lediglich kleine Unterschiede zwischen der Mantel-Haenszel-Methode und den beiden iterativen Methoden [Greenland, 1990], was den Aufwand für die letzteren nicht rechtfertigen kann.

Die stratifizierte Log-Rank-Analyse, die stratifizierte Cox-Regression-Modellierung und die Invers-Varianz-Methode stellen weit verbreitete Methoden zur Synthese von Zeit-bis-Ereignis-Endpunkten im FEM bei MA-IPDs dar. Mathematiktheoretische, simulations-basierte und empirische Vergleiche zeigten ähnliche Ergebnisse für diese Methoden im FEM bei MA-IPD mit wenigen Primärstudien und keiner bis niedriger Heterogenität [Tudur Smith, 2007].

1.2.4.2.2. Random-Effects-Modell

Die Synthese im Random-Effects-Modell (REM) berücksichtigt sowohl die Intra-Studien-Varianzen als auch die Inter-Studien-Varianz. Unter der Annahme, dass die in der MA eingeschlossenen Primärstudien eine zufällige repräsentative Stichprobe aller vorliegenden und zukünftigen Primärstudien zu einer Fragestellung darstellen, kann die zusammengefasste Effektgröße der MA im REM als der Mittelwert der Effektgrößen aller Primärstudien und die Inter-Studien-Varianz der MA als die Varianz dieses Mittelwerts betrachtet werden [Sutton, 1998].

Die Invers-Varianz-Methode ist der einzige verwendete Ansatz zur Synthese im REM. Dabei wird die Summe der gewichteten Effektgrößen mit der Summe der Gewichte dividiert. Die Gewichtung der Effektgröße jeder Primärstudie erfolgt im REM durch die Multiplikation der Effektgröße mit der Reziproke der Summe ihrer Inter- und Intra-Studien-Varianzen [Cooper, 1994]. Während die Intra-Studien-Varianz meistens mit einer einfachen Rechenformel aus den Daten jeder Primärstudie geschätzt wird, bestehen mehrere Schätzverfahren mit unterschiedlicher Rechenkomplexität für die Inter-Studien-Varianz. Die Vielfalt der Ansätze zur Synthese im REM richtet sich nach den unterschiedlichen Schätzern für die Inter-Studien-Varianz und für die Varianz der Inter-Studien-Varianz. Dies wird in den nächsten Unterabschnitten ausführlich behandelt.

1.2.4.2.2.1. Inter-Studien-Varianz

Die Inter-Studien-Varianz in MA wird manchmal die „Zwischen-Studien-Varianz", die „Random-Effects-Varianz" oder die „Heterogenitäts-Varianz" genannt und oft mit Tau-Quadrat (τ^2) bezeichnet [Sidik, 2007; Böhning, 2004; Makambi, 2004]. In der biometrischen Literatur werden für sie mehrere Schätzer entwickelt. Der Schätzer von Inter-Studien-Varianz nach DerSimonian und Laird (τ^2_{DL}) [DerSimonian, 1986] gilt als der am meisten verwendete [Higgins, 2006; Sutton, 1998; Brockwell, 2001; Thompson, 1999]. τ^2_{DL} ist die Chi-Quadrat-Statistik des Q-Tests abzüglich ihrer Freiheitsgrade, dividiert durch eine standardisierte Abweichung des Mittelwerts der Gewichte aller Primärstudien von der Varianz dieser Gewichte. Das Gewicht jeder Primärstudie wird, wie im FEM, durch die Reziproke der Intra-Studien-Varianz geschätzt. τ^2_{DL} weist negative Werte auf, wenn die beobachtete *Q*-Statistik ihren unter der Nullhypothese der Homogenität erwarteten Wert übersteigt [Cooper, 1994], diese negativen Werte werden gleich Null gesetzt. Bei allen iterativen und nicht-iterativen Schätzern der Inter-Studien-Varianz liegen negative Werte außerhalb des Parameterraums [Viechtbauer, 2007a]. τ^2_{DL} ist ein nicht-iterativer „Moment-of-Estimate"-Schätzer der Inter-Studien-Varianz, der nicht die Normalverteilung der Effektgrößen von Primärstudien für ihre mathematische Ableitung annimmt. Dieser Schätzer ist einfach zu berechnen [DerSimonian, 1986; Sutton, 1998].

Ein zweiter ähnlicher nicht-iterativer „Moment-of-Estimate"-Schätzer der Inter-Studien-Varianz, der ebenfalls einfach zu berechnen ist, wurde von Hedges und Olkin vorgeschlagen (τ^2_{HO}). Er bezieht aber die Gewichte der Primärstudien nicht ein und wird in der Psychologie

oft, in der Medizin jedoch selten verwendet [Hedges, 1985; DerSimonian, 1986; Sutton, 1998; Sidik, 2007]. Wie bei τ^2_{DL} werden negative Werte von τ^2_{HO} gleich Null gesetzt.

Ein dritter nicht-iterativer Schätzer der Inter-Studien-Varianz, der durch eine Parametrisierung der gesamten Varianz im REM abgeleitet wurde, wurde von Sidik und Jonkman vorgeschlagen (τ^2_{SJ}) [Sidik, 2005a].

Ein vierter nicht-iterativer Schätzer der Inter-Studien-Varianz, der auf den Wert der τ^2_{HO} bei der Parametrisierung von τ^2_{SJ} baut, wurde ebenso von Sidik und Jonkman (τ^2_{SJHO}) entwickelt [Sidik, 2007]. Durch die Parametrisierung können beide Schätzer von Sidik und Jonkman nur positive Werte für τ^2 ergeben.

Es existieren mehrere iterative Schätzer für die Inter-Studien-Varianz, die via „Maximum-Likelihood"- (ML), das „Restricted Maximum-Likelihood"-Verfahren (REML) und empirische oder voll Bayesianische Modellierung zu ermitteln sind [Spiegelhalter, 2000; Sutton, 1998; Thompson, 1999; Hardy, 1996; DerSimonian, 1986; Hedges, 1985; Raudenbush, 1985]. In aller Regel sind die iterative Schätzverfahren, ML und REML, von der Annahme der Normalverteilung der Effektgrößen von Primärstudien abhängig, sie brauchen Anfangswerte (Seeds) für die Iteration und Kriterien für die Konvergenz. Zudem soll der Schätzer von τ^2 für negative Werte bei jeder Iteration überprüft und gegebenenfalls auf Null gesetzt werden. Die Null wird oft als der Anfangswert für τ^2 benutzt, dennoch wurde empfohlen, als Anfangwert für die Iterationen einen nicht-iterativen Schätzer von τ^2 zu verwenden [DerSimonian, 2007].

τ^2_{REML} nach Raudenbush [Raudenbush, 1985] unterscheidet sich vom Schätzer der Inter-Studien-Varianz in MAs nach dem ML-Verfahren (τ^2_{ML}) dadurch, dass beim ersten (τ^2_{REML}) anders als beim zweiten (τ^2_{ML}) für die Schätzung von θ und τ^2 aus denselben Daten der Primärstudien adjustiert wird [DerSimonian, 1986]. Daher kann τ^2_{REML} als das iterative Äquivalent zu τ^2_{DL} betrachtet werden [Sutton, 1998; Thompson, 1999; DerSimonian, 1986; Raudenbush, 1985].

Basierend auf einer generellen „Moment-of-Estimate"-Methode für die Schätzung der Inter-Studien-Varianz in MAs von Laborstudien [Kacker, 2004] wurde von Paule und Mandel ein iterativer Schätzer für τ^2 (τ^2_{PM}) vorgeschlagen [Paule, 1982; DerSimonian, 2007]. τ^2_{PM} setzt keine Normalitätsannahme voraus, produziert bei Verletzung der Annahme einen „robusteren" Schätzer als τ^2_{DL} und entspricht bei der Erfüllung der Annahme dem Schätzer

der Inter-Studien-Varianz in MAs nach dem REML-Verfahren (τ^2_{REML}) [DerSimonian, 2007; Rukhin, 2000].

Ein iterativer empirischer Bayes-Schätzer wurde von Morris (τ^2_M) vorgeschlagen [Morris, 1983]. Er weist rechnerische Ähnlichkeit zu τ^2_{SJ} auf [Sidik, 2007], erfordert aber ein iteratives Schätzverfahren.

Alle vorher genannten Schätzer für die Inter-Studien-Varianz sind an verschiedenen Skalen und Maßen der Effektgröße anwendbar. Schätzer spezifisch für stetige Endpunkte, z.B. standardisierte und unstandardisierte Mittelwertdifferenz, sind eher in der Psychologie und Bildungsforschung weit verbreitet und werden von anderen Autoren detailliert behandelt [Viechtbauer, 2007a; Viechtbauer, 2007b; Hall, 2002; Hartung, 2001a; Hedges, 1985]. Bei stetigen Endpunkten zeigten die bisherigen Simulationsstudien, dass τ^2_{DL} und τ^2_{REML} besser als τ^2_{ML} und weitere Schätzer, z.B. τ^2 nach Hunter-Schmidt [Hunter, 1990], abschnitten [Viechtbauer, 2007a; Viechtbauer, 2007b].

Bei Simulationsstudien mit dem log-OR als Endpunktmaß erzielen bei niedriger bis moderater Inter-Studien-Varianz τ^2_M und τ^2_{SJHO} den niedrigsten Bias und bei hoher Inter-Studien-Varianz τ^2_{SJ} die besten Ergebnisse [Sidik, 2007]. Diese Simulationsstudien zeigen auch, dass τ^2_{DL} , τ^2_{ML} und τ^2_{REML} den Wert der Inter-Studien-Varianz unterschätzen und die Unterschätzung mit steigenden Werten von τ^2 umso stärker ausfällt. Dieselben Simulationsstudien beobachten eine Tendenz zur Überschätzung der Inter-Studien-Varianz durch τ^2_{HO}.

Eine empirische Studie mit einer MA zeigt eine Unterschätzung der Inter-Studien-Varianz durch τ^2_{ML} im Vergleich zu τ^2_{REML} [Thompson, 1999]. Eine weitere empirische Studie, die 8 MAs als Vergleichsgrundlage verwendete, zeigt, dass im Vergleich zu τ^2_{DL} und τ^2_{REML}, die ähnlich waren, τ^2_{ML} niedrigere und τ^2_{UE} inkonsistente Werte aufwiesen [DerSimonian, 1986].

1.2.4.2.2.2. Unsicherheit der Inter-Studien-Varianz

Wie jeder aus empirischen Daten abgeleitete Schätzer weist die Inter-Studien-Varianz (τ^2) statistische Unsicherheit auf. Allerdings zeigten Simulationsstudien, dass die Variabilität von τ^2 den Punktschätzer der zusammengefassten Effektgröße der MA nur geringfügig beeinflusst [Hardy, 1996; Viechtbauer, 2007a].

Mittels des „Profile-Likelihood“-Ansatzes und unter der Normalitätsannahme der Daten entwickelten Hardy und Thompson ein Konfidenz-Intervall für τ^2 (KI_{HT}) und unter seiner Berücksichtigung ein Konfidenz-Intervall für die zusammengefasste Effektgröße im REM (θ_{REM}) [Hardy, 1996]. Diese zusammengefasste Effektgröße kann für binäre, ordinale und stetige Endpunkte berechnet werden [Hardy, 1996].

Gegenüber der Synthese im FEM schreibt die Synthese im konventionellen REM bei hohem Wert von τ^2_{DL} den kleinen Primärstudien höhere Gewichte zu und den großen Primärstudien niedrige. Dieses wurde als Überbewertung kleiner Primärstudien zulasten großer Primärstudien kritisiert [Greenland, 1994]. Basierend auf einer annähernden Gamma-Verteilung für die *Q*-Statistik leiten Biggerstaff und Tweedie ein Konfidenz-Intervall für τ^2 (KI_{BT}) ab und unter seiner Berücksichtigung entwickeln sie neue Gewichte für die Primärstudien im REM [Biggerstaff, 1997]. Im Vergleich zu den Gewichten im REM, die die Punktschätzer von τ^2_{DL} einbeziehen, sind die neuen Gewichte für große Primärstudien höher und für kleine Primärstudien niedriger, was zu einer zusammengefassten Effektgröße führt, die zwischen den im FEM und im REM nach der Methode von DerSimonian und Laird zusammengefassten Effektgrößen liegt [Biggerstaff, 1997]. Ähnliche Gewichte für Primärstudien werden in beiden REM bei großer Anzahl von Primärstudien (>20) erwartet [Biggerstaff, 1997; Sutton, 1998]. Allerdings basiert das KI_{BT} lediglich auf einer Approximation der Q-Statistik an die Gamma-Verteilung [Viechtbauer, 2007a].

Basierend auf der Annahme, dass eine Parametrisierung der Inter-Studien-Varianz von Sidik und Jonkman Chi-Quadrat verteilt ist, wurde ein nicht-iteratives Konfidenz-Intervall für τ^2_{SJ} entwickelt (KI_{SJ}) [Sidik, 2007]. Da $\tau^2_{SJ} \geq 0$ ist, schließt sein Konfidenz-Intervall nur positive Werte ein.

Ebenfalls von Biggerstaff und Tweedie wurden unter der Annahme der Normalverteilung zwei aus iterativen Verfahren zu schätzende Konfidenz-Intervalle für τ^2 eingeführt. Das eine basiert auf der „Maxmium-Likelihood“-Methode ($KI_{BT\text{-}ML}$) und das andere auf der „Restricted-Maximum-Likelihood“-Methode ($KI_{BT\text{-}REML}$) [Biggerstaff, 1997; Viechtbauer, 2007a].

Ein iterativer Schätzer des Konfidenz-Intervalls für τ^2, der auf einer generalisierten Q-Statistik $\{\sum (\theta_i-\theta)^2 / \sigma_i^2+\tau^2\}$ basiert, wurde von Viechtbauer vorgeschlagen (KI_{VB}) [Viechtbauer, 2007a]. Bei Simulationsstudien wies das KI_{VB} eine höhere Wahrscheinlichkeit auf, den wahren Wert von τ^2 einzuschließen, als die nicht-iterativen KI_{BT} und KI_{VB} sowie als das iterative $KI_{BT\text{-}ML}$ und $KI_{BT\text{-}REML}$ [Viechtbauer, 2007a].

Knapp, Biggerstaff und Hartung beobachteten, dass bei stetigen Endpunkten die untere Grenze des Konfidenz-Intervalls τ^2 und bei binären Endpunkten die obere Grenze hohe Werte aufweisen [Knapp, 2006]. Basierend auf einer standardisierten Gesamtvarianz für REM {$1/(1-k)\ \sum w_i / w\ (\theta_i-\theta)^2$, wobei $w_i= 1/(\sigma_i^2+\tau^2)$, $w= \sum w_i$} schlugen sie eine Korrektur für die jeweilige Grenze des Konfidenz-Intervalls von τ^2 (KI_{KBH}) vor, wobei das KI_{KBH} annährend Chi-Quadrat-verteilt ist [Knapp, 2006].

Parametrische und nicht-parametrische Bootstraps stellen weitere iterative Verfahren (Resampling) dar, die für die Entwicklung von Konfidenz-Intervallen für jeden Schätzer der Inter-Studien-Varianz angewendet werden können [Turner, 2000; Switzer, 1992]. Die Konfidenz-Intervalle nach der parametrischen Bootstrap-Methode ($KI_{B\text{-}PAR}$) setzen keine Normalität von τ^2 voraus und nach der nicht-parametrischen Bootstrap-Methode ($KI_{B\text{-}NPAR}$) nehmen sie zusätzlich keine Normalitätsannahme der Effektgrößen von Primärstudien an [Viechtbauer, 2007a]. Allerdings zeigten Simulationsstudien, dass das $KI_{B\text{-}PAR}$ und das $KI_{B\text{-}NPAR}$ suboptimale statistische Eigenschaften aufweisen [Viechtbauer, 2007a]. Wie in anderen iterativen Schätzverfahren werden die negativen Werte von τ^2 gleich Null gesetzt.

1.2.4.2.3. Auswahl des Synthese-Modells

Statistisches Testen auf und biometrische Quantifizierung von Heterogenität in MA fungieren bislang federführend bei der Auswahl des Synthese-Modells. Diese Ansätze weisen erhebliche Einschränkungen auf. Hauptsächlich führt die niedrige Anzahl von RCTs in MAs zur hohen Wahrscheinlichkeit für Fehler zweiter Art bei verwendeten Heterogenitäts-Tests und -Maßen [Huedo-Medina, 2006; Cooper, 1994]. Existierende Heterogenität kann dadurch unentdeckt bleiben.

Die Auswahl des Synthese-Modells soll sich nicht nur auf das Testen und Schätzen statistischer Heterogenität stützen, sondern sich auch nach der systematisierten Einschätzung der methodischen und klinischen Heterogenität zwischen den Primärstudien richten. Leider wurde über die Grundlage der Selektion eines Synthese-Modells bei einem beträchtlichen Anteil von MAs nicht berichtet. Eine SR, die 38 MAs von klinischen Studien zur Infektion mit Helicobacter pylori einschloss, fand heraus, dass etwa die Hälfte der MAs über das Ergebnis eines Heterogenitäts-Tests berichtete. Über die Hälfte der 11 MAs, die das Signifikanzniveau des Tests genauer angaben, fassten trotz statistischer Heterogenität die Ergebnisse meta-analytisch zusammen. Lediglich 40% der MAs berichteten über die Auswahl des Synthesemodells und 26% begründeten diese [Huang, 2004].

Evidenz-Synthesen im FEM und REM unterscheiden sich in Bezug auf die statistische Folgerung (Inferenz). Dabei liefert FEM lediglich über die in der MA eingeschlossenen Primärstudien Aussagen, während REM Aussagen über diese und zukünftige Primärstudien ermöglicht [Bailey, 1987; Hedges, 1998]. Auf der einen Seite nimmt das FEM an, dass die in einer MA eingeschlossenen Primärstudien alle Primärstudien zu der Fragestellung darstellen, was dem Anspruch der MA entspricht, auf einer SR aller Primärstudien zu basieren [Villar, 2001]. Allerdings kann diese Annahme aufgrund des hoch prävalenten Publication-Bias nicht aufrechterhalten werden. Auf der anderen Seite fußt das REM auf der Annahme, dass die in einer MA eingeschlossenen Primärstudien eine zufällige Stichprobe aus der Verteilung aller Primärstudien darstellen und dass sie für alle Primärstudien repräsentativ sind. Aussagen über die Generalisierbarkeit der Ergebnisse von MA auf Patienten in zukünftigen Primärstudien sind daher inferenzstatistisch durch das REM möglich [Schmidt, 2007]. Allerdings kann die Repräsentativität von Primärstudien, die in einer MA eingeschlossen sind, für alle Primärstudien erst angenommen werden, wenn die eingeschlossenen Primärstudien die klinische Heterogenität der Patienten und der Interventionen in hohem Maße abbilden. So kann zum Beispiel: eine MA von Primärstudien mit niedrigem Frauenanteil, wie bei den Mega-RCTs für Statine [Baigent, 2005], nicht den Anspruch der Repräsentativität erheben. In der Regel schließt eine RCT eine relativ homogene Patientenstichprobe und standardisierte Interventionen ein und die klinische Variabilität ist eher zwischen als innerhalb der RCTs zu finden. Daher kann die Annahme der Repräsentativität im REM sich bei erhöhter Anzahl der in der MA eingeschlossenen Primärstudien verstärken.

Bei einer Inter-Studien-Varianz von Null sind die Ergebnisse des FEM und REM per se identisch. Die Verwendung dreier Standardabweichungen bei der Bildung des Konfidenz-Intervalls der zusammengefassten Effektgröße im FEM kann zur Verringerung der Unterschiede zwischen FEM und REM führen [Sutton, 1998]. Sie wurde von Peto, einem der stärksten Gegner des REM, propagiert [Peto, 1987].

Da die Varianz der zusammengefassten Effektgröße im REM aus der Summe der Intra- und Inter-Studien-Varianzen besteht, weist das REM bei vorliegender Heterogenität ein weiteres Konfidenz-Intervall auf als das FEM. Auch bei nicht signifikantem Heterogenitäts-Test und geringer Heterogenität kann das Konfidenz-Intervall im REM breiter sein als im FEM [Villar, 2001]. Da bei der Gewichtung im REM die Inter-Studien-Varianz berücksichtigt wird, erhalten bei großer Inter-Studien-Varianz kleine Primärstudien mit großer Intra-Studien-Varianz im REM mehr Gewicht als im FEM, große Primärstudien mit kleiner Intra-Studien-Varianz

erhalten weniger Gewicht im REM als im FEM. Demzufolge kann bei MAs mit großer Inter-Studien-Varianz und kleinen Primärstudien, die einen großen Interventionseffekt aufweisen, die zusammengefasste Effektgröße im REM höher sein als im FEM. Daher kann bei vorliegender Heterogenität das REM bezüglich der Breite des Konfidenz-Intervalls von zusammengefasster Effektgröße in aller Regel konservativer sein als das FEM; dennoch kann das REM hinsichtlich des Punktschätzers der zusammen-gefassten Effektgröße in manchen Fällen liberaler sein als das FEM [Poole, 1999; Villar, 2001].

Es kann eine steigende Anerkennung des Mangels an Homogenität zwischen den Primärstudien in der biomedizinischen Forschung und eine vermehrte Anwendung des REM für deren Synthese beobachtet werden [Sutton, 1998]. Diese Tendenz scheint sich allerdings auf MA-APD zu beschränken, da eine Übersicht von 36 MA-IPDs, die RCTs eingeschlossen haben, zeigte, dass das FEM in 81% der MA-IPDs als Synthese-Modell verwendet wurde. Das kann damit zusammenhängen, dass die MA-IPDs die Modellierung von Heterogenitäts-quellen bei der Synthese besser ermöglichen als die MA-APDs (s. Abschnitt 1.2.3).

Eine empirische Studie, die 125 MAs von RCTs einschloss, zeigte, dass bei 119 MAs das REM ein breiteres Konfidenz-Intervall für die zusammengefasste OR als das FEM ergab und bei 71 MAs das REM ein breiteres Konfidenz-Intervall für die zusammengefasste Risiko-Differenz als das FEM aufwies [Engels, 2000].

Eine empirische Studie, die alle 86 Cochrane MAs von RCTs mit RR als Endpunktmaß im Bereich pränataler Medizin einschloss, fand heraus, dass die zusammengefasste Effektgröße breitere Konfidenz-Intervalle im REM als im FEM aufwiesen und dieser Unterschied sich mit erhöhter Heterogenität vergrößerte [Villar, 2001]. Aus 21 MAs mit signifikanten Q-Tests auf dem Niveau von 10% wiesen sechs MAs Punktschätzer für die zusammengefasste Effektgröße im REM auf, die um mehr als 0,1 Einheiten auf der log-OR-Skala größeren Nutzen zugunsten experimenteller Interventionen zeigten als die Punktschätzer für die zusammengefasste Effektgröße im FEM [Villar, 2001]

Empirische Vergleiche zwischen dem FEM nach der Peto-Methode und dem REM nach der DerSimonian-Laird-Methode wurden für 22 MAs von RCTs durchgeführt [Berlin, 1989]. Lediglich in 3 MAs ergaben das FEM und das REM unterschiedliche Schlussfolgerungen, wobei das FEM einen signifikanten Nutzen für experimentelle Interventionen zeigte und das REM nicht [Berlin, 1989; Dickersin, 1992].

Eine weitere empirische Studie, die 169 MAs in der Fachzeitschrift „Psychological Bulletin“ ohne Einschränkung des Studiendesigns der Primärstudien identifizierte, zeigte, dass, obwohl das FEM in 76% der MAs verwendet wurde, über die Jahre dennoch das REM zunehmend häufiger zur Anwendung kam [Schmidt, 2007]. Innerhalb dieser empirischen Studie wurden 68 MAs, die im FEM durchgeführt wurden, im REM erneut durchgeführt. Die zusammen-gefassten Effektgrößen im FEM wiesen im Durchschnitt 51% schmalere Konfidenz-Intervalle auf als die im REM [Schmidt, 2007].

1.2.4.3. Weitere Aspekte der Synthese in Meta-Analyse

1.2.4.3.1. Intra-Studien-Varianz

In der geläufigen Praxis biomedizinischer Forschung wird die geschätzte Varianz des Interventionseffekts innerhalb jeder Primärstudie als die Intra-Studien-Varianz bei der Gewichtung in MA verwendet [DerSimonian, 1986; Thompson, 1999; Higgins, 2002a; Kulinskaya, 2004]. Dies wurde als problematisch erachtet bei MAs, die durch eine große Primärstudie mit einem engen Konfidenz-Intervall dominiert werden, d.h. eine Primärstudie weist im Vergleich zu den anderen ein sehr viel höheres Gewicht auf [Li, 1994; nach: Sutton, 1998]. Mathematische Ableitungen zeigten, dass in solchen MAs die Schätzung der Intra-Studien-Varianz mit der konventionellen Formel zur Unterschätzung der Varianz der zusammengefassten Effektgröße im FEM führt, d.h. zur Überschätzung der Summe der Gewichte, was in einer verfälschten Signifikanz der zusammengefassten Effektgröße resultieren kann [Li, 1994; nach: Sutton, 1998].

Anderseits scheint die Annahme der Gleichheit von Varianzen der Interventions- und der Kontrollgruppe (Homoskedastizität) innerhalb einer großen Primärstudie, z.B. einer RCT mit ausreichender Power für kleine bis moderate Effektgrößen, plausibel zu sein. Allerdings ist diese Annahme für kleine Primärstudien relativ stark und wird im generellen Kontext der MAs kritisiert [Böhning, 2002; DerSimonian, 1986]. Es wird jedoch in der Literatur nur selten vorgeschlagen, wie mit vermuteter Heteroskedastizität in den Primärstudien bei der MA umzugehen ist [DerSimonian, 1986; Van Houwelingen, 1993; Chang, 2001; Kulinskaya, 2004]. Zu diesen Vorschlägen gehört eine „Full-Likelihood-Methode“ für binomial-verteilte Endpunkte, wobei keine Approximation der Normalverteilung vorgenommen wird, die erwarteten Werte einer Vierfeldertafel durch ihre Randverteilungen bestimmt werden und die

zusammengefasste Effektgröße ein breiteres Konfidenz-Intervall als die Synthese unter der Annahme der Homoskedastizität aufweist [Van Houwelingen, 1993].

Eine „smoothed“ Intra-Studien-Varianz für die logarithmierten RR und OR wurde von Berkey entwickelt, wobei die Korrelation zwischen dem Punktschätzer und seiner Varianz reduziert wird [Berkey, 1995]. Basierend auf den Schätzern von Berkey wurde eine „smoothed“ Intra-Studien-Varianz für die Risikodifferenz hergeleitet [Knapp, 2003]. Bestimmte Effektmaße wie die Mittelwertdifferenz sind von ihrer Varianz stochastisch unabhängig und benötigen daher keine „smoothed“ Intra-Studien-Varianz [Knapp, 2003]. Simulationsstudien zeigten allerdings, dass die Verwendung von „smoothed“ oder konventionellen Intra-Studien-Varianzen die relativen Gewichte der Primärstudien im REM nicht ändert [Knapp, 2003].

1.2.4.3.2. Gewichtung kleiner Primärstudien

Da kleine Primärstudien mit großen Effektgrößen und signifikanten Ergebnissen zugunsten experimenteller Intervention bislang mit höherer Wahrscheinlichkeit publiziert werden als kleine Primärstudien mit kleinen Effektgrößen und nicht signifikanten Ergebnissen, werden sie eher in MAs identifiziert und eingeschlossen. Die Zuweisung hoher Gewichte zu kleinen Primärstudien und niedriger Gewichte zu großen Primärstudien bei beobachteter Heterogenität im REM wird von manchen Biometrikern kritisiert [Greenland, 1994]. Es wird angenommen, dass große Primärstudien, die oft hohe Präzision der Effektschätzer und gute methodische Qualität aufweisen, zugunsten von kleinen Primärstudien, die oft niedrige Präzision der Effektschätzer und schlechte methodische Qualität aufweisen, abgewertet werden [Thompson, 1991; Greenland, 1994; Kjaergard, 2001]. Allerdings erfolgt in der gängigen Praxis der MA die Gewichtung der Primärstudien im REM nach der Intra-Studien-Varianz und der Inter-Studien-Varianz, nicht jedoch nach der methodischen Qualität. Daher sollen kleine Primärstudien mit guter methodischer Qualität von MAs nicht ausgeschlossen bleiben und sie dürfen bei vorliegender Heterogenität in MA durch das REM mehr Gewicht bekommen.

1.2.4.3.3. Verteilung der Effektgrößen

In den meisten MAs im FEM oder im REM werden zur Konstruktion des Konfidenz-Intervalls für die zusammengefasste Effektgröße, die Perzentile der Normalverteilung verwendet [DerSimonian, 1986]. Dies fußt auf der Annahme, dass die Effektgrößen von Primärstudien

normal verteilt sind, was in der Regel mit mathematischer Vereinfachung verbunden ist [Goldstein, 2003]. Bei kleiner Anzahl von Primärstudien in der MA wurde diese Annahme als stark eingestuft [Brockwell, 2001; Böhning, 2002]. Zudem wurde anhand der Synthese einer kleinen empirischen MA und eines großen Multizentren-Trial argumentiert, dass die Breite des Konfidenz-Intervalls zusammengefasster Effektgröße (θ) mehr durch die Relation zwischen der zusammengefassten Effektgröße und der Inter-Studien-Varianz (τ^2) als durch die Anzahl von Primärstudien und die Präzision von τ^2 beeinflusst wird [Hardy, 1996].

Anders als bei dem konventionellen REM wird von Sidik und Jonkman ein Konfidenz-Intervall der zusammengefassten Effektgröße nach den Perzentilen der t-Verteilung vorgeschlagen (KI_{SJ}) [Sidik, 2002]. Simulationsstudien zeigen, dass bei kleiner Anzahl von Primärstudien das KI_{SJ} mit höherer Wahrscheinlichkeit den wahren Wert einschließt als das Konfidenz-Intervall mittels der Perzentile der Normalverteilung [Sidik, 2002]. Allerdings wurden die mathematischen Ableitungen von Sidik und Jonkman zur Approximation der t-Verteilung heftig kritisiert [Copas, 2003].

Basierend auf τ^2_{DL} wurde eine standardisierte Gesamtvarianz für das REM vorgeschlagen $\{1/(k - 1) \sum w_i / w (\theta_i - \theta)^2$, wobei $w_i = 1/ (\sigma_i^2 + \tau^2)$, $w = \sum w_i\}$ und ein Test für die statistische Signifikanz der zusammengefassten Effektgröße im REM, das nach den Perzentilen der t-Verteilung entscheidet [Hartung, 2001b]. Im Vergleich zum Test im konventionellen REM nach der DerSimonian-Laird-Methode ergaben Simulationsstudien kleinere Fehler erster Art für den Test von Hartung und Knapp [Hartung, 2001b].

Bei nicht normal verteilten Effektgrößen und hoher statistischer Heterogenität in MA wurden die Perzentile einer schiefen t-Verteilung für das Konfidenz-Intervall der zusammengefassten Effektgröße vorgeschlagen [Lee, 2008]. Es wurden ebenfalls parametrische Misch-Normalverteilungen [Carrol, 1999] und Bayesianische nicht-parametrische Ansätze [Ohlssen, 2007] für nicht-normal verteilte Effektgrößen empfohlen. Allerdings können etablierte Tests auf Normalität nicht direkt an MAs angewendet werden und die Anpassung von mehreren Verteilungen oder nicht Normalverteilungen für MAs sind rechnerisch komplex [Lee, 2008].

1.2.4.3.4. Synthese mit Kovariablen

Die Gewichtung der kleinsten Quadrate im Meta-Regressions-Modell kann im FEM oder im REM erfolgen. Da unterschiedliche Schätzverfahren für Inter-Studien-Varianzen zu unterschiedlichen Gewichtungen der Primärstudien im REM führen können, sind

unterschiedliche Schätzungen des Einflusses von Kovariablen anhand von Meta-Regression zu erwarten. Fehler bei der Schätzung der Inter-Studien-Varianz können die Varianz der Koeffizienten einer im Meta-Regressions-Modell eingeschlossenen Kovariablen (β_V) beeinflussen und somit das Testen auf Signifikanz der Kovariablen verfälschen. Bislang wurden sehr begrenzt modifizierte Schätzer für β_V entwickelt und untersucht, die für Fehler bei der Schätzung der Inter- und Intra-Studien-Varianzen adjustieren [Sidik, 2005b; Knapp, 2003; Thompson, 1999; Berkey, 1995].

Obwohl die Schätzung der Intra-Studien-Varianz in großen Primärstudien oft mit vernachlässigbarem Fehler verbunden ist, ist der Schätzer der Inter-Studien-Varianz in kleinen MAs, d.h. MAs mit niedriger Anzahl von Primärstudien, oft nicht präzise [Sidik, 2005b; Thompson, 1999]. Knapp und Hartung schlugen einen modifizierten Schätzer für β_V vor, der auf der Summe der gewichteten kleinsten Quadrate geteilt durch die Freiheitsgrade basiert [Knapp, 2003]. Hierbei können für die Gewichtung der kleinsten Quadrate im Meta-Regressions-Modell unterschiedliche Schätzer für τ^2 und σ^2 benutzt werden. Die Freiheitsgrade werden durch die Anzahl der Primärstudien minus der Anzahl der Kovariablen berechnet. Auf β_V basierend wurde ein Signifikanztest für eine Kovariable in der Meta-Regression entwickelt, wobei die Perzentile der t-Verteilung für die kritischen Werte des Tests verwendet werden [Knapp, 2003]. Simulationsstudien zeigten, dass der modifizierte Test für eine Kovariable nach Knapp und Hartung besser das gesetzte Signifikanzniveau einhielt als ein konventionelles Testverfahren, das aber die Fehler bei der Schätzung von τ^2 nicht berücksichtigt und nach den Perzentilen der Normalverteilung entscheidet [Knapp, 2003].

Sidik und Jonkman entwickelten einen „robusten“ Schätzer für β_V [Sidik, 2005b], der aus τ^2_{SJ} abgeleitet wurde [Sidik, 2005a]. Simulationsstudien ergaben bessere Eigenschaften für den Schätzer von β_V nach Knapp und Hartung als für den nach Sidik und Jonkman, wobei beide besser als der konventionelle Schätzer für β_V abschnitten [Sidik, 2005b].

1.2.4.4. Fazit

Aufgrund der niedrigen Anzahl von Primärstudien in den meisten MAs mangelt es allen Heterogenitäts-Tests an Power. Außerdem mangelt es allen Schätzern für das Heterogenitäts-Maß und für die Inter-Studien-Varianz an Präzision [Huedo-Medina, 2006; Ioannidis, 2007c]. Da alle Schätzer, auch die von Inter-Studien-Varianz in MA, Variabilität aufweisen, kann ein Einblick in die Stabilität der Synthese im REM gewonnen werden, durch

eine graphische Darstellung variabler Werte für die Inter-Studien-Varianz, die in einem Konfidenz-Intervall der τ^2, z.B. nach der Profile-Likelihood-Methode [Hardy, 1996], eingeschlossen sind, gegenüber den daraus resultierenden Schätzern der zusammengefassten Effektgröße im REM. In aller Regel ist die Anlehnung nur an den Punktschätzer eines Heterogenitäts-Maßes, einschließlich des I^2, zur Selektion des Synthese-Modells oder zum Verzicht auf eine statistische Synthese sehr problematisch. Aufgrund der niedrigen Anzahl von Primärstudien ist die Schätzung des I^2 oft mit großer statistischer Unsicherheit verbunden [Ioannidis, 2007c]. Ein breites Konfidenz-Intervall des I^2, das Werte für hohe und niedrige Heterogenität beinhaltet, ist schwer zu interpretieren.

Es wurden vier iterative und vier nicht-iterative Schätzer für die Inter-Studien-Varianz in MA in der biometrischen Literatur identifiziert und beschrieben. In aller Regel sind iterative Schätzverfahren rechnerisch komplexer als die meisten nicht-iterativen und es liegt nahe zu vermuten, dass sie für die Mehrheit der Meta-Analytiker nicht einfach durchzuführen sind. Die Schätzung der Inter-Studien-Varianz nach DerSimonian und Laird basiert, wie die meisten Schätzer, auf der Q-Statistik, nimmt, anders als die meisten Schätzer, nicht eine Normalverteilung der Effektgrößen an, ist nicht-iterativ und einfach zu berechnen. Dieser Schätzer wird bisher am meisten für das REM benutzt. Trotz weitgehender Recherche konnten keine extensiven empirischen Studien oder Simulationsstudien mit Belegen gegen seine weitere Anwendung gefunden werden.

Die Mehrheit der Untersuchungen über das Testen, das Schätzen von Heterogenität in MA und über die darauf basierende Auswahl des Synthese-Modells wurde seit 1986 veröffentlicht [DerSimonian, 1986]. Empirische Vergleiche und Simulationsvergleiche zwischen den verschiedenen Schätzern der Inter-Studien-Varianz, der Heterogenitäts-Maße und ihren Varianzen sind bislang von begrenzter Zahl, sie basieren auf bestimmten nicht ausreichend repräsentativen MAs oder sind durch eher restriktive oder unrealistische Annahmen belastet [Sidik, 2007]. Simulationsstudien zur Evaluation verschiedener Schätzer für die Inter-Studien-Varianz und die Heterogenitäts-Maße sollen auf realistischen Annahmen über die Verteilung von Primärstudien und von MAs konzipiert werden, damit sie repräsentative Daten simulieren. Da bislang keine ausreichenden Daten zu Merkmalen von Primärstudien (Effektgrößen, Intra-Studien-Varianz) und von MAs (Anzahl der Primärstudien, Inter-Studien-Varianz) [Simmonds, 2005; Mallett, 2003; Engels, 2000; Sterne, 2000] in der medizinischen Forschung vorliegen, können abgeschlossene simulierte MAs zur Validierung der Schätzer nur begrenzt realitätsnah sein.

SRs der empirischen Vergleiche und der Simulationsvergleiche zwischen den verschiedenen Schätzern der Inter-Studien-Varianz, der Heterogenitäts-Maße und ihren Varianzen sind geboten [Sutton, 2008] und können für die Konzipierung und Durchführung weiterer Vergleiche verwendet werden. Kritisch zu sehen ist der beobachtete Umstand, dass mit wenigen Ausnahmen [Sidik, 2005b; DerSimonian, 2007] die Ergebnisse der Mehrheit der bisherigen Simulationsstudien zur Evaluation eines neu vorgeschlagenen Schätzers für einen Parameter in MA zugunsten des neuen Schätzers ausfallen. Dem kann mit ähnlicher Skepsis begegnet werden, wie den positiven Ergebnissen von RCTs zugunsten der Intervention des Studien-Sponsoren. Trotz weitgehender Recherche konnten keine Kriterien zur kritischen Bewertung von Simulationsstudien für MAs gefunden werden. Leitlinien zum Design und zur Berichterstattung von Simulationsstudien in der medizinischen Biometrie sind eher selten und nicht umfassend, sie weisen keinen Konsens auf und wurden bislang selten befolgt [Burton, 2006; Demirtas, 2007]. Allerdings können sie für die retrospektive Bewertung der Simulationsstudien von MAs nützlich sein. Die Notwendigkeit einer externen kritischen Bewertung der methodischen Qualität besteht sowohl für empirische Studien als auch für Simulationsstudien.

1.2.5. Untersuchung von Heterogenität in Meta-Analyse

MA wird zunehmend nicht nur als ein Verfahren der Evidenz-Synthese betrachtet, sondern auch als ein Ansatz zur Heterogenitäts-Analyse verwendet [Sutton, 2008]. Da MA eine beobachtende Studie ist (s. Abschnitt 1.1.4.1), kann sie nur zur explorativen Untersuchung der Zusammenhänge zwischen potenziellen Heterogenitätsquellen und dem Interventions-effekt benutzt werden. Eine Assoziation zwischen einem möglichen Effektmodifikator (Kovariable) und dem Interventionseffekt, die in einer MA gefunden wurde, kann durch weitere Kovariablen verzerrt werden [Thompson, 1999]. Die Heterogenitäts-Analyse im Rahmen von MAs ist für Confounding anfällig und kann fälschlicherweise zur Identifikation einer Patientensubgruppe führen, die von einer Intervention mehr, weniger oder keinen Nutzen bekommt. Dies hat zur Folge, dass die Intervention für diese Subgruppe der Patienten irrtümlich überhaupt nicht bzw. mehr oder weniger verschrieben wird als für andere Subgruppen, obwohl sie sich in Wahrheit nicht von diesen unterscheidet. Der Verfasser vertritt ausdrücklich die Ansicht, dass konfirmatorische Ergebnisse bezüglich eines Effektmodifikators nur im Rahmen von RCTs mit hoher methodischer Qualität und ausreichender statistischer Power erreicht werden können. Das Hauptziel der Heterogenitäts-Analyse in MAs soll die Generierung von Hypothesen für zukünftige RCTs sein.

1.2.5.1. Ansätze im Überblick

Es bestehen mehrere Ansätze zur Analyse von Heterogenität in MAs. Jeder Ansatz weist Vorteile und Nachteile auf, die in Betracht gezogen werden sollen. Ein Überblick über die Ansätze wird in Tab. 5 dargestellt. Da Subgruppen-Analyse und Meta-Regression als die am meisten verwendeten Ansätze in MAs gelten, wird in den nächsten Abschnitten auf sie näher eingegangen. Abschließend werden gebräuchliche, graphische Darstellungen und Methoden zur Verringerung von Heterogenität in MA kurz beschrieben.

Tab. 5 Ansätze zur Analyse von Heterogenität in Meta-Analyse von RCTs

Ansatz	Beschreibung	Nachteile	Vorteile
Schwellenwert-Analyse	Ausschluss von Primärstudien, die eine Ausprägung bestimmter Kovariablen (Ausschlussprägung) aufweisen.	(i) Informationsverlust durch ausgeschlossene Primärstudien (ii) Schwierige Implementierung bei kleiner Zahl von Primärstudien (iii) Die Bestimmung der Kovariablen und/ oder deren Ausschlussprägung kann arbiträr, Primärstudien-geleitet, a priori unbegründet oder subjektiv sein. (iv) Die Kategorisierung nicht binärer Kovariablen kann die Auswahl der Ausschlussprägung erschweren.	(i) Lediglich Primärstudien mit vordefinierter, Evidenz-basierter, und/ oder „plausibler“ Kovariable und deren Ausschlussprägung (z.B. Primärstudien „minimaler“ Qualität) können in die MA eingeschlossen werden. (ii) Die statistische Formalisierung der Auswahl des Schwellenwerts (z.B. Mittelwert plus eine Standardabweichung, Median oder Interquartil-Bereich) ist möglich.
Sensitivitäts-Analyse	Durchführung zweier MAs vor und nach dem Ausschluss von Primärstudien mit der Ausprägung bestimmter Kovariablen (Ausschlussprägung).	(i) Informationsverlust durch ausgeschlossene Primärstudien (ii) Schwierige Implementierung bei kleiner Zahl von Primärstudien (iii) Die Bestimmung der Kovariablen und/ oder deren Ausschlussprägung kann arbiträr, Primärstudien-geleitet, a priori unbegründet oder subjektiv sein. (iv) Die Kategorisierung nicht binärer Kovariablen kann die Auswahl der Ausschlussprägung erschweren.	(i) Lediglich Primärstudien mit vordefinierter, Evidenz-basierter, und/ oder „plausibler“ Kovariable und deren Ausschlussprägung (z.B. Primärstudien „bester“ Qualität) können in eine der beiden MAs eingeschlossen werden. (ii) Die statistische Formalisierung der Auswahl des Schwellenwerts (z.B. Mittelwert plus eine Standardabweichung, Median oder Interquartil-Bereich) ist möglich. (ii) Gesamt- und Teilmenge von Primärstudien werden in zwei MAs einbezogen und verglichen (d.h. weniger Verlust an Primärstudien).

Ansatz	Beschreibung	Nachteile	Vorteile
Graphische Darstellung	Darstellung zweier oder mehrerer MAs von Primärstudien einschließlich der Punktschätzer und der Konfidenz-Intervalle (Y-Achse) gegenüber den Ausprägungen bestimmter Kovariablen (X-Achse). Darstellung von Primärstudien, einschließlich der Punktschätzer und der Konfidenz-Intervalle (Y-Achse), gegenüber den Ausprägungen bestimmter Kovariablen (X-Achse).	(i) Visuelle Verzerrungen können auftreten. (ii) Die Subjektivität bei der Interpretation kann hoch sein. (iii) Die Bestimmung der Kovariablen und/ oder deren Strata kann arbiträr, primärstudiengeleitet, a priori unbegründet oder subjektiv sein. (iv) Die Kategorisierung nicht binärer Kovariablen kann die Auswahl der Strata erschweren.	(i) Die Visualisierung der Heterogenität ist möglich. (ii) Die Exploration der Heterogenität ist möglich.
Kumulative Meta-Analyse	Sequenzielles, meta-analytisches Kombinieren von Primärstudien nach Anzahl der Ausprägungen bestimmter Kovariablen (Strata).	(i) Die Bestimmung der Kovariablen und/ oder deren Strata kann arbiträr, Primärstudien-geleitet, a priori unbegründet oder subjektiv sein. (ii) Die Kategorisierung nicht binärer Kovariablen kann die Auswahl der Strata erschweren.	(i) Die univariate Analyse der Evolution der zusammengefassten Effektgröße bezüglich einer Kovariablen ist möglich. (ii) Trotz der multiplen MAs besteht keine Multiplizitäts-Problematik. (iii) Die Exploration der Heterogenität ist möglich.
Gewichtung	Gewichtung der Primärstudien mit dem Produkt der Präzision der Primärstudie und der Ausprägung bestimmter Kovariablen	(i) Die statistische Rechtfertigung ist mangelhaft. (ii) Die Bestimmung der Kovariablen und/ oder deren Ausprägungen kann arbiträr, primärstudiengeleitet, a priori unbegründet oder subjektiv sein.	(i) Eine Qualitäts-Komponente oder ein Qualitäts-Score kann zur Gewichtung beitragen. (ii) Die Berücksichtigung mehrerer Heterogenitätsquellen bei der Synthese (z.B.Präzsion und methodische Qualität) ist möglich.

Ansatz	Beschreibung	Nachteile	Vorteile
Bayesianische Hierarchische Modellierung	Modellierung der a priori-Verteilungen über der Ausprägung bestimmter Kovariablen mit der primärstudien-internen Ausprägung der Kovariablen.	(i) Experten-basierte a priori-Verteilungen über der Ausprägung bestimmter Kovariablen sind subjektiv. (ii) Die Elizitation der a priori-Verteilungen von Experten ist schwer.	(i) Die Subjektivität kann transparent modelliert werden. (ii) Unsicherheiten sind für alle Parameter modellierbar. (iii) Die Empirie-basierte a priori-Verteilung ist eher objektiv. (iv) Sensitivitäts-Analysen bei variablen a priori-Verteilungen sind zur Robustheitsprüfung möglich.
Subgruppen-Analyse	Durchführung zweier oder mehrer MAs von Primärstudien je nach Anzahl der Ausprägungen bestimmter Kovariablen (Strata) und statistischer Vergleich der MAs.	(i) Falsch-negative Ergebnisse infolge niedriger Power des statistischen Vergleichs sind bei starker Heterogenität und/ oder bei kleiner Zahl von Primärstudien mit lediglich aggregierten Patientendaten wahrscheinlich. (ii) Falsch-positive Ergebnisse infolge statistischer Vergleiche bei multiplen Subgruppen-Analysen und/ oder bei durch Transformation multi-skalierten Effektgrößen sind wahrscheinlich (ausgeprägte Multiplizitäts-Problematik). (iii) Die Bestimmung der Kovariablen und/ oder deren Strata kann arbiträr, Primärstudien-geleitet, a priori unbegründet oder subjektiv sein. (iv) Die Kategorisierung nicht binärer Kovariablen kann die Auswahl der Strata erschweren.	(i) Die Formalisierung des statistischen Vergleichs der Primärstudien mittels Testen und Schätzen ist möglich. (ii) Das Aufweisen hoher statistischer Power bei einer großen Zahl von Primärstudien und/ oder verfügbaren individuellen Patientendaten ist möglich. (iii) Die Substratifizierung bei einer großen Zahl von Primärstudien und/ oder von verfügbaren individuellen Patientendaten ist möglich.

Ansatz	Beschreibung	Nachteile	Vorteile
Meta-Regression	Durchführung uni- und multivariater Regressionen von Primärstudien	(i) Falsch-negative Ergebnisse infolge niedriger Power des statistischen Vergleichs sind bei starker Heterogenität und/ oder bei kleiner Zahl von Primärstudien mit lediglich aggregierten Patientendaten wahrscheinlich. (ii) Die Bestimmung der Kovariablen und/ oder deren Ausprägungen kann arbiträr, primärstudiengeleitet, a priori unbegründet oder subjektiv sein. (iii) Die Kategorisierung nicht binärer Kovariablen kann die Auswahl der Ausprägungen erschweren.	(i) Die Formalisierung des statistischen Vergleichs der Primärstudien mittels Testen und Schätzen ist möglich. (ii) Die Berechnung der für eine oder mehrere Kovariablen adjustierten Effektgröße ist möglich. (iii) Uni- und multivariate Analysen, einschließlich der Wechselwirkung zwischen Kovariablen, bei großer Zahl von Primärstudien und/ oder verfügbaren individuellen Patientendaten sind möglich. (v) Falsch-positive Ergebnisse infolge statistischer Vergleiche mittels multivariaten Analysen sind weniger wahrscheinlich als bei multiplen Subgruppen-Analysen.

1.2.5.2. Subgruppen-Analyse

Ein interdisziplinärer, wissenschaftlicher Diskurs (Epidemiologie, Medizin, Biostatistik) über Chancen und Risiken der Subgruppen-Analysen in RCTs ist intensiv seit den 1980er Jahren zu beobachten [Buyse, 1989; Schneider, 1989; Yusuf, 1991; Gheorghiade, 1991; Oxman, 1992; Feinstein, 1998; Assmann, 2000; Parker, 2000; Pocock, 2002; Cook, 2004; Rothwell, 2005; Cuzick, 2005; Lagakos, 2006; Hernández, 2006]. Bei einer Subgruppen-Analyse werden die Primärstudien bei MA-APD und die Patienten bei MA-IPD nach einem in jeder Subpopulation identischen oder ähnlichen Merkmal (Kovariable) aufgeteilt, woraufhin eine MA für jede Subgruppe getrennt durchgeführt wird und für den Unterschied zwischen den Subgruppen mittels Interaktions- oder Homogenitäts-Test geprüft wird. Subpopulationen können nach Merkmalen der Patienten und/ oder der Primärstudien gebildet werden.

1.2.5.2.1. Zielsetzungen

Subgruppen-Analysen weisen drei Ziele auf: Erstens: die Stärkung der internen Validität des durchschnittlichen Interventionseffekts durch die Überprüfung der Konsistenz des Effekts in Strata der Patienten oder der Primärstudien, da dies den Kausalitätsnachweis verbessert. Zweitens: die Erhöhung der externen Validität durch die Schätzung der Vergleichbarkeit der Subgruppen der Patienten, die in den Primärstudien eingeschlossen waren mit denen, die in der Routineversorgung eingeschlossen werden, da dies die Übertragbarkeit der Ergebnisse der Primärstudien in der Praxis verbessert. Drittens: die Förderung des Maßschneiderns des Interventionseffekts durch die Überprüfung der Differenzierbarkeit des Interventionseffekts nach Charakteristika der Patienten. Während die Evaluation der Konsistenz eine überragende Anerkennung überwiegend durch methodische und regulatorische Instanzen genießt [Brookes, 2001; ICH, 1998b], erhält die Untersuchung klinischer Heterogenität zu wenig Unterstützung von Gesundheitsversorgern und Interventionsherstellern [Kravitz, 2004; Rothwell, 2005; Kraemer, 2006].

1.2.5.2.2. Quantitative und qualitative Subgruppenunterschiede

Falls ähnliche Effekte in verschiedenen Subgruppen der in einer RCT eingeschlossenen Patienten gefunden werden, gewinnt der durchschnittliche Effekt an interner Validität, an Generalisierbarkeit auf größere Patientenkollektive und an Extrapolierbarkeit auf individuelle Patienten. Falls quantitative Variationen bezüglich des Effekts, d.h. Unterschiede in dem

Ausmaß, aber nicht in der Richtung des Effekts, gefunden werden, kann unter ihrer Berücksichtigung der durchschnittliche Effekt in der entsprechenden Gesamtpopulation geschätzt werden. Zudem können solche Unterschiede von klinischer Relevanz sein. Qualitative Unterschiede, d.h. Unterschiede im Ausmaß und der Richtung des Effekts, wurden bisher selten gesehen [Glasziou, 1998].

1.2.5.2.3. A priori Festlegung der Subgruppen-Analysen

Die meisten Methodiker befürworten eine prospektiv geplante Subgruppen-Analyse, die im Protokoll oder Amendment der MA festgeschrieben sein soll. Das Protokoll soll einem Genehmigungsverfahren mit hoher Qualität unterzogen werden. Das vor der tatsächlichen Studiendurchführung der MA publizierte Protokoll macht die Behauptung einer a priori Bestimmung der subgruppenbildenden Kovariblen nachvollziehbar. Eine a posteriori definierte Subgruppen-Analyse soll immer als solche gekennzeichnet sein, darf nur als explorativ betrachtet und kann lediglich für die Hypothesenbildung verwendet werden.

Die a priori-Bestimmung der Subgruppen-Analysen dient der Reduzierung ihrer Anzahl und der Begründung ihrer Auswahl, wobei letztere durch vorherige Evidenz gestützt sein soll. Die Praktizierung der „fishing expeditions“, des „data dredging“, des „torturing the data until it confesses“ [Kraemer, 2006, S. 1288] und des „HARKing: Hypothesizing After the Results are Known“ [Kerr, 1998, S. 196], um signifikante Subgruppenunterschiede zu finden, soll vermieden werden.

1.2.5.2.4. Selektion, Erhebung und Aufbereitung der Kovariablen

Im Rahmen einer Subgruppen-Analyse sollen die Kovariable, ihre Ausprägungen sowie ihr Auswertungsverfahren a priori bestimmt werden. Die Auswahl der subgruppen-bildenden Kovariablen soll sich möglichst auf die Ergebnisse von SRs von nicht-randomisierten klinischen Studien oder SRs von beobachtenden Studien stützen. Die Vordefinierung der Kovariablen nur durch Expertenmeinung ist oft mit Verzerrungen verbunden. Erhebliche Zweifel sind allerdings angebracht an der Vorbestimmung von Kovariablen in MAs von RCTs, falls diese auf den Ergebnissen der RCTs basieren, die in die MA eingeschlossen sind. Die Prämisse, datengebundene Selektionsverfahren zu vermeiden, ist dann verletzt.

Valide Erhebung und reliable Erfassung der Kovariablen sollen angestrebt werden. Die Kovariablen sollten erhoben werden, bevor die Studienteilnehmer den Interventionen exponiert werden (an der „Baseline“), um Kontamination der Werte der Kovariablen mit dem Interventionseffekts zu vermeiden. Die Erhebung und Dokumentation der Kovariablen soll bei klinischen Studien mit sogenannten „Run-In“- oder „Wash-Out“-Perioden vor jeglichem Einsatz der experimentellen Interventionen erfolgen.

Untersucht werden häufig vereinzelte (z.B. Alter, Geschlecht) oder kombinierte soziodemographische Kovariablen (z.B. Sozioökonomischer Status, Ethnizität) und biomedizinische Kovariablen (z.B. Blutfett, Blutdruck, Krankheitsphase, Schweregrad der Erkrankung, Interventionsdosis), die als etablierte oder vermutete prognostische Faktoren des Interventionseffekts betrachtet werden können. Diesen Kovariablen soll durch die Ergebnisse vorheriger empirischer Studien eine zentrale Rolle im pharmakologischen Wirkungsmechanismus oder im pathophysiologischen Krankheitsverlauf zugeschrieben werden.

Besonderer Aufmerksamkeit bedürfen Subgruppen-Analysen bezüglich nicht kategorisierter Kovariablen und Endpunkte. Das betrifft die Kategorisierung stetiger Variablen sowie die Skalierung und die Zusammenlegung der Ausprägungen ordinaler und nicht-ordinaler Variablen. Beim Fehlen eines Empirie-gestützten Leitfadens können mehrere Kategorisierungen einer Kovariablen nach dem Median, nach einem graphisch explorierten Schwellenwert und nach dem minimalen p-Wert-Ansatz [Mazumdar, 2000] unternommen und verglichen werden.

1.2.5.2.5. Testen auf Subgruppenunterschiede

Falls die Daten innerhalb der Subgruppen relativ homogen sind aber eine beträchtliche Variation zwischen den Subgruppen zu beobachten ist, kann eine wichtige Heterogenitäts-quelle identifiziert werden. Subgruppenunterschiede sollen mittels sogenannter Interaktions- oder Homogenitäts-Tests geprüft werden [Matthews, 1996; Altman, 1996; Altman, 2003]. Die Ergebnisse von getrennten Tests für den Interventionseffekt in jeder Subgruppe liefern keine Information darüber, ob sich der Interventionseffekt zwischen den Subgruppen unterscheidet.

1.2.5.2.6. Fehler zweiter Art

Aufgrund hoher statistischer Power weisen MAs bei der Erkennung von Subgruppenunterschieden größere Potenziale auf als Primärstudien. Dies gilt verstärkt bei MAs, die IPDs bearbeiten und deren statistische Heterogenität gering ist. Oft wird dennoch die Subgruppen-Analyse durch die kleine Zahl von Primärstudien in MA-APDs, durch die geringe Anzahl von Patienten in MA-IPDs und durch die hohe Heterogenität des Effekts eingeschränkt. In dem Fall verfügt der Interaktions-Test über niedrige statistische Power um den „wahren" Unterschied zwischen den Strata zu entdecken (sogenannter β-Fehler). Diese Einschränkung gilt umso mehr, wenn Stratifizierungen nach mehr als einer Primärstudien-Ebenen-Variablen beabsichtigt werden, da die Anzahl von Primärstudien pro Stratum sich dadurch verringert. Demzufolge sollen negative Testergebnisse bei geringer Datenmenge und/ oder großer Heterogenität in der Regel mit Vorsicht interpretiert werden.

1.2.5.2.7. Fehler erster Art

Die effiziente Nutzung von Ergebnissen kostenintensiver RCTs verstärkt die Neigung vieler Prüfärzte [Wang, 2007], mehrere Subgruppen-Analysen und Kovariate-Adjustierungen zu unternehmen. Man erhofft sich daraus, empirische Informationen für die Patientenversorgung und die zukünftige, experimentelle und nicht-experimentelle Forschung zu erlangen [Wang, 2007]. Allerdings erhöht sich beim multiplen Testen auf Unterschiede zwischen den Strata innerhalb einer MA die Wahrscheinlichkeit einen α-Fehlers zu begehen, d.h. signifikante Ergebnisse über den Unterschied zwischen der Behandlungs- und der Kontrollgruppe zu finden, wenn in Wahrheit jedoch kein Unterschied besteht. Die Irrtumswahrscheinlichkeit, mindestens einen falsch-positiven Subgruppenunterschied zu finden, übersteigt 50% bei der Durchführung von 14 Tests bei demselben Signifikanzniveau von 5%. Demzufolge kann empfohlen werden, eine sehr begrenzte Zahl von Subgruppen-Analysen durchzuführen und gegebenenfalls eine Adjustierung des Signifikanzniveaus, z.B. nach der Bonferroni-Methode, vorzunehmen [EMEA, 2002; Schulz, 2005].

1.2.5.2.8. Stratifizierte Randomisierung

Bei stratifizierter Randomisierung werden die Studienteilnehmer erst in Strata aufgeteilt, dann werden sie nach einer für jedes Stratum unabhängigen Randomisierungsliste den Studieninterventionen zugeteilt. Stratifizierte Randomisierung kann zur Überprüfung einer

markanten Hypothese zur klinischen Heterogenität des Interventionseffekts benutzt werden. Stratifizierte Randomisierung kann allerdings auch zur Reduzierung von Bias verwendet werden. Zum Beispiel, ist es für das Studienzentrum in multizentrischen RCTs üblich vor der Randomisierung zu stratifizieren, um für mögliche Unterschiede bezüglich der Durchführung der Studie in den Zentren, z.B. aufgrund von unterschiedlicher Expertise der Prüfärzte oder technischer Ausstattung der Zentren, bei der Auswertung zu adjustieren [Kraemer, 2005]. Außerdem kann stratifizierte Randomisierung bei kleinen Primärstudien zur Stärkung statistischer Power und damit zur Verringerung der erforderlichen Fallzahl führen [Kernan, 1999].

1.2.5.2.9. Diskrepanz des Effekts in der Gesamtpopulation und in den Subgruppen

Die Interpretation der Diskrepanz des Interventionseffekts bezüglich der Lage, der Größe und der Varianz in der Gesamtpopulation auf der einen Seite und in den Subgruppen auf der anderen Seite kann nur mit großer Unsicherheit betrieben und soll weder unter- noch überschätzt werden. In der Regel soll jede Diskordanz zugunsten des durchschnittlichen Interventionseffekts entschieden werden, zumal Ergebnisse zu Subgruppen-Analysen eher den Fehlern erster und zweiter Art exponiert sind.

1.2.5.2.10. Aussagekraft und Stellenwert von Subgruppen-Analysen

Aussagekräftige Subgruppen-Analysen können nur validen Primärstudien entnommen werden. Daher ist eine adäquate Subgruppen-Analyse mittels IPDs aus allen RCTs guter Qualität optimal.

Subgruppen-Analysen werden oft in einzelnen RCTs und in MAs durchgeführt. Sofern sie allerdings nicht den in den Abschnitten 1.2. 5.2.3 bis 1.2.5.2.7 beschriebenen methodischen Anforderungen entsprechen, sind sie ausschließlich für die Generierung von Hypothesen für kommende RCTs zu verwenden.

Aus dem Blickwinkel der Evolution wissenschaftlicher Erkenntnisse darf die Rolle von Subgruppen-Analysen als ein Hypothesen extrahierendes Verfahren nicht unterschätzt werden. Die gängige Praxis, Evidenz niedriger Qualität und nicht selten verzerrte Expertenmeinungen für die Hypothesenentwicklung bezüglich klinischer Heterogenität zu

verwenden, soll vermieden werden. Stattdessen sollen die Ergebnisse der Subgruppen-Analysen in MAs von RCTs für die Hypothesengenerierung verwendet werden.

In der einschlägigen Literatur, einschließlich des bekannten „Consumer's Guide to Subgroup Analysis" [Oxman, 1992], wird empfohlen, die Konsistenz von Ergebnissen der Subgruppen-Analyse in weiteren Primärstudien zu prüfen, was allerdings in Widerspruch zur Evidenz-Basierung bei der a priori-Bestimmung der Subgruppen-Analyse stehen kann. Die Forderung nach „biologischer Plausibilität" für Subgruppen-Analyse [Oxman, 1992], die weder klar konzeptualisiert noch detailliert operationalisiert ist, wird für die Stärkung des Kausalitäts-nachweises erhoben. Sie birgt allerdings die Gefahr, potenziell verzerrte Evidenz, z.B. Expertenmeinung, Tier-Experimente, dafür zu verwenden.

1.2.5.3. Meta-Regression

Bei der Meta-Regression fungieren die Effektgrößen von Primärstudien als abhängige Variable und eine Kovariable (univariate Meta-Regression) bzw. mehrere Kovariablen (multivariate Meta-Regression) auf Primärstudien-Ebene als unabhängige Variable (Prädiktoren). Die Regression unabhängiger Variablen auf die abhängige Variable erfolgt gewichtet. Da Primärstudien-Ebene-Variablen bei dieser Art der Regressions-Analyse einbezogen werden, werden sie Meta-Regression genannt. Effektgrößen in diversen Maßen (RR, OR, Risikodifferenz) und in transformierten Skalen (z.B. logarithmierte Maße), und Prädiktoren in beliebigen Skalen (binär, kategorial, stetig) können in die Regression einbezogen werden.

1.2.5.3.1. Gewichtung in Meta-Regression

Die Gewichtung der kleinsten Quadrate im Meta-Regressions-Modell kann im FEM, d.h. nach dem Kehrwert der Varianz innerhalb der Primärstudien [Hedges, 1995; Sutton, 1998], oder im REM, d.h. nach dem Kehrwert der Varianz innerhalb der Primärstudien plus der Varianz zwischen den Primärstudien [Cooper, 1994; Sutton, 1998], erfolgen. Meta-Regression mit Gewichtung nach dem REM wird auch „Random-Effects-Regression" genannt und schließt einen Faktor für durch Prädiktoren unerklärte Variabilität ein. Dieser Faktor wird im Meta-Regressions-Modell mit Gewichtung nach dem FEM nicht eingebaut [Sutton, 1998]. Dabei wird oft angenommen, dass die Gesamtvariabilität durch die

Kovariablen erklärbar ist. Da eine solche Annahme als stark einzustufen ist (s. Abschnitt 1.2.4.2), soll die Random-Effects-Regression vorgezogen werden.

Bei der Verwendung von Software zur statistischen Auswertung ohne nachvollziehbare Programmierung soll nach dem Gewichtungsverfahren auch für die Standardfehler der Koeffizienten von Meta-Regression recherchiert werden. Generalisierte Lineare Modelle (GLM) sind flexibel nicht nur in Bezug auf das Gewichtungsmodell, sondern auch hinsichtlich der Verteilungsauswahl, z.B. der t-Verteilung bei wenigen Daten, und Skalentransformierung, z.B. logarithmische Skala bei schief verteilten Daten [Dobson, 2001].

1.2.5.3.2. Automatisierung der Selektion der Kovariablen

Für den Zeitpunkt und die Vorgehensweise zur Selektion, Erhebung und Aufbereitung der Kovariablen gelten bei Meta-Regression alle oben genannten Ausführungen zur Subgruppen-Analyse (s. Abschnitte 1.2.5.2.3 und 1.2.5.2.4). Selektions-Bias bei der Auswahl der Kovariablen kann in verzerrter Schätzung der im Modell aufgenommenen Parameter resultieren. Bei Mangel an Evidenz-basierter Auswahl der Kovariablen kann auf statistikgeleitete Selektion von Prädiktoren zurückgegriffen werden. Die Anwendung von Verfahren zur Vorwärts-, Rückwärts- oder schrittweisen Selektion von Prädiktoren in multivariater Regression lehnt sich oft an p-Werte an, weist erhebliche mathematische Probleme auf und soll möglichst vermieden oder kritisch betrachtet werden [Derksen, 1992; Harrel, 2001]. Die medizinische Begründung der Prädiktorenauswahl ist der statistischen Selektion vorzuziehen. Eine Simulationsstudie schreibt den Rückwärts-Eliminations-Verfahren bessere Eigenschaften zu als den Vorwärts-Einschluss-Verfahren [Harrel, 2001]. Eine weitere Simulationsstudie zeigte, dass der „Least Absolute Shrinkage and Selection Operator“-Ansatz (Lasso-Ansatz) besser funktioniert als die Regression mit und ohne statistische Selektion bei „small to moderate number of moderate-sized effects“ [Tibshirani, 1996, S. 286].

Die Aufnahme von Interaktionen zwischen den Kovariablen im Regressions-Modell orientiert sich am Ein- oder Ausschluss betreffender Prädiktoren in demselben Modell. In additiven Regressions-Modellen sind einzelne Kovariablen additiv und Interaktionsterms multiplikativ. Daher kann das Einschließen von Kovariablen und ihren Interaktionsterms in einem solchen Modell die Interpretation erschweren, was der Fall ist in gängiger Software für automatisiertes Selektionsverfahren, z.B. vorwärts, rückwärts oder schrittweise [Chen, 2007]. Dennoch sind Interaktionsterms und ihre Prädiktoren ins Modell einzuschließen. Die Ansätze

unterscheiden sich bezüglich der Reihenfolge beim Einschluss von Kovariablen und Interaktionstermen im Modell [Chipman, 1996; Nelder, 1998; Chen, 2007].

1.2.5.3.3. Fehler zweiter und erster Art

Im Vergleich zur Subgruppen-Analyse weist die Meta-Regression meistens eine geringfügig größere statistische Power auf und kann mehrere Heterogenitätsquellen (Primärstudien-Ebene-Variablen) gleichzeitig untersuchen. Zur Einhaltung des Signifikanzniveaus bei automatisierter Prädiktorenselektion kann ein Schwellenwert für den p-Wert festgelegt werden, der den Ein- oder Ausschluss der Kovariablen im Regressions-Modell bestimmt [Chen, 2007].

1.2.5.3.4. Baseline-Risiko als Prädiktor

Das Baseline-Risiko kann durch die Patientenmerkmale am Studienanfang ermittelt werden. Als Surrogat für das Baseline-Risiko wurden die beobachtete Ereignisrate des Studienendpunktes in der Kontroll- und in der Interventionsgruppe vorgeschlagen. Insbesondere die Ereignisrate in der Kontrollgruppe, als ein potenzieller Indikator für klinische Heterogenität, wird als Prädiktor mit der Effektgröße auch mittels Meta-Regression untersucht [Schmid, 1998; Arends, 2000; Dohoo, 2007]. Aufgrund der Abhängigkeit der Schätzung des Interventionseffekts von der Ereignisrate in der Kontroll- und Interventionsgruppe kann das Einbeziehen einer der beiden Ereignisraten als eine Kovariable im Meta-Regression-Modell zur massiven Verzerrung durch sogenannte „Regression to the Mean“ führen [Senn, 1994; McIntosh, 1996; Thompson, 1997; Walter, 1997]. Alternativ wurden Bayesianische Ansätze und Maximum-Likelihood-Methoden (via EM-Algorithmus) für die Problematik vorgeschlagen [McIntosh, 1996; Thompson, 1997].

1.2.5.3.5. Ecological-Fallacy

Eine weitere Einschränkung der Meta-Regression ist die sogenannte „Ecological-Fallacy“ [Morgenstern, 1982]. Sie bezieht sich auf die Regression durchschnittlicher Werte eines Patientencharakteristikums, z.B. Mittelwert der Low-Density-Lipoproteine in jeder Primärstudie, auf die Effektgröße, wobei die Variablität des patientenbezogenen Prädiktors innerhalb der Primärstudien keine Berücksichtigung findet. Die Übertragung von Ergebnissen

aus aggregierten Daten auf individuelle Patienten ist problematisch, d.h. die Ergebnisse einer Primärstudie, die Patienten mit einem Durchschnittsalter von 50 Jahren aufweist, nicht ohne weiteres auf 50-jährige Patienten angewendet werden können [Lau, 1998, S. 125]. Individual-Ebene-Variablen dürfen in aller Regel nur mit individuellen Patientendaten untersucht werden. Da die Regression eines patientenbezogenen Prädiktors mittels aggregierter Daten niedrige statistische Power aufweist, kann ein signifikanter Regressionskoeffizient diesbezüglich, einen in weiteren Daten zu überprüfenden Hinweis auf einen möglichen Zusammenhang liefern [Lambert, 2002]. Da die methodische Qualität eine Primärstudien-Ebene-Variable ist, wird die Ecological-Fallacy hier nicht relevant.

1.2.5.3.6. Multi-Kolinearität und Regressionsdilution

Wenn zwei Prädiktoren eine hohe lineare Korrelation aufweisen, große Veränderung der Regressionskoeffizienten beim Ein- oder Ausschluss eines der beiden Prädiktoren beobachtet wird oder große Standardabweichungen der Koeffizienten gesehen werden, kann dies auf das Problem der Multi-Kolinearität hindeuten [Song, 2001]. Bei kleiner Datenmenge bezüglich Zahl und Größe der Primärstudien, kann als Maßnahme gegen Multi-Kolinearität entweder einer der betreffenden Prädiktoren aus dem Modell genommen oder es können mehr Daten generiert werden [Van den Poel, 2004].

Regressionsdilution entsteht bei hoher Variabilität einer Kovariablen, die dem Zufall, den Messfehlern oder der statistischen Varianz zuzuschreiben ist [Carroll, 1995]. Instabile Prädiktoren, wie der Blutdruck, sind für Regressionsdilution anfällig.

1.2.5.4. Graphische Methoden

Es existieren mehrere graphische Methoden zur Exploration der Heterogenität in MAs. Obwohl die Graphiken einfach zu erstellen sind, kann eine visuelle Verzerrung bei ihrer Interpretation aber oft nicht ausgeschlossen werden. Formale statistische Methoden sollen beobachtete graphische Muster evaluieren. Anhand von graphischen Methoden können die Effektgrößen der Primärstudien gegenüber einer Variablen auf Primärstudien-Ebene dargestellt werden. Im Folgenden werden oft verwendete graphische Methoden vorgestellt.

1.2.5.4.1. Forest-Plot

Der Forest-Plot ist die am meisten verwendete graphische Methode in MAs. Sie zeigt die Punktschätzer und die entsprechenden Konfidenz-Intervalle der Effektgrößen von Primärstudien sowie der zusammengefassten Effektgröße. Die Ergebnisse der Primärstudien können nach unterschiedlichen Primärstudien-Ebenen-Variablen geordnet werden, wobei die Punktgröße proportional zum Stichprobenumfang der Primärstudie dargestellt wird. Die Ordnung kann nach ihrer Effektgröße, ihrem Stichprobenumfang oder ihrer methodischen Qualität erfolgen. Eine solche Darstellung kann auf einen möglichen Zusammenhang der Intervention mit einer Primärstudien-Ebenen-Variablen hindeuten. In einer sogenannten kumulativen MA werden die Primärstudien nach einer spezifizierten Primärstudien-Ebenen-Variable geordnet, dann werden die ersten zwei Primärstudien kombiniert und sukzessive die dritte, vierte, usw. hinzugenommen. Geordnet nach dem Veröffentlichungsjahr ist der früheste Zeitpunkt, zu dem die Kumulation der Evidenz statistisch signifikant wird, zu erkennen [Song, 2001].

1.2.5.4.2. L´Abbé-Plot

Der L´Abbé-Plot ist ein Streudiagramm, bei dem jeder Punkt eine Primärstudie vertritt, wobei auf der x-Achse die Ereignisrate in der Kontrollgruppe und auf der y-Achse die Ereignisrate in der Interventionsgruppe dargestellt werden. Der Primärstudienarm (Kontroll- oder Interventionsgruppe), bei dem der größere Anteil der Heterogenität liegt, kann durch diese Darstellung erkannt werden. Die Punktgröße wird proportional zum Stichprobenumfang der Primärstudie dargestellt. Falls die Ereignisrate in der Kontroll- und Interventionsgruppe gleich ist, wird die Studie mit einem Punkt auf der soliden diagonalen Linie dargestellt werden. Falls eine Studie unter dieser Linie dargestellt wird, deutet dies auf eine niedrige Ereignisrate in der Interventionsgruppe hin [Song, 2001].

1.2.5.4.3. Funnel-Plot

Eine graphische Darstellung des Stichprobenumfangs der Primärstudien gegenüber deren Effektgrößen soll ein Trichterform (Funnel-Plot) aufweisen. Wenn der Funnel Plot asymmetrisch ist, kann das auf einen Publication-Bias aufgrund des Fehlens kleiner Primärstudien mit nicht signifikanten Ergebnissen hindeuten. Allerdings kann ein asymmetrischer Funnel-Plot auch auf methodische oder klinische Heterogenität der

Primärstudien zurückzuführen sein [Song, 2001]. Die adjustierte Rangkorrelationsmethode von Begg und Mazumdar [Begg, 1994] und die Methode der linearen Regression von Egger und Kollegen [Egger, 1997] sind statistische Methoden zum Detektieren der Asymmetrie von Funnel-Plots und die sogenannte Trim-and-Fill-Methode von Duval und Tweedie dient der Berechnung einer für Publication-Bias adjustierten zusammengefassten Effektgröße [Duval, 2000].

1.2.6. Ansätze zur Reduzierung der Heterogenität in Meta-Analyse

1.2.6.1. Veränderung der Skala der Endpunktmessung

Werden unterschiedliche Heterogenitäts-Maße bei unterschiedlichen Skalen der Endpunktmessung beobachtet, wird empfohlen die Skala mit der niedrigsten Heterogenität für die MA zu verwenden. Eine Untersuchung der Heterogenität bei 125 MAs zeigt, dass die Synthese von OR und von Risikodifferenzen bei den meisten MAs zu ähnlichen Schlussfolgerungen führte. Allerdings wurde statistisch signifikante Heterogenität öfter bei Risikodifferenzen als bei ORs beobachtet [Engels, 2000].

1.2.6.2. Ausschluss von Ausreißern

Durch den sukzessiven Ausschluss von Primärstudien mit extremen Effektgrößen (Ausreißern) kann man eine statistisch signifikante Heterogenität eliminieren. Allerdings stellt der Ausschluss von Primärstudien auf der Grundlage ihrer Ergebnisse und nicht auf der Grundlage ihrer Qualität, eine erhebliche Verzerrungsquelle der MA dar.

1.3. Kritische Bewertung von randomisierten kontrollierten Studien

Die kritische Bewertung der methodischen Qualität von Primärstudien soll ein essenzieller Bestandteil von SRs und MAs sein [Higgins, 2006], es gilt der Grundsatz: „*many bad studies dont't make a good one*“ [Parmigiani, 2002, S. 125]. Es ist müßig, sich mit der Qualitätsbewertung zu begnügen, wenn nicht auch deren Ergebnisse bei der Auswertung von SRs berücksichtigt werden. Der Einschluss von RCTs mit niedriger methodischer Qualität in eine MA kann zur Entwertung von deren Schlussfolgerungen führen („garbage in,

garbage out“) [Pildal, 2007]. Der Anteil der SRs, die ihre Studien einer Qualitätsbewertung unterzog, unterscheidet sich stark zwischen den medizinischen Subdisziplinien (s. Tab. 6). Qualitätsbewertung wurde öfter in Cochrane SRs als in Fachzeitschriften-basierten SRs durchgeführt [Jadad, 1998; Moher, 1999b; Jüni, 2000; Jadad, 2000; Moja, 2005; de Craen, 2005; Collier, 2006; Jørgensen, 2006; Moher, 2007; Delaney, 2007]. Dies kann mit dem uneingeschränkten Publikationsumfang, die die Cochrane SRs genießen, zusammenhängen. Allerdings findet die Qualitätsbewertung bei der Auswertung oder Interpretation von SRs weiterhin selten Berücksichtigung (bei jeweils 12%, 51% bzw. 53% der SRs in den Untersuchungen von Moher, 1999b; Moja, 2005 und de Craen, 2005], dabei schneiden die Cochrane SRs nicht besser ab als Fachzeitschriften-basierte SRs.

Obwohl RCTs die beste Evidenzqualität zur Wirksamkeit von Interventionen liefern, sind sie vor systematischen Fehlern (Bias) nicht gefeit. Thomas C. Chalmers ist ein anerkannter Pioniere der Thematisierung von Bias in RCTs in den 1970ern [Berk, 1999]. Moher und Kollegen definierten die Qualität von RCTs als *„the confidence that the trial design, conduct, and analysis has minimized or avoided biases in its treatment comparisons“* [Moher, 1995, S. 63]. Allerdings bleiben die Ansätze zur Bewertung von Qualität primärer Studien im Rahmen von SRs/ MAs umstritten.

Tab. 6 Epidemiologie der Qualitätsbewertung in SRs und MAs [Eigene Darstellung]

Autor, Jahr	Medizingebiet	Quelle der SRs/ MAs	Anzahl der SRs/ MAs	Anteil der SRs/ MAs mit Qualitäts-bewertung (%)
Mulrow, 1987	Divers	Fachzeitschriften	50	2
Jadad, 1996a	Analgetische Interventionen	Fachzeitschriften und Datenbank	80	39
Jadad, 1998	Divers	Datenbanken	75	64
Moher, 1999b	Divers	Fachzeitschriften und Datenbank	240	48
Petticrew, 1999	Divers	Datenbank	480	52
McAlister, 1999	Divers	Fachzeitschriften	158	9
Jüni, 2000	Divers	Fachzeitschriften	133	41
Jadad, 2000	Asthma	Fachzeitschriften und Datenbanken	50	28
Mackay, 2003	Psychotherapie und Beratung	Datenbanken	255	46
Bader, 2004	Zahnheilkunde	Datenbanken	131	54
Huang, 2004	Infektion mit Helicobacter pylori	Datenbanken	38	26
Moja, 2005	Divers	Fachzeitschriften und Datenbank	965	89
de Craen, 2005	Divers	Fachzeitschriften und Datenbank	73	88
Golder, 2006a	Unerwünschte Ereignisse	Datenbanken	256	41
Shea, 2006	Skelettmuskelkrankheiten	Datenbank	57	72
Collier, 2006	Dermatologie	Datenbanken	38	79
Mignini, 2006	Tiergrundlagenforschung	Datenbanken	30	50
Moher, 2007	Diverse	Datenbank	300	67
Delaney, 2007	Intensiv- und Notfallmedizin	Datenbanken	128	48

1.3.1. Komponenten der Fragestellung

1.3.1.1. Patientenkollektive

Bei der Bewertung der Patientenkollektive, die in einer RCT eingeschlossen sind, sollen Änderungen in der Nomenklatur (z.B. Moynihan's Syndrom/ Dyspepsia/ Magen- und Duodenalulcus, Angina Pectoris/ Koronare Herzkrankheit) und Änderungen in der ätiologischen Klassifikation der Krankheit (z.B. Magen- und Zwölffingerdarmgeschwür als idiosynkratische/ stressbedingte/ gewürzebedingte/ bakterielle Krankheit) beachtet werden [Feinstein, 2001]. Weiterhin soll die Validität der verwendeten diagnostischen Verfahren zur Bestimmung der Patientenkollektive überprüft werden (z.B. Tuberkulin-Hauttest versus Interferon-Gamma-Test nach der Infektion mit Mycobakterium tuberkulosis [Menzies, 2007]). Modifikationen der diagnostischen Kriterien sollen berücksichtigt werden. Die Einführung der kardialen Troponine für die Diagnose von Myokardinfarkt [Antman, 2000] und die Empfehlung eines reduzierten Nüchternglukosewerts für die Diagnose von Diabetes Mellitus [DECODE, 1998] können zum Einschluss von Patienten mit niedrigem kardiovaskulärem Risiko in RCTs führen.

Die Neigung, gesündere Patientenkollektive für RCTs zu rekrutieren und einzuschließen, stellt die Repräsentativität der Patientenkollektive in Frage. Die COURAGE-Studie [Boden, 2007] kann als eins der aktuellen Beispiele für eng gefasste Einsschlusskriterien gesehen werden. Weitgefasste Einschlusskriterien, wie sie in der SANAD-Studie [Marson, 2007] zu beobachten sind, sollen angestrebt werden.

1.3.1.2. Interventionsgruppen

1.3.1.2.1. Standardisierung nicht-pharmakologischer Intervention

Als schwierig gelten im Rahmen von kontrollierten klinischen Studien die Standardisierung nicht-pharmakologischer Interventionen, z.B. Chirurgie, Psychotherapie, Chirotherapie, und komplexer Interventionen, z.B. Prävention und Gesundheitsförderung in sozialen Settings, Organisationsprogramme. Das UK „Medical Research Council" entwickelte ein „Framework for the Development and Evaluation of Randomised Controlled Trials for Complex Interventions" [MRC, 2000]. Die „Komplexität" einer Intervention nimmt zu mit Erhöhung der

Schwierigkeiten bei der Bestimmung, Isolierung, Optimierung und Normierung, also der Standardisierung der „aktiven“ Komponenten [Campbell, 2000; Hawe, 2004; Murchie, 2007; Peters-Klimm, 2007].

Die in klinischen Studien ermittelten Effekte von durch soziale Interaktionen (Inter-individuen oder Inter-Gruppen) oder professionelle Interaktionen (z.B. multidisziplinäre Prüfärzte) modifizierbaren Interventionen können zur Kontamination zwischen den Interventionseffekten führen. Diese Kontamination resultiert oft in der Unterschätzung des Nutzens und der Überschätzung des Risikos der experimentellen Intervention und ist generell mit zusätzlichen Unsicherheiten bei der Interpretation der Ergebnisse verbunden [Murphy, 2006; Torgerson, 2001]. Bei erwarteter hoher Rate von inter-individueller Kontamination können Cluster-RCTs verwendet werden, wobei die Randomisierungs- und Auswertungseinheit das Setting oder die jeweilige Teilnehmergruppe ist [Hahn, 2005; Eldridge, 2004; Donner, 2001]. Allerdings weisen Cluster-RCTs im Vergleich zu Individuum-basierten RCTs in der Regel höhere Fallzahl, höhere Rekrutierungshürden und höhere Studienaustrittsraten auf. Zumal eine Cluster-RCT durch unvollständiges Allocation Concealment und mangelnde Rekrutierbarkeit aller Mitglieder eines Clusters vor der Randomisierung für Selection-Bias empfänglich ist. Pseudo-Cluster-RCTs sind Studien bei denen lediglich ein großer Anteil der Mitglieder des Clusters, die der experimentellen Intervention zufällig zugewiesen sind, diese Intervention erhalten und wenige Mitglieder die Kontrollintervention bekommen. Dies findet vice versa in den Kontroll-Clustern statt. Pseudo-Cluster-RCTs wurden zur Reduzierung der Fallzahl (Erhöhung der Trial-Effizienz), Verbesserung der Rekrutierung und des Allocation Concealment vorgeschlagen [Teerenstra, 2006; Borm, 2005].

Expertise-basierte RCTs wurden für verbesserte Standardisierung der durch Erfahrungen von Prüfärzten beeinflussbaren Interventionen, z.B. Chirurgie, vorgeschlagen [Devereaux, 2005], wenngleich die darin geforderte große Expertise zu begrenzter externer Validität führen kann. Statistisches Monitoring und Rückmeldung für Prüfärzte bezüglich nicht standardisierter Intervention (z.B. Psychotherapie) wurde als eine kostengünstige Alternative zu Schulungen und Manualen vorgeschlagen [Iberg, 1991].

Nicht-pharmakologische Interventionen, die in Clusters durchgeführt werden, z.B. Programme zur Gesundheitsförderung in Schulen, oder die durch die Expertise der Prüfärzte variieren (z.B. Psychotherapie), sollen durch Schulungen, Audits und externe Gutachter standardisiert und evaluiert werden.

1.3.1.2.2. Selektion der Kontroll-Gruppe

Der Biometriker Francis Galton schlug bereits 1872 vor, eine Kontrollgruppe einzusetzen, um die Wirksamkeit von Gebeten zu ermitteln [Dehue, 2000]. Ein gelegentlicher Einsatz von Kontrollgruppen kann bis ins 6. Jh. v. Chr. (im Buch Daniel) zurückverfolgt werden. Die systematische Verwendung einer Kontrollgruppe als methodische Anforderung an Experimente wurde erst seit Anfang des 20. Jahrhunderts beobachtet [Dehue, 2000].

Die Selektion einer „angemessenen" Kontrollintervention ist die unabdingbare Voraussetzung für die Validität jedes kontrollierten Experiments. Die Validität einer klinischen Studie mit einer ineffektiven Kontrollintervention kann weder mittels weiterer höchstqualitativer Dimensionen des Studiendesigns noch anhand aufwändiger, methodischer und biometrischer Analyse-Verfahren restauriert werden. Die „International Conference on Harmonisation" entwickelte einen ausführlichen und differenzierten Leitfaden zur Selektion von Kontrollgruppen in klinischen Studien [ICH, 2000].

Seit langem werden überwiegend ethische und regulatorische Debatten geführt werden über den Einsatz von keiner Intervention, Placebo/ Schein-Intervention oder aktiver Intervention in der Kontrollgruppe [Albin, 2005; Moerman, 2002; Miller, 2002; Temple, 2000; Ellenberg, 2000]. Ausführliche Diskussionen wurden geführt, wenn die Evidenzlage bezüglich der Nutzen-Risiko-Bilanz einer als Standard zu bezeichnenden aktiven Kontrollintervention schlecht ist oder bislang keine aktive Intervention existiert. Die Selektion der Kontroll-Gruppe wird engagiert debattiert in der Chirurgie wegen der zu bedenkenden Risiko-Nutzen-Bilanz von Schein-Prozeduren, wie im Falle der Parkinson-Krankheit [Frank, 2005; Albin, 2005; Angelos, 2007]. Einen ebenso ausführlichen Diskurs gibt es über die Zweitlinienkrebstherapie, weil für das Zulassungsverfahren in Europa „keine Intervention" als Kontrollintervention akzeptiert ist, diese Regelung in den USA dagegen abgelehnt wird [Ellenberg, 2000]. Allerdings führt diese Debatte unmittelbar zur relativen Vernachlässigung eines für die methodische Validität der Studien unabdingbaren konzeptuellen und Empirie-gestützten Diskurses über die „Angemessenheit" einer Kontrollintervention.

1.3.1.2.2.1. Assay-Sensitivity

Das Potenzial einer Studie, eine effektive Intervention von weniger effektiver oder nicht effektiver Intervention zu unterscheiden, wird als „Assay-Sensitivity“ bezeichnet. Dies hängt von der Validität des Studiendesigns, einschließlich einer konkurrenten angemessenen Kontrollgruppe, und der statistischen Power der klinischen Studien ab. Falls eine aktive Kontrollintervention verfügbar ist, sichert ein paralleles dreiarmiges Studiendesign (Placebokontroll-, Aktivkontroll-, und Experiment-Gruppe) sowohl in Überlegenheits-, Nicht-Überlegenheits- und Äquivalenz-RCTs die Assay-Sensitivity am besten [ICH, 2000; Kieser, 2007]. Dreiarmige klinische Studien sind allerdings weiterhin eine Seltenheit, wobei die meisten wurden für die Evaluation onkologischer Therapeutika durchgeführt [Bidoli, 2007; Baum, 2002; Le Chevalier, 1994].

Die Verwendung von nicht in der klinischen Studie (externe Kontrollen) oder von im Zeitraum vor der Studie (historische Kontrollen) eingeschlossenen Patienten als Kontrollgruppe schwächt die Validität des Vergleichs ab und ist nur in begrenzten und im Vorfeld zu begründenden Situationen zulässig [ICH, 2000].

1.3.1.2.2.2. Inaktive Kontrollintervention

Obwohl „The Powerful Placebo“ schon 1955 im Rahmen der ersten bekannten und mit Recht kritisierten MA zu therapeutischen Interventionen thematisiert wurde [Beecher, 1955; nach: Egger, 2005; Boussageon, 2006], besteht bedauernswerterweise bislang keine eindeutige Definition, weder von Placebo, noch vom Placebo-Effekt [Meissner, 2007; Hrobjartsson, 2004; Macedo, 2003; Gotzsche, 1994], noch vom Nocebo-Effekt [Olshansky, 2007; Barsky, 2002; London, 2002]. Placebo-kontrollierte klinische Studien ohne effektive und erhaltende Verblindung laufen Gefahr, dass die Effektgröße in der Placebogruppe durch den Placebosuspekt unterschätzt und dass anderweitige Nutzen oder Risiken in der Placebogruppe zu vermehrtem Placebosuspekt führen. Herstellung, Aufrechterhaltung und Überprüfung der Verblindung von Patienten, Prüfärzten und weiteren an der Patientenversorgung beteiligten Studienpersonals ist für die Minimierung eines möglichen Placebo-Effekts entscheidend. Daher soll ein Placebo der experimentellen Intervention in Bezug auf die äußere Erscheinung (Größe, Form, Farbe, Gewicht, Konsistenz, Geschmack, Geruch) und das Administrationsregime (Dosis, Frequenz, Weg) möglichst ähneln, sich jedoch hinsichtlich der effektiven Komponente und des Wirkungsmechanismusses weitgehend unterscheiden. Diese Unterscheidung setzt das Vorhandensein von Evidenz

oder einer Hypothese über die effektive Komponente und den Wirkungsmechanismus der experimentellen Intervention voraus [Brinkhaus, 2008].

Die Verblindung inerter Kontrollintervention erwies sich oft als nicht trivial bei Interventionen, die nicht-pharmakologisch sind, z.B. Chirurgie [Boutron, 2007], bei Interventionen mit unsicherem Wirkungsmechanismus, z.B. Akupunktur, Wirbelsäulen-Manipulation [Dincer, 2003; Ernst, 2001], mit suboptimaler Standardisierung, z.B. Bewegung, oder mit komplexen Multi-Interventionen, z.B. Heimpflege. Weiterhin zeigte eine RCT, die Schein-Akupunktur mit Placebo-Pillen bei Patienten mit Armschmerzen verglich, eine signifikante, wenn auch kleine und auf patientenberichteten Endpunkte eingeschränkte, Überlegenheit der Schein-Akupunktur [Kaptchuk, 2006]. Die „Placebo Quality Checklist“, begleitet von einem Fragenraster zum Monitoring der Verblindung, wurde vor kurzem publiziert [Brinkhaus, 2008].

Obwohl der Verfasser den Einsatz einer Nicht-Intervention oder eines Placebos als Kontrollintervention [Leber, 2000], insbesondere in dem dafür auszuwählenden dreiarmigen Trial-Design, in den meisten Evaluationen neuer Interventionen sowohl methodisch als auch ethisch für vertretbar hält, kann ein zu häufiger Einsatz von inaktiven Kontrollinterventionen beobachtet werden. Eine SR, die 136 klinische Studien zu neuen Therapeutika gegen multiple Myeloma berücksichtigte, fand heraus, dass inaktive Interventionen bei 60% der industriefinanzierten, klinischen Studien und bei 21% der nicht-industriefinanzierten klinischen Studien als Komparatoren fungierten [Djulbegovic, 2000]. Aus Vermarktungsmotiven wurden im renommierten „New England Journal of Medicine“ in drei klinischen Studien für Patienten mit diabetischer Nephropathie ACE-Hemmer, statt mit verfügbaren aktiven Kontrollinterventionen, mit Placebos verglichen [Hostetter, 2001; Parving, 2001; Brenner, 2001; Lewis, 2001].

1.3.1.2.2.3. Aktive Kontrollintervention

Interessenbelastet und als wissenschaftliche Verfehlung zu bewerten ist die Verwendung einer pharmakodynamisch impotenten oder mit erhöhten Risiken verbundenen aktiven Kontrollintervention. Fokussiertes Augenmerk soll gelegt werden auf aktive Kontroll-interventionen mit suboptimaler Dosierung [Woods, 2005; Safer, 2002; Geddes, 2000], mit ineffektiven [Schroeder, 2004; Johansen, 1999], oder riskanten Ingredienzen [Ahmad, 1992], mit ungeeignetem Administrationsweg [Johansen, 1999] oder Darreichungszeitpunkt [Safer, 2002; Christiansen, 1996]. Eine SR, die 30 empirische Studien einschloss, fand heraus, dass

von der pharmazeutischen Industrie finanzierte klinische Studien und MAs vierfach häufiger Ergebnisse zugunsten der Sponsorintervention zeigten als von anderen finanzierte klinische Studien und MAs. Unangemessene Kontrollintervention und Publication-Bias wurden als mögliche Gründe für diese Verzerrung beobachtet [Lexchin, 2003].

Die Äquivalenz oder Nicht-Unterlegenheit einer experimentellen Intervention gegenüber einer aktiven Kontrollintervention kann erst geprüft werden, wenn eine „Kopf-an-Kopf-Überlegenheit" der aktiven Kontrollintervention gegenüber Placebo oder gegenüber keiner Intervention in mindestens einer vorher durchgeführten zweiarmigen RCT nachgewiesen wurde [Gelfand, 2006; Temple, 2000a]. Die Überlegenheitsannahme neuer Interventionen über Placebo beim Vorhandensein einer Äquivalenz zwischen neuen und alten aktiven Interventionen erwies sich oft als unhaltbar und die Verwendung von indirekten nicht-randomisierten Vergleichen oder historischen Placebo-kontrollierten Studien stellten sich nicht selten als nicht valide heraus [Wang, 2002; Otto, 2002; Leber, 1989].

1.3.1.3. Endpunkte

Bei der Evaluation einer Intervention geht es nicht nur um die Selektion der Kontroll-intervention, sondern auch um die Auswahl der Endpunkte, die man schätzt und testet. Die Zielsetzungen bestimmen dabei die Endpunkte. Die Klassifikation eines Endpunkts kann jedoch nach unterschiedlichen Schemata erfolgen. In Tabelle 7 ist ein vom Verfasser entwickeltes Klassifikationsschema für Endpunkte dargestellt, das 12 Dimensionen beeinhaltet.

Fragestellunggeleitete ist datengeleiteter Selektion der Endpunkte vorzuziehen [Schünemann, 2006]. Daher sollen primäre und sekundäre Endpunkte im Studienprotokoll vor der Studiendurchführung festgelegt werden. Primäre Endpunkte, die im Rahmen von RCTs mit hoher methodischer Qualität erhoben werden, können konfirmatorische Ergebnisse liefern. Vor Studienbeginn bekannte Interventionsrisiken sollen zumindest als Sekundärendpunkt definiert werden (gezielte Exploration). Tertiäre Endpunkte werden lediglich zwecks der ungezielten Exploration oder der Deskription untersucht.

Tab. 7 Klassifikationsschema für Endpunkte in klinischen Studien [Eigene Darstellung]

Nr.	Dimension	Klassen	Implikation
1	Aussagekraft	primär / sekundär / tertiär	Zielsetzungen der Studie
2	Relevanz	relevant / kontrovers / irrelevant: (validiert / nicht validiert)	Orientierung (Surrogate Endpunkte)
3	Objektivität	objektiv / subjektiv: (validiert / nicht validiert)	Objektivität (Patient Reported Outcomes)
4	Stabilität	stabil / instabil	Intra-Individuum-Variabilität
5	Erhebungsperspektive	patientenberichtet / Prüfarzt-geleitet“ / Para-klinisch-erhoben	Relevanz/ Objektivität
6	Ereignishäufigkeit	sehr selten / selten / häufig	Häufigkeit
7	Erhebungszeitpunkte	angemessen / unangemessen	Studiendauer
8	Erhebungsintervalle	Vorher-Nachher / wiederholt mit Äquidistanz / wiederholte Erhebungen ohne Äquidistanz	Messwiederholung
9	Externe Validierung	systematisierte Adjudikation / keine Adjudikation	Externe Validierung
10	Verblindung der Erhebung	effektiv und erhaltend verblindet / effektiv und nicht aufrechterhaltend verblindet / unverblindet	Objektivität
11	Messskala	binär / nominal / ordinal / stetig / Zeit bis zum binären Ereignis	Test- und Schätzverfahren
12	Kombinierte Endpunkte	angemessen / unangemessen	Effizienz / konkurrierende Risiken

Inzwischen besteht in der sich in unterschiedlichen Etablierungsphasen befindenden Evidenz-basierten Medizin (EbM) und im Health Technology Assessment (HTA) ein relativ großer Konsens bezüglich der Verwendung von patientenrelevanten Endpunkten (PREPs) in klinischen Studien. Die wachsende Inanspruchnahme zugänglicher Gesundheitsinformation, die steigende Emanzipationskultur seitens der Patienten [Arora, 2000; Little, 2001; Whelan, 2003] und Diskrepanzen zwischen Patienten und Ärzten bezüglich der Nutzen-Risiko-Abwägungen [Devereaux, 2001; Man-Son-Hing, 2005] führten zum Proklamieren von Konzepten und zu Begriffen wie „patient-oriented-outcomes“ [Berger, 1999], „personal significance“ [Sweeney, 1998] und „patient-centered-outcomes“ [Guyatt, 2004].

Als wenig umstrittene PREP gelten Mortalitäten, patientenrelevante Morbiditäten, Krankheitslasten, z.B. Hospitalisierung, Interventionsbürden, Lebensqualität und patientenberichtete Zufriedenheit einschließlich gewünschter Autonomie [Bastian, 2006; Schünemann, 2006]. Bei den meisten PREPs, mit Ausnahme von Mortalität, ist es schwer, die Relevanz des Schweregrades festzulegen.

Bei der Bestimmung der Relevanz eines Endpunktes ist eine Orientierung an folgenden Empfehlungen möglich: das “Scottish Intercollegiate Guidelines Network“ (SIGN) empfahl

eine Vorabrecherche nach quantitativer und qualitativer empirischer Evidenz über Endpunkte [SIGN, 2008], das „United States Preventive Services Taskforce“ (USPSTF) schlug eine 4-punktige Stufungsskala der „Wichtigkeit“ von Endpunkten vor [Harris, 2001], das „National Institute for Health and Clinical Excellence“ (NICE) in Großbritannien und das „Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen“ (IQWiG) in Deutschland strebt die Einbeziehung von Patientenorganisationen an [Schünemann, 2006; Bastian, 2006]. Weiterhin können die Ergebnisse sogenannter „Views Studies“ [Harden, 2004; Kent, 1996; Dicker, 1995] und systematischer Methode „Influence Diagrams“ [Bravata, 2005; Owens, 1997] zur Berücksichtigung der Patientenperspektive verwendet werden. Relevante Endpunkte können durch Konsensverfahren in jedem medizinischen Fachgebiet und unter Betetiligung der betroffenen Patientenoragnisationen bestimmt werden.

Die Patientenrelevanz eines Endpunktes verringert sich, je stärker er sich an den Belangen der Krankheit, z.B. Senkung pathophysiologischer Werte, der Ärzte oder des Systems, z.B. Mengenerhöhung oder Fertigkeitenverbesserung, orientiert. Er kann dann als Surrogatendpunkt bezeichnet werden. Zahlreiche Beispiele für Surrogattrugschlüsse sind in der Literatur zu finden, wobei die Ergebnisse von klinischen Studien für Surrogat-Endpunkte und für patientenrelevante Endpunkte sich widersprachen [Mühlhauser, 2006; Carter, 2002; Smith, 2002, Albert, 1998]. Die Validierung von Surrogatparametern an PREP gilt als schwierig und setzt in der Regel valide, lange andauernde und große klinische Studien voraus [Buyse, 1998; Molenberghs, 2002; Lassere, 2007].

Als „hart“ werden Endpunkte bezeichnet, die objektiv, relevant und stabil, z.B. Gesamtmortalität, sind. Weiche Endpunkte weisen diese Kriterien nicht auf.

Obwohl die Verwendung wiederholter, äquidistanter Erhebungszeitpunkte eine übliche Praxis in klinischer Forschung darstellt, wurde ihre Effizienz für manche Studien hinterfragt [Overall, 1994; Vickers, 2003]. Durch mathematische Ableitungen unter bestimmten Annahmen wurden „Optimal Time-Points“ (OTPs) für die Zeitwahl der Messungen vorgeschlagen, wobei die OTPs als sehr abhängig vom Zeitverlauf der Korrelation (Kovarianzstruktur) eingestuft wurden [Winkens, 2005]. Auf umfangreiche Literatur zur Analyse longitudinaler Daten [Zeger, 1992; Diggle, 1998; Choi, 2005] kann zurückgegriffen werden.

Die Evaluation von Interventionen bezüglich seltener oder spät eintretender Endpunkte geht einher mit höherer Fallzahl oder Beobachtungsdauer und steigenden Rekrutierungs- und

Verfolgungskosten. Diese erhöhten Hürden rechtfertigen jedoch nicht die Lockerung von Maßstäben zur Selektion valider und relevanter Endpunkte [EMEA, 2006; Braitman, 2002].

Ausgeprägte Erwartungen von Patienten und Studienpersonal können die Beurteilung der Endpunktmessung bei unverblindeten patientenberichteten Endpunkten beeinflussen [Boutron, 2006].

Die Adjudikation von Endpunkten durch lokale oder zentrale, verblindete Reviewergruppen wurde vermehrt in Multi-Zentren-RCTs [Kirwan, 2007; Mahaffey, 2002; Eisenbud, 2001; Mahaffey, 2001] verwendet. Die Adjudikationsmethoden der „Women's Health Initiative" können in dieser Hinsicht als Vorbild gesehen werden [Curb, 2003].

Bei Endpunkten mit mehr als einer Komponente ist entscheidend, dass die Komponenten vergleichbare Relevanz, vergleichbare Ereignisraten, z.B. kardiovaskuläre Mortalität und Myokardinfarkt, und dieselbe Effektrichtung, z.B. Risikoreduktion, aufweisen [Ferreira-González, 2007; Freemantle, 2007; Freemantle, 2003]
.
Zusammenfassend lässt sich konstatieren: die Bewertung eines Studienendpunktes soll bei effektiv und dauerhaft verblindeten Patienten erfolgen. Die Bewertung soll mit einem validierten, reliablen und hoch standardisierten Endpunktleitfaden durchgeführt werden. Die für die Endpunkterhebung zuständigen Prüfärzte sollen effektiv und dauerhaft verblindet und nicht an der Interventionszuteilung oder –administration beteiligt sein. Weiterhin sollen sie über ähnliche Qualifikation und Berufserfahrung verfügen [Brehaut, 2007; Poses, 1997]. Konsolidierend soll der Endpunkt anhand weiterer zentral gesammelter objektiver Endpunkte, z.B. Radiologie, Laborparameter, und von Endpunktgutachtern unabhängiger zentraler Endpunktbeschluss-Komitees (Blinded Adjudication Outcomes Committee) validiert werden [Petersen, 2006; Curb, 2003; Walter, 1997].

1.3.2. Quantität des Interventionseffektes

1.3.2.1. Effektgröße

Die Bestimmung der minimalen, klinisch relevanten Differenz (MKRD) des Interventionseffekts wird in den Fachkreisen kontrovers diskutiert. Offen bleiben oft Fragen wie: Aus wessen Perspektive darf und soll die MKRD definiert werden (Arzt/ Patient, Individuum/ Population)? Wie ist die Variabilität bezüglich der Bestimmung von MKRD zu

berücksichtigen? [Naylor, 1994; van Walraven, 1999; Guyatt, 2002; Barrett, 2005; Brozek, 2006]. Kleine MKRDs, z.B. Absolutrisikoreduktion von 2 pro 100, können für schwerwiegende und relevante Endpunkte, wie Mortalität und hoch beeinträchtigende Morbiditäten mehr Akzeptanz bei Patienten und Ärzten finden als große MKRDs. Große MKRDs, z.B. Absolut-risikoreduktion von 20 pro 100, sind für weniger schwerwiegende Endpunkte, wie z.B. Husten, erforderlich. Das Konstrukt MKRD kann sich bei binären Endpunkten wie Mortalität oder Morbidität durch das Einbeziehen von Schmerzhaftigkeit oder Schweregrad als recht komplex erweisen.

Umstrittenere MKRDs beziehen sich oft auf durch stetige oder ordinale Maße erfasste und von Patienten berichtete Endpunkte (Patient Reported Outcomes= PROs) [Wyrwich, 2007; Pusic, 2007; Tubach, 2005]. Die US „Food and Drug Administration“ (FDA) stellte in einem Leitlinienentwurf vier Verfahren für die Derivation von MKRDs für PROs dar: Abgleich mit MKRDs für nicht-PROs, Kontrastieren der MKRDs unter verschiedenen PROs, Verwendung empirischer Befunde oder Verteilungsapproximation [FDA, 2006]. Weiterhin sollen stetige MKRDs sorgfältig kategorisiert werden. Die Anteile der Patienten pro vordefinierter Kategorie (klein, moderat, groß) einer MKRD sollen berichtet werden [Guyatt, 2004; Man-Son-Hing, 2002; Chan, 2001].

Von Bedeutung ist MKRD nicht nur für die Bewertung und Interpretation der Effektgröße, sondern auch für die prospektive Fallzahlkalkulation und die a priori-Festlegung der klinischen Relevanz [Schünemann, 2005; Goodman, 1994; Naylor, 1994]. Die Bestimmung des relevanten Unterschieds zwischen den Interventionsgruppen (das Delta) ist ein zentraler Bestandteil der Fallzahlkalkulation. Bei klinischen Studien wurden bis dato die MKRDs oft post hoc oder ad libitum und das Delta nach selektiertem Forschungsstand oder nach Machbarkeit definiert; öfter noch wurde über die beiden Konstrukte nichts berichtet [Man-Son-Hing, 2002; Chan, 2001]. Zurzeit gelten die MKRD zwischen Clopidogrel und Aspirin für ischämischen Schlaganfall [Gorelick, 2006; Albers, 2004], zwischen Coxiben und nichtsteroidalen Antirheumatika plus Protonenpumpen-Hemmern für gastrointestinale Komplikationen [Antman, 2007; Moore; 2006] sowie zwischen Warfarin und Aspirin für Schlaganfall bei Vorhofflimmern in der medizinischen Literatur als hoch umstritten [Hart, 2007; Gage, 2004].

Obwohl für die Bestimmung von MKRDs kein Ansatz ohne Einschränkungen zu erwarten ist, soll in die Entwicklung von konsensfähigen und patientenorientierten MKRDs investiert werden. MKRDs können durch Konsensverfahren in jedem medizinischen Fachgebiet und unter Betetiligung der betroffenen Patientenorganisationen entwickelt werden. Die Abwägung

von statistischen und klinischen Anforderungen an MKRD soll im Rahmen eines solchen Verfahrens sorgfältig erfolgen und transparent berichtet werden.

1.3.2.2. Effektunsicherheiten

Die Schätzung der Effektgröße einer Intervention ist, wie jede Schätzung, mit statistischer Unsicherheit verbunden. Für jede Effektgröße soll ein Konfidenz-Intervall und der exakte p-Wert dargestellt werden. Die statistische Unsicherheit kann nur in begründeten Fällen von den bestehenden Konventionen (zweiseitige Konfidenz-Grenzen, 95%- Konfidenz-Intervall) abweichen. Eine einseitige Konfidenz-Grenze ist lediglich für Einbahnhypothesen angemessen, z.B. Fertilitätserhöhung. Allerdings sind Einbahnhypothesen in der Medizin selten und fußen oft auf nicht stichhaltigen Erwartungen [Bland, 1994]. Einseitiges Testen ignorierte das Risikopotenzial einer Intervention. Das Cardiac Arrhythmia Suppression Trial (CAST) illustrierte die Illusion über einseitiges Nutzenpotenzial, da sich eine erhöhte Mortalität herausstellte [CAST Investigators, 1989]. Obwohl einseitiges Testen zur Reduzierung der notwendigen Fallzahl führt, wird eine Zahlhalbierung nicht erwartet [Moye, 2002].

Bei suboptimaler Datenlage, z.B. kleine Stichprobe, hohe Fehlangaben- oder hoher Ausreißeranteil, und bei unsicheren Annahmen zum statistischen Modell können „exakte“ [Agresti, 2003], „robuste“ oder nicht-parametrische [Altman, 2000] Methoden angewendet werden. Zur Vermeidung des Fehlers erster Art soll nur ein Test, entweder für einen primären Endpunkt oder für einen kombinierten Endpunkt, durchgeführt werden. Bei der Suche nach konfirmatorischen Ergebnissen ist eine Adjustierung des Signifikanzniveaus für multiples Testen unabdingbar. Falls mehrere Tests nur zum Zweck der Exploration durchgeführt werden, ist eine Adjustierung nicht notwendig [EMEA, 2002].

1.3.3. Qualität des Designs, der Durchfühung und der Analyse

Es bestehen drei Ansätze zur Qualitätsbewertung von RCTs: der Qualitäts-Komponenten-Ansatz, der Qualitäts-Score-Ansatz und der Qualitäts-Checklisten-Ansatz. Randomisierungs-methode, Allocation Concealment, Verblindung sowie Studienaustritte zählen zu den oft zitierten Qualitäts-Komponenten, deren Angemessenheit die interne Validität von Studien beeinflussen kann. Zudem sind Allocation Concealment, Verblindung und Studienaustritte die am häufigsten bewerteten Komponenten in SRs [Moja, 2005]. Die vier oben genannten

Qualitäts-Komponenten spielen eine zentrale Rolle bei der Reduzierung verschiedener Biasarten in RCTs (s. Tab. 8) und sind häufig ein Bestandteil von Qualitäts-Scores und -Checklisten.

Tab. 8 Bias und Gegenmaßnahmen in RCTs [Eigene Darstellung]

Bias	Definition	Gegenmaßnahmen
Selection-Bias	Systematische Unterschiede bzgl. „prognostischer Faktoren“ zwischen den Interventionsgruppen	Adäquate Randomisierungsmethode, Allocation Concealment, adäquate Fallzahl
Performance-Bias	Systematische Unterschiede bzgl. nicht-experimenteller Interventionen zwischen den Interventionsgruppen	Verblindung des für die Patientenversorgung zuständigen Studienpersonals
Detection-Bias	Systematische Unterschiede bzgl. der Bewertung von Endpunkten zwischen den Interventionsgruppen	Verblindung des für die Bewertung der Endpunkte zuständigen Studienpersonals
Attrition-Bias	Systematische Unterschiede bzgl. Studienaustritten zwischen den Interventionsgruppen	Dokumentation von Zahl und Gründen für Studienaustritte und Minimierung von Barrieren

Der Anteil von RCTs mit angemessenen Qualitäts-Komponenten variiert stark zwischen den Medizingebieten (s. Tab. 9), tendiert allerdings zur Verbesserung nach der Veröffentlichung des CONSORT-Statement 1996 [Moher, 2001; Plint, 2006]. Drei wichtige Aspekte bei der Bewertung von Qualitäts-Komponenten sollen allerdings nicht mehr in der methodischen Literatur vernachlässigt werden. Erstens: der Einfluss der Verblindung auf die interne Validität der RCT kann sich je nach Fragestellung unterscheiden. Der Einfluss von Endpunktverblindung als Maßnahme gegen Detection-Bias entfällt bei der Anwendung der Gesamtmortalität als Endpunkt. Zweitens: die Interdependenz zwischen Qualitäts-Komponenten soll bei deren Bewertung miteinbezogen werden. Eine RCT mit einer Fallzahlkalkulation, die für erwartete Studienaustritte Rechnung getragen hat, scheint weniger für Attrition-Bias anfällig zu sein. Drittens: es sollen in bestimmten Fragestellungen auch andere als die bisher oft abgehandelten Qualitäts-Komponenten berücksichtigt werden, z.B. die Standardisierung nicht-pharmakologischer Interventionen, bei denen die Expertise des Studienpersonals, z.B. in der Chirurgie oder in der Psychiatrie, sich stark auf die Ergebnisse auswirkt.

Tab. 9 Prozentualer Anteil der RCTs mit angemessenen Qualitäts-Komponenten [Eigene Darstellung]

Autor, Jahr	Medizingebiet	RCT-Zahl	Randomisierungs-methode	Allocation Concealment	Verblindung	Studienaustritte
Schulz, 1994b	Gynäkologie & Geburtshilfe	206	32	23	-	-
Schulz, 1996	Gynäkologie & Geburtshilfe	110	-	-	28.2 (DoB)	34.5 (With)
Kjaergard, 1999	Hepatologie	235	51.5	34	34 (DoB)	70.2 (With)
Adetugbo, 2000	Dermatologie	68	1	7	47 (DoB)	6 (ITT)
Latronico, 2002	Intensivmedizin	173	26.6	7	10.4 (DoB)	49.1 (With)
Montenegro, 2002	Periodontologie	177	16.5	6.5	55 (OutB)	11 (ITT)
Harrison, 2003	Orthodontie	155	50.3	2.6	6.5 (DoB)	28.4 (With)
Mills, 2004	Klinische Pharmakologie	193	17	3	26 (NR)	79 (ITT)
Strippoli, 2004	Nephrologie	430	-	7.4	7.4 (OutB)	29.7 (ITT)
Shang, 2005	Homöopathie	110	25	45	92 (DoB)	30 (ITT)

DoB= Double-blinding; OutB= Outcome-blinding; With= Reporting on withdrawals/ drop-outs;
ITT= Intention-to-treat analysis; NR= Not reported

1.3.3.1. Randomisierungsmethode

Die Hauptstärke von RCTs ist die Randomisierung, da sie bei adäquater Fallzahl dem Selection-Bias vorbeugt und ihn bei kleinerer Fallzahl reduziert. Allerdings kann die Methode zur Generierung der Randomisierungsliste (Randomisierungsmethode) von unterschiedlicher Qualität sein. Eine angemessene Randomisierungsmethode garantiert bei angemessener Fallzahl, dass die Wahrscheinlichkeit, mit der jeder Patient zur einen oder anderen Interventionsgruppe zugeordnet wird, gleich ist. Als angemessene Randomisierungs-methoden, die die Trennung von Gruppenzuordnung und Patientenmerkmalen gewährleisten können, gelten der Computer-basierte Zufallsgenerator, die Tabelle von Zufallszahlen, das Münzwerfen und die Karteneinmischung. Computer-basierte Zufallszahlen sind robuster und einfacher zu implementieren und zu dokumentieren als andere adäquate Verfahren. Alternation (Abwechslung), Geburtsdatum, Krankenaktennummer und Sozialversicherungs-nummer hingegen sind unangemessene Randomisierungsmethoden, sogenannte Quasi-Randomisierung [Schulz, 2002a].

1.3.3.2. Allocation Concealment

Ungeachtet der Randomisierungsmethode kann jegliche Abweichung von der strikten Einhaltung der Zuordnungssequenz zu Selection-Bias führen. Wenn der Allokationsplan zugänglich ist, kann das für die Rekrutierung zuständige Studienpersonal Patienten mit besserer oder schlechterer Prognose nach ihren Präferenzen einem bestimmten Studienarm

zuordnen. Daher gilt es als entscheidend für die Vermeidung von Manipulation der Zufallsliste durch die Prüfärzte, die Unvorhersehbarkeit der Gruppenzuordnung der Patienten zu gewährleisten, bis sie tatsächlich randomisiert werden. Zufallsgenerierte Allokationstabellen, insbesondere ohne kleine Blockbildung, bleiben eher verborgen als quasirandomisierte Pläne. Das Allocation Concealment (Verblindung der Randomisierung) kann erfolgen durch kodierte und identische Behälter aus einer unabhängigen Arzneimittelvergabestelle, mit Seriennummern versehene Arzneimittelpackungen, opake und versiegelte Briefumschläge oder zentrale Randomisierung, bei der in einem vom Studienzentrum entfernt liegenden Randomisierungszentrum die Gruppenzuordnung durch Telefon, Fax o. ä. kommuniziert wird [Schulz, 2002b].

1.3.3.3. Verblindung

Während Allocation Concealment in jeder RCT durchführbar ist, erwies sich die Verblindung der Studienbeteiligten bezüglich der administrierten Intervention manchmal als schwierig, z.B. bei nicht-pharmakologischen Interventionen. Wenn Patienten und Prüfärzte keine Kenntnis über die zugewiesene Intervention erhalten, spricht man von Doppel-Verblindung. Die Doppel-Verblindung, insbesondere bei Placebo-kontrollierten Studien, verringert den Detection- und den Peformance-Bias. Endpunktverblindung findet statt, wenn die Untersuchung der Endpunkte von einer Person durchgeführt wird, die die Gruppenzugehörigkeit der Patienten nicht kennt (Vorbeugung von Detection-Bias).

1.3.3.4. Attrition

Als angemessene Berücksichtigung der Studienaustritte (Attrition) gilt die Berichterstattung über deren Zahl und Gründe und die Anwendung des „Intention-To-Treat-Prinzips“ bei der Auswertung. Die Berücksichtigung der Studienaustritte führt zur Verminderung des Attrition-Bias.

Ob das Studiendesign einer RCT bezüglich der Qualitäts-Komponenten angemessen ist, kann anhand der in Tabelle 10 zusammengestellten Formulierungen in ihrer Publikation erschlossen werden.

Tab. 10 Kriterien zur Beurteilung von Qualitäts-Komponenten in RCTs [modifiziert nach: Siersma, 2006]

Component	Adequate	Inadequate
Generation of the allocation sequence	Computer-generated, random number table, coin tossing, card shuffling or similar	Case record numbers, social insurance number, birth dates or not described
Concealment of the allocation sequence	Central randomization, sequentially numbered sealed opaque envelopes, serially administrated coded identical drug containers or similar	Based on an open allocation sequence, alternation, or not described
Double blinding Trial	described as double blind, or outcome assessor and patients described as blinded	Tablets versus injection or similar, or not described
Intention-to-treat analysis	All randomized participants were included in the analysis in the group to which they originally were assigned	Some participants were excluded from the analysis or not described

1.3.3.5. Qualitäts-Scores und -Checklisten

Ein Qualitäts-Score vergibt einen numerischen Wert zu jeder Qualitäts-Komponente und summiert die Werte über alle Komponenten hinaus. In der Regel weisen hohe Qualitäts-Scores auf hohe methodische Qualität hin. Qualitäts-Checklisten vergeben keine numerischen Werte zu Qualitäts-Komponenten und fassen die Werte nicht zusammen. Eine systematische Übersicht von Moher und Kollegen identifizierte 25 Scores und 9 Checklisten zur Bewertung der methodischen Qualität von RCTs, von denen der erste, der TC Chalmers-Score, 1981 entwickelt wurde [Moher, 1995]. Allerdings variiert die Zahl der Komponenten pro Score von 3 bis 34 und zentrale Komponenten bekommen unterschiedliches Gewicht in den jeweiligen Scores. Zudem beinhalten manche Scores Items, wie das „Informed Consent“ oder die Generalisierbarkeit der Ergebnisse, die kaum mit der methodischen Qualität zu assoziieren sind. Da die Anwendung unterschiedlicher Scores zu divergierenden Qualitätsurteilen derselben RCT führen kann [Jüni, 1999], ist Vorsicht bei deren Auswahl geboten. Die nach Cochrane, Jadad und Schulz genannten Bewertungsinstrumente gelten als die am meisten verwendeten Qualitäts-Scores/ -Checklisten und es wurde ihre Validität und Reliabilität, wenn auch unzureichend, untersucht [Moja, 2005; Clark, 1999]. Auf der anderen Seite wird berechtigterweise zunehmend der Verlust oder das Verbergen wichtiger Informationen zu einzelnen Qualitäts-Komponenten mit unterschiedlichen Gewichten durch die Aggregation in einem Score kritisiert [Greenland, 2001].

1.3.3.6. Performanz

Die kontrollierten Bedingungen, unter denen klinische Studien verlaufen, zielen darauf ab, die durch die Randomisierung erreichte Vergleichbarkeit der Interventionsgruppen zu erhalten, in Bezug auf Ko-Interventionen und Ko-Morbiditäten sowie auf Beobachtung, Bewertung, Erfassung und Entscheidung. Obwohl die von der „International Conference on Harmonisation“ 1995 vorgeschlagene „Good Clinical Practice: Consolidated Guideline (GCP; E6)“ eine vereinbarte Richtlinie zwischen den Pharmaherstellern und den Zulassungsbehörden in der Europäischen Union, USA und Japan darstellte, dehnte sie sich zunehmend von regulatorischen auf nicht regulatorische klinische Studien aus. Während GCP zwecks der Performanz-, Beobachtungs- und Datenqualität dem im Rahmen von Datenmanagement mit hohen Kosten betriebenen, klinischen Monitoring, Audits und Dokumentationen ein Hauptaugenmerk widmete, vernachlässigte sie weitere Qualitäts-Komponenten, wie die Sicherung des Allocation Concealment und die Minimierung von Studienaustritten [Grimes, 2005]. Adäquate und erhaltende Verblindung der Patienten und des für deren Versorgung während der klinischen Studie zuständigen Personals trägt zur Verminderung des sogenannten Performance-Bias bei [Higgins, 2006].

Die Interaktion zwischen Studienteilnehmern kann in vielfältigen Placebo- und Nocebo-Effekten resultieren [Olshansky, 2007]. Eine 25 RCTs einschließende SR fand heraus, dass emotionale und wahrnehmungsunterstützende Arzt-Patient-Interaktion zu positiveren Endpunkten führte [Di Blasi, 2001]. Dies kann auf den möglichen Einfluss der Inner- und Außer-Trial-Umstände auf den Interventionseffekt hinweisen. Auf der anderen Seite sind manche Performanzbedingungen einer klinischen Studie, wie intensives Follow-Up und Dokumentation, im klinischen Alltag nicht erfüllbar.

1.3.4. Nutzen-Risiko-Abwägung

Abwägungen zwischen Nutzen und Schaden einer Intervention im Rahmen der Plannung von RCTs stießen oft auf Unvorhersehbarkeit, Unter-Standardisierung, relative Zeitversetzung und niedrige Inzidenz der Risiken [Ioannidis, 2006; Cuervo, 2003]. Das in der Umwelt- und Gesundheitspolitik gängige Vorsorgeprinzip [European Environment Agency, 2001], das mnemonische BRAND-Schema (Benefits, Risks, Alternatives, Nothing, Decision) und verschiedene methodische und statistische Ansätze [Wittes, 2007; Shaffer, 2006; Hart, 2005; Philips, 2004; Straus, 2002; Lilford, 1996; Jennison, 1993] können zur Nutzen-Risiko-

Abwägung verwendet werden. Sorgfältige Nutzen-Risiko-Abwägungen wurden durchgeführt für Hormonersatztherapie [Griffiths, 2005; Minelli, 2004], Mammographie [Fenton, 2004], Antidepressiva [Gunnell, 2004] und Antikoagulanzien [Holbrook, 2007].

1.3.5. Vorzeitiger Abbruch

Das „Data Safety and Monitoring Board“ einer klinischen Studie legt vor Studienbeginn Kriterien und Prozeduren fest, die für eine Früheinstellung des Experiments notwendig sind (Stopping Rules). Die „Stopping Rules“ schließen die „Monitoring Methods“, „Stopping Boundaries“, „Adjusted Analysis for Interim Monitoring and Truncation“ ein [Pocock, 2005; Grant, 2005; Walker, 2004]. Eine klinische Studie sollte erst vorzeitig beendet werden, wenn parallele Zwischenauswertungen des Nutzens und des Risikos in Durchkreuzung vorher festgelegter Beendungsgrenzen resultieren. Dabei sollen „hinreichende“ Ereignisse in Bezug auf den Nutzen-Effekt und den Risiko-Effekt eingetreten und für das multiple Testen adjustiert werden [Goodman, 2007; Mueller, 2007; Bassler, 2007]. Eine SR, die 143 vorzeitig aufgrund augenscheinlichen Nutzens, beendete RCTs einschloss, zeigte eine steigende Inzidenz solcher trunkierter RCTs (tRCT), eine Unterberichterstattung der Frühbeendungs-Maßstäbe, eine unplausible Größe des Nutzen-Effekts, insbesondere bei niedriger Zahl der Nutzenereignisse, was auf eine Überschätzung des Nutzens hindeutete, eine Verwendung umstrittener Komposit-Endpunkte für das Monitoring, die Trunkation und einen Mangel an adäquaten Risikodaten [Montori, 2005b]. Eine SR identifizierte 96 SRs, die mindestens eine aufgrund des Nutzens trunkierte RCT einschlossen, und fand heraus, dass 46% der SRs mehr als 1 tRCT einschlossen, 71% die vorzeitige Beendung nicht erwähnten, bei 34% der berechneten MAs (n= 47) tRCTs mehr als 40% des MA-Gewichts ausmachten und dass insgesamt eine Unterschätzung seitens der SRs für die erwartete Überschätzung durch tRCTs bestand [Bassler, 2007]. Da die meisten tRCTs in Fachzeitschriften hohen Impacts veröffentlicht werden und damit leicht für SRs lokalisierbar sind und da ihre Inzidenz sich erhöht, kann ihr Einschluss in MA zur Überschätzung des Nutzens und zur Unterschätzung des Risikos führen [Montori, 2005b; Bassler, 2007].

1.3.6. Präferenzen in klinischen Studien

Aus Informationen, Erwartungen und Wertungen von Patienten und Ärzten entstehende ausgeprägte Präferenzen für eine Intervention führen oft zur Nicht-Teilnahme an RCTs. Innovative Studiendesigns versuchen den Präferenzeneffekt durch (Teil-)Randomisierung zu

schätzen (Rücker-, Wennberg-, Zelen-Designs) [Rücker, 1989; Wennberg, 1993; Zelen, 1979] oder Teilnehmer mit Präferenzen konkurrent mit randomisierten Patienten zu beobachten und zu vergleichen (Comprehensive Cohort-Design) [Olschewski, 1985; Brewin, 1989]. Weiterhin strittig bleiben die Methoden zur Eruierung von Präferenzen, da letztere ein häufig komplexes, dynamisches und durch die Messung selbst beinflussbares Konstrukt sind. Eine SR, die 34 RCTs zum Einfluss von Präferenzen der Studienteilnehmer und des -personals einschloss, zeigte, dass Präferenz die Rekrutierung reduzierte. Weiterhin konnten keine Merkmalsunterschiede zwischen randomisierten und aufgrund starker Präferenzen nicht-randomisierten Patienten gefunden werden. Ein Einfluss der Präferenzen auf den Austritt aus der Studie konnte nicht festgestellt werden. Ein Einfluss der Präferenzen von Ärzten und Patienten auf die Endpunkte der Patienten konnte beobachtet werden. Allerdings wies dieser Einfluss unterschiedliche Richtungen auf [King, 2005].

2. Der Einfluss von Bias in randomisierten kontrollierten Studien auf die Ergebnisse von Meta-Analysen: Eine systematische Review der empirischen Studien

2.1. Hintergrund

Die methodische Qualität ist ein multidimensionales Konzept, das in dieser SR auf die interne Validität beschränkt wird. Die meisten Experten gehen davon aus, dass die methodische Qualität von RCTs im Wesentlichen mit der Angemessenheit der Randomisierungsmethode, des Allocation Concealments, der Verblindung sowie der Berücksichtigung von Studienaustritten zusammenhängt (s. Abschnitt 1.3.3). Ein Zusammenhang zwischen der methodischen Qualität und der geschätzten Effektgröße von RCTs wurde in zahlreichen meta-epidemiologischen Studien untersucht. Allerdings fanden die Studien diesbezüglich keine konsistenten Ergebnisse. Ein Einfluss von Bias wurde sogar auf die Ergebnisse von Tierexperimenten gefunden. Tierexperimente in der Notfallmedizin ohne Randomisierung und ohne Verblindung überschätzen die Effektgröße um das Fünffache [Bebarta, 2003]. Zudem deuten empirische Studien zum Vergleich der Ergebnisse von Tierexperimenten und humanen klinischen Studien an, dass die Variation ihrer Ergebnisse zu der selben Intervention mit unterschiedlicher Verteilung von Bias in beiden Studienarten zusammenhängen kann [Perel, 2007]. Die Angemessenheit aller Qualitäts-Komponenten soll im Kontext der Fragestellung untersucht werden.

Dieser Abschnitt stellt eine vom Verfasser durchgeführte SR der empirischen Vergleiche zwischen den Effektgrößen von RCTs mit hoher methodischer Qualität (RCTs-HQ) und den Effektgrößen von RCTs mit niedriger methodischer Qualität (RCTs-NQ) vor. Ein weiterer Schwerpunkt der SR liegt in der Berücksichtigung von zufallsbedingter und klinischer Heterogenität bei meta-epidemiologischen Studien.

2.2. Zielsetzungen

Der Zweck dieser SR ist, die Studien, die RCTs-HQ mit RCTs-NQ verglichen, systematisch zu suchen und kritisch zu bewerten. Folgende Fragestellungen werden im Rahmen dieser SR untersucht:

- Unterscheiden sich die zusammengefassten Effektgrößen von RCTs-HQ und RCTs-NQ?
- Wie wird die methodische Qualität von RCTs erhoben und ihre Variationen untersucht?
- Wird die zufallsbedingte Variation zwischen den RCTs bei der Synthese berücksichtigt?
- Werden mögliche klinische Ursachen der Heterogenität zwischen den RCTs untersucht?

2.3. Methodik

2.3.1. Suchstrategie

Im Cochrane Methodology Register (Ausgabe 2, 2006) wurde nach den Schlüsselwörtern "bias in trials", einschließlich der Teilmengen: "general", "relationship to trial quality", "random allocation", "blinding", "follow up", "intention to treat vs. on-treatment analysis", "small trial bias" gesucht. Die Suchfunktion „related articles" in Medline wurde für sechs Schlüsselstudien verwendet [Schulz, 1995; Moher, 1998; Jüni, 1999; Kjaergard, 2001; Balk, 2002; Egger, 2003] (Dezember 2006). Zudem wurde in den Literaturverzeichnissen der in der SR eingeschlossenen Studien nach relevanten Studien recherchiert. Es wurden bei der Suche keine Zeitfenster- oder Sprachenbeschränkungen angewendet.

2.3.2. Auswahl der Studien

Eingeschlossen wurden Studien, die:

(i) die Effektgrößen von RCTs-NQ vs. RCTs-HQ verglichen,

(ii) eine oder mehrere MAs einschlossen und

(iii) die Qualität von RCTs anhand von Scores oder einer der folgenden individuellen Komponenten bewerteten: Randomisierungsmethode, Allocation Concealment, jegliche Art der Verblindung und Berücksichtigung der Studienaustritte.

Ausgeschlossen wurden Studien, die das „vote counting“ verwendeten [Sutton, 1998]. Die Auswahl der Studien wurde nicht nach Art der Population, der Intervention oder der Endpunkte eingeschränkt.

Titel und Zusammenfassungen gefundener Studien wurden gesichtet. Es erfolgte eine stärker sensitive als spezifische Sichtung, d.h. bei Unklarheiten wurde der Volltext gesucht. Auch relevante Studien, für die keine Volltext-Publikation gefunden wurde, wurden in die Übersicht einbezogen. Alle Studien wurden von einer Person ausgewählt. Allerdings wurde die Sichtung von Titeln und Abstracts und die Selektion von Studien mit 3 Monaten Abstand vom Verfasser wiederholt. Zahl und Gründe für den Ausschluss von Studien wurden dokumentiert (s. Abbildung 1).

2.3.3. Extraktion der Daten

Daten wurden anhand einer a priori entwickelten Standardform vom Verfasser extrahiert. Für die Dateneingabe wurde Excel benutzt. Folgende Daten wurden aus den eingeschlossenen Studien extrahiert:

- Allgemeine Studienmerkmale: Erster Autor, Publikationsjahr, medizinisches Spezialgebiet
- Datengrundlage der empirischen Studien: Zahl der MAs, Zahl der RCTs
- Punktschätzer und Konfidenz-Intervalle der zusammengefassten Effektgrößen von RCTs-NQ und RCTs-HQ
- Methoden zur Bewertung methodischer Qualität von RCTs (Qualitäts-Scores, individuelle Qualitäts-Komponenten) und Zahl der Qualitäts-Begutachter (eine Person, mehr als eine Person)
- Definition der Kriterien zur Angemessenheit von individuellen Qualitäts-Komponenten und Festlegung des Grenzwerts bei Qualitäts-Scores, nach denen RCTs-NQ und RCTs-HQ getrennt wurden
- Methoden zur Berücksichtigung der methodischen Qualität in MAs (Subgruppen-Analyse, Meta-Regression, Schwellenwert-Analyse, Gewichtung, kumulative Meta-Analyse, graphische Darstellung, Bayesianische Hierarchische Modellierung)
- Modell der Synthese (FEM, REM) und Heterogenitäts-Test
- Confounding durch weitere Design-Merkmale, durch Erkrankung und durch Intervention

- Untersuchung der klinischen Heterogenität (interventionsbezogene Variationen, patientenbezogene Variationen)
- Berücksichtigung der Multiplizität

2.3.4. Statistische Auswertung

Um den Einfluss von Bias auf die Ergebnisse von RCTs zu schätzen, wurde der Quotient (das Ratio) von Odds-Ratios (ROR, d.h. die Effektgröße von RCTs-NQ zur Effektgröße von RCTs-HQ) kalkuliert oder aus Publikationen extrahiert. Der Logarithmus von ROR ist die Differenz der Logarithmen von Interventionseffekten in RCTs-NQ und RCTs-HQ. Die Varianz des Logarithmus von ROR ist die Summe der Varianzen der Logarithmen von Interventionseffekten in RCTs-NQ und RCTs-HQ. Bei negativen Ereignissen, z.B. Mortalität, deutet ein ROR kleiner eins auf eine Überschätzung des Interventionseffekts durch RCTs-NQ und ein ROR größer eins auf eine Unterschätzung des Effekts durch RCTs-NQ hin. Die RORs für Qualitäts-Scores und für jede individuelle Qualitäts-Komponente wurden anhand des REM von DerSimonian und Laird kombiniert [DerSimonian, 1986]. Vergleiche aus 9 Studien wurden von der MA der RORs ausgeschlossen, da 3 von ihnen erwünschte Ereignisse als Endpunkt benutzten [Khan, 1996a; Khan, 1996b; Als-Nielsen, 2003a], 3 stetige Zielvariablen verwendeten [McAlindon, 2000; Wang, 2004; Ni Mhurchu, 2005], in 2 Studien etwa die Hälfte der RCTs von der MA ausgeschlossen blieben [Verhagen, 2000; Verhagen, 2002] und in 1 Studie RCTs-NQ und RCTs-HQ auf derselben Datengrundlage basierten [Tierney, 2005]. Das ROR wurde in 8 Studien berichtet [Schulz, 1995; Moher, 1999b; Linde, 1999; Kjaergard, 2001; Balk, 2002; Egger, 2003; Shang, 2005; Siersma, 2006]. Allerdings wurde für 1 Studie das Inverse des RORs vom Verfasser errechnet, da die Publikation die Effektgröße von RCTs-HQ zur Effektgröße von RCTs-NQ berichtet [Balk, 2002]. In allen weiteren eingeschlossenen Studien wurde das ROR, wie oben beschrieben, im Rahmen dieser SR kalkuliert. Für 1 Studie wurde das zusammengefasste Relative Risiko (RR) aus der zusammengefassten Relative Risk Reduction errechnet [Nieuwenhoven, 2001]. 4 Studien teilten die RCTs in 3 Qualitätskategorien ein [Caubet, 1997: Potter, 1998; Gluud, 2001 (nur bei Doppel-Verblindung); Nowak, 2004], wobei der Verfasser die Effektgrößen von RCTs mittlerer und niedriger Qualität mittels des REM zusammenlegte und bei weiterer Auswertung als RCTs-NQ betrachtete.

Gewichtete Regression wurde benutzt, um die Unterschiede zwischen RORs in Bezug auf Medizingebiet und Intervention zu untersuchen. Dafür wurde ein Generalisiertes Lineares Modell (GLM) mit einer Log-Link-Funktion und im REM gewichteten kleinsten Quadraten

ausgeführt. Es ist anzunehmen, dass RORs für Erkrankungen innerhalb eines Medizingebiets und für Interventionen innerhalb einer MA homogener sind als RORs für Erkrankungen innerhalb mehrerer Medizingebiete und für Interventionen innerhalb mehrerer MAs. Anhand von 2 univariaten gewichteten Meta-Regression-Modellen wurden die Unterschiede untersucht zwischen RORs, die auf nur einer MA, bzw. nur einem Medizingebiet basierten, und RORs, die sich auf mehrere MAs und Medizingebiete bezogen. Die Gewichtung der kleinsten Quadrate in der Meta-Regression erfolgt nach dem REM.

Um Verzerrung durch Erkrankung zu vermeiden, wurden, wenn möglich, nach Medizingebiet stratifizierte empirische Vergleiche miteinbezogen. Allgemeine Merkmale, Datengrundlage und Methoden zur Bewertung und Berücksichtigung methodischer Qualität von RCTs wurden beschrieben und Modelle der Synthese, die Verwendung des Heterogenitäts-Tests und die Untersuchung klinischer Heterogenität wurden dargestellt.

Die Heterogenität zwischen den Studien wurde mit dem Cochran-Test untersucht und durch das Heterogenitäts-Maß I^2 quantifiziert [Higgins, 2002a]. Alle statistischen Auswertungen wurden mit der freien Software R (Version 2.2.0) durchgeführt.

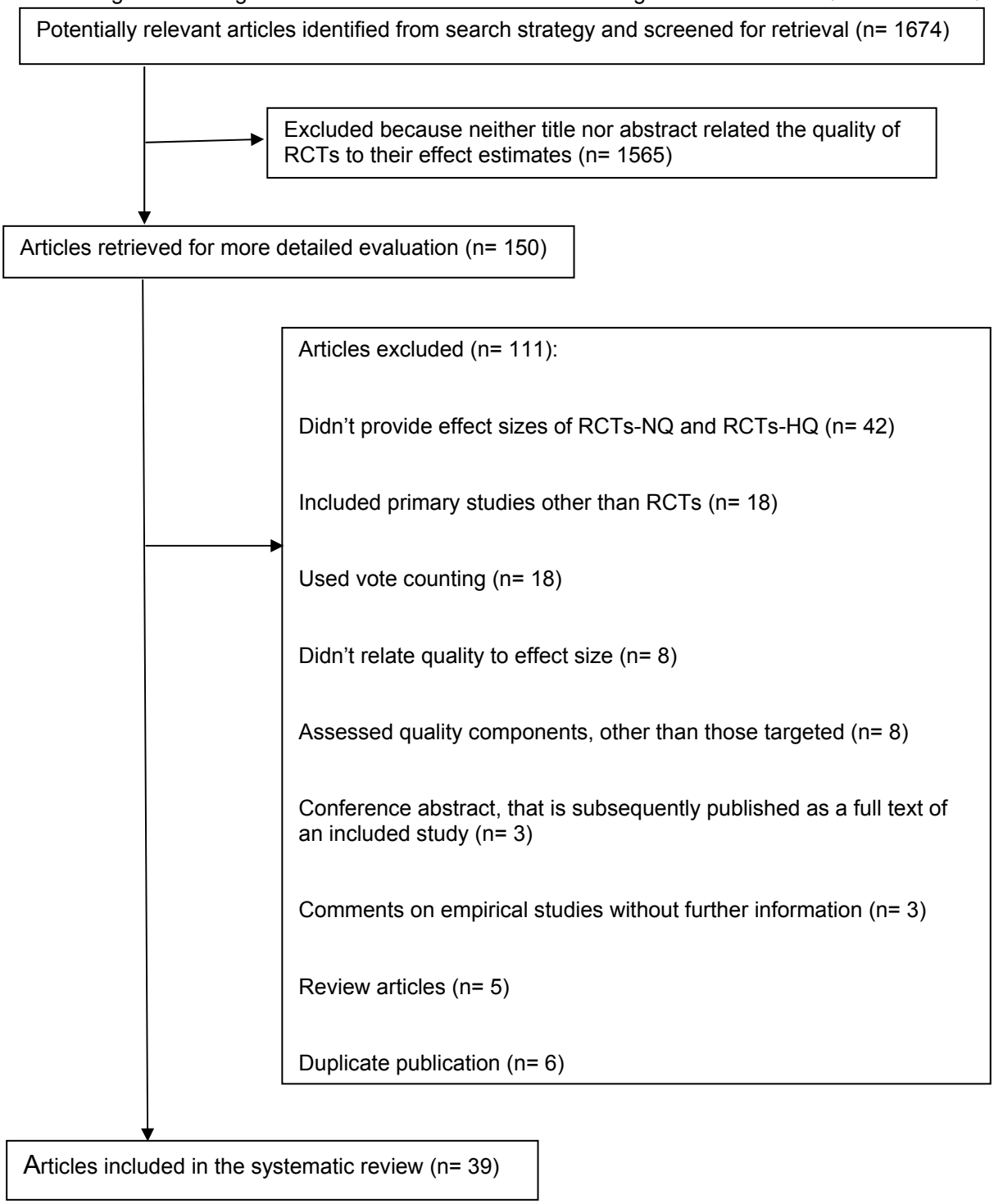
2.4. Ergebnisse

2.4.1. Ergebnisse der Suchen

Die Suchen ergaben 1674 Zitationen. Nach der Sichtung von Titeln und Zusammenfassungen wurden 150 als möglicherweise relevant eingestuft und für 138 von ihnen wurden Volltext-Publikationen beschafft. Für die verbleibenden 12 Zitationen waren nur Zusammenfassungen zugänglich.

Es wurden 39 Studien, einschließlich einer nicht publizierten empirischen Studie vom Verfasser, die auf 242 MAs und 2845 RCTs basieren, in die SR eingeschlossen. 111 Zitationen wurden ausgeschlossen, die Gründe für den Ausschluss sind in Abbildung 1 dargelegt. Ein Literaturverzeichnis für alle ausgeschlossenen Studien ist im Anhang 1 zu finden.

Abbildung 1: Flussdiagramm zur Auswahl der Studien zum Vergleich von RCTs-HQ und RCTs-NQ



2.4.2. Allgemeine Merkmale der eingeschlossenen Studien

39 empirische Studien, die auf 242 MAs und 2845 RCTs basierten, wurden in diese SR eingeschlossen. Die eingeschlossenen Studien basierten auf einer oder mehreren MAs und deckten ein oder mehrere Medizingebiete ab (s. Tab. 11). Etwa 80% der Studien (n= 32) berichteten die Effektgrößen von RCTs-NQ und RCTs-HQ für einzelne MAs und verringerten damit das Confounding durch Intervention. 7 Studien (18%) bezogen RCTs aus mehreren MAs zur Schätzung des Einflusses von methodischer Qualität [Schulz, 1995; Khan, 1996b; Moher, 1999b; Kjaergard, 2001; Balk, 2002; Egger, 2003; Siersma, 2006]. 7 Studien (18%) verwendeten MAs aus diversen Medizinfeldern und setzten sich damit einem Confounding durch Erkrankung aus [Moher, 1999b; Linde, 1999; Kjaergard, 2001; Balk, 2002; Egger, 2003; Shang, 2005; Siersma, 2006]. Allerdings wurden in 2 von ihnen auch nach medizinischem Spezialgebiet stratifizierte Effektgrößen von RCTs-NQ und RCTs-HQ berichtet [Balk, 2002 (4 Gebiete); Egger, 2003 (6 Gebiete)]. Bei weiteren Auswertungen wurden lediglich die nach Medizinfeldern stratifizierten Schätzungen verwendet. 10 Studien (26%) verglichen RCTs-NQ und RCTS-HQ für mehr als einen Endpunkt: [Siragusa, 1996 (2 Endpunkte); McAlister, 1998 (3 Endpunkte); D'Amico, 1998 (2 Endpunkte); Nieuwenhoven, 2001 (2 Endpunkte); Gluud, 2001(2 Endpunkte); Als-Nielsen, 2003a (2 Endpunkte); Wang, 2004 (3 Endpunkte); Villari, 2004 (2 Endpunkte); Abraham, 2004 (2 Endpunkte); Ni Mhurchu, 2005 (2 Endpunkte)]. Ioannidis zog Vergleiche für 2 Interventionen [Ioannidis, 1997], Fergusson, McAlindon und Roderick für jeweils 7, 2 und 6 Interventionen [Fergusson, 2000; McAlindon, 2000; Roderick, 2005]. Tierney präsentierte Vergleiche für 14 Erkrankungen [Tierney, 2005].

Tab. 11 Allgemeine Merkmale der eingeschlossenen Studien

Year	Study	Medical Field	MAs	RCTs
1995	Schulz	Reproductive Medicine	33	250
1996	Khan I	Reproductive Medicine	1	9
1996	Khan II	Reproductive Medicine	9	34
1996	Siragusa	Cardiovascular Disease	1	13
1997	Ioannidis	Infectious Disease	2	15
1997	Caubet	Oncology	1	13
1998	Ortiz	Rheumatology	1	8
1998	McAlister	Surgery	1	6
1998	D'Amico	Infectious Disease	2	32
1998	Potter	Nutrition	1	26
1999	Moher	Miscellaneous	11	127
1999	Jüni	Surgery	1	17
1999	Linde	Miscellaneous	1	89
2000	Verhagen I	Rheumatology	1	22
2000	Fergusson	Surgery	7	91
2000	McAlindon	Rheumatology	2	15

Tab. 11 (Forts.) Allgemeine Merkmale der eingeschlossenen Studien

Year	Study	Medical Field	MAs	RCTs
2001	Kjaergard	Miscellaneous	14	190
2001	Nieuwenhoven	Infectious Disease	1	32
2001	Wilkes	Emergency Medicine	1	42
2001	Gluud	Gastroenterology	1	16
2002	Verhagen II	Cardiovascular Disease	1	17
2002	Balk	Miscellaneous	26	276
2003	Egger	Miscellaneous	39	304
2003	Als-Nielsen	Gastroenterology	1	11
2004	Panpanich	Infectious Disease	1	14
2004	Nowak	Oncology	1	10
2004	Wang	Rheumatology	1	17
2004	Villari	Infectious Disease	1	53
2004	Abraham	Gastroenterology	3	39
2005	Silva Filho	Infectious Disease	1	38
2005	Shang	Miscellaneous	2	220
2005	Poeze	Emergency Medicine	1	30
2005	Roderick	Cardiovascular Disease	6	86
2005	Tierney	Oncology	14	92
2005	Ni Mhurchu	Nutrition	1	14
2005	Rambaldi	Gastroenterology	1	13
2006	Mukhtar	Cardiovascular Disease	1	13
2006	Siersma	Miscellaneous	48	523
2006	Aher	Paediatrics	1	28

MAs= Meta-Analyses; RCTs= Randomised Clinical Trials

2.4.3. Endpunkte und Datengrundlagen der Studien

Bei 92% der Studien (n= 36) fungierten unerwünschte Ereignisse, z.B. Mortalität, als Zielvariable. Lediglich die 2 Studien von Khan und 1 Studie von Als-Nielsen wiesen ein positives Ereignis als Endpunkt auf [Khan, 1996a; Khan, 1996b; Als-Nielsen, 2003a]. Mit Ausnahme von 4 Studien (10%) [Verhagen, 2000; McAlindon, 2000; Wang, 2004; Ni Mhurchu, 2005], die stetige Zielvariablen verwendeten, benutzten alle Studien binäre Endpunkte. Der Median der Anzahl der in den Studien eingeschlossenen RCTs betrug 28 (Interquartil-Bereich: 14 – 87.5).

2.4.4. Bewertung der methodischen Qualität von RCTs

In 41% der Studien (n= 16) wurden sowohl individuelle Komponenten als auch Scores zur Bewertung der Qualität von RCTs verwendet. 5 Studien (13%) verwendeten zusammengelegte Qualitäts-Komponenten [Potter, 1998; Nowak, 2004; Shang, 2005;

Rambaldi, 2005, Aher, 2006], die der Verfasser als Scores in dieser Arbeit behandelt. 13 Studien (33%) benutzten nur individuelle Qualitäts-Komponenten. In 10 Studien (26%) wurden lediglich Qualitäts-Scores eingesetzt (s. Tab. 12). 21 von 29 Studien (72%), die individuelle Komponenten verwendeten, führten die Kriterien zur Beurteilung ihrer Angemessenheit auf. 3 weitere Studien [Wilkes, 2001; Villari, 2004; Ni Mhurchu, 2005] verwiesen auf Standard-Definitionen (Schulz-Definition [Schulz, 1995], Cochrane-Definition [Higgins, 2006]). 6 Studien legten keine Definition der Qualitäts-Komponenten dar [Siragusa, 1996; Ioannidis, 1997; Jüni, 1999; Linde, 1999; Wang, 2004; Poeze, 2005]. In 20 von 26 Studien (81%), die Qualitäts-Scores benutzten, wurde der Grenzwert zur Trennung von RCTs-NQ und RCTs-HQ im Methodikteil definiert oder als zuvor festgelegt gekennzeichnet. 5 Studien definierten den Grenzwert erst im Ergebnisteil [Caubet, 1997; McAlister, 1998; Gluud, 2001; Villari, 2004; Aher, 2006]. Die Studie von Wang bestimmt keinen Grenzwert für den Qualitäts-Score, da er als stetige Variable in das Regressions-Modell einbezogen wurde [Wang, 2004]. 34 Studien (87%) berichteten, dass die Qualität von RCTs oder die Datenextraktion im Allgemeinen von mehr als einer Person bewertet wurde. In 2 Studien wurde sie von einer Person beurteilt [Schulz, 1995; Mukhtar, 2006] und in weiteren 3 Studien war die Zahl der bewertenden Personen unklar [Ioannidis, 1997; D'Amico, 1998; Tierney, 2005].

Tab. 12 Bewertung der methodischen Qualität von RCTs

Study	Score	Components
Schulz	None	4 [Rand; AllCon; DoB; With]
Khan I	Jadad	4 [Rand; AllCon; DoB; With]
Khan II	Jadad	4 [Rand; AllCon; DoB; With]
Siragusa	None	1 [OutB]
Ioannidis	None	1 [DoB]
Caubet	Chalmers	None
Ortiz	Jadad	None
McAlister	Jadad	None
D'Amico	None	2 [AllCon; DoB]
Potter	Potter	None
Moher	Jadad	3 [Rand; AllCon; DoB]
Jüni	Jadad [3 points]	3 [AllCon; OutB; With]
Linde	Jadad; Linde	4 [Rand; AllCon; DoB; With]
Verhagen I	Delphi	3 [Rand + AllCon (combined); DoB + OutB (combined); ITT]
Fergusson	Jadad	None
McAlindon	Chalmers	None
Kjaergard	Jadad	4 [Rand; AllCon; DoB; With]
Nieuwenhoven	Cook	3 [Rand; AllCon; OutB]
Wilkes	None	2 [AllCon; OutB]
Gluud	Jadad	4 [Rand; AllCon; DoB; ITT]
Verhagen II	Delphi	4 [Rand + AllCon (combined); DoB + OutB (combined); With; ITT]

Tab. 12 (Forts.) Bewertung der methodischen Qualität von RCTs

Study	Score	Components
Balk	None	4 [Rand; AllCon; DoB; With]
Egger	None	2 [AllCon; DoB]
Als-Nielsen	None	3 [Rand; AllCon; DoB]
Panpanich	Nonene	1 [AllCon]
Nowak	Nowak	None
Wang	Downs	2 [DoB (vs. SnB); ITT]
Villari	Jadad; Chalmers	3 [Rand; AllCon; DoB]
Abraham	Jadad	None
Silva Filho	Jadad, Maastricht, Delphi, (Cochrane) (combined)	None
Shang	Shang	4 [Rand; AllCon; DoB; ITT]
Poeze	Chalmers	1 [OutB]
Roderick	None	2 [AllCon; OutB]
Tierney	None	1 [ITT]
Ni Mhurchu	None	1 [AllCon]
Rambaldi	Rambaldi	2 [AllCon; DoB]
Mukhtar	Jadad	1 [AllCon]
Siersma	None	4 [Rand; AllCon; DoB; ITT]
Aher	Aher	None

Rand= Generation of random sequence; AllCon= Allocation concealment; DoB= Double-blinding; OutB= Outcome-blinding; With= Reporting on withdrawals/ drop-outs; ITT= Intention-to-treat analysis

2.4.5. Methoden zur Berücksichtigung methodischer Qualität von RCTs

14 Studien (36%) wandten mehr als eine Methode zur Berücksichtigung methodischer Qualität von RCTs an. Die Subgruppen-Analyse war der am meisten verwendete Ansatz (n= 33). In 14 Studien (36%) wurde Meta-Regression benutzt. Multivariate Regressions-Modelle zur gleichzeitigen Kontrolle individueller Qualitäts-Komponenten wurden in 6 Studien [Schulz, 1995; Khan, 1996b; Moher, 1999b; Linde, 1999; Nieuwenhoven, 2001; Siersma, 2006] geführt und bei der MA von RORs einbezogen. 7 Studien führten univariate Regressionen für einzelne Qualitäts-Komponenten durch [Jüni, 1999; Kjaergard, 2001; Balk, 2002; Wang, 2004; Villari, 2004; Shang, 2005; Mukhtar, 2006] und in einer weiteren Studie blieb dies unklar [Ioannidis, 1997] (s. Tab. 13). Logistische Regression wurde in 4 Studien [Schulz, 1995; Khan, 1996b; Moher, 1999b; Kjaergard, 2001] angewendet. Multi-Ebenen-Regression (Bayesianische Hierarchische Modelle) wurde in 2 Studien verwendet [Balk, 2002; Siersma, 2006].

2.4.6. Modelle der Synthese und Heterogenitäts-Tests

Während 14 Studien (36%) die RCTs im FEM und 12 Studien (31%) im REM kombinierten, führten 13 Studien (33%) die Synthese in beiden Modellen durch (s. Tab. 13). 11 Studien (28%) berichteten nicht über die Anwendung eines Heterogenitäts-Tests.

Tab. 13 Methoden der Berücksichtigung, Modelle der Synthese und Heterogenitäts-Tests

Study	Methods for Incorporation	Model of Synthesis	Test of Heterogeneity
Schulz	MR	FEM	Reported
Khan I	SB	FEM	Reported
Khan II	MR	FEM	Not Reported
Siragusa	SB	FEM	Reported
Ioannidis	SB; MR	FEM; REM	Reported
Caubet	SB	REM	Not Reported
Ortiz	SB	FEM	Reported
McAlister	SB	FEM; REM	Reported
D'Amico	SB	FEM	Reported
Potter	SB	FEM; REM	Reported
Moher	SB; MR; TH; WT	FEM	Reported
Jüni	SB; MR	FEM; REM	Not Reported
Linde	MR; CM	REM	Reported
Verhagen I	SB; GR	REM	Not Reported
Fergusson	SB	REM	Not Reported
McAlindon	SB	REM	Reported
Kjaergard	MR	REM	Not Reported
Nieuwenhoven	SB; MR	FEM	Not Reported
Wilkes	SB	FEM; REM	Reported
Gluud	SB	FEM; REM	Reported
Verhagen II	TH; WT; CM; GR	FEM	Not Reported
Balk	SB; MR	REM	Reported
Egger	SB; TH	REM	Not Reported
Als-Nielsen	SB	FEM; REM	Reported
Panpanich	SB	REM	Reported
Nowak	SB	FEM	Reported
Wang	SB; MR	FEM; REM	Reported
Villari	SB; MR	FEM; REM	Reported
Abraham	SB	FEM	Not Reported
Silva Filho	SB	REM	Reported
Shang	SB; MR	REM	Reported
Poeze	SB	FEM; REM	Reported
Roderick	SB	FEM	Reported
Tierney	SB	FEM	Reported
Ni Mhurchu	SB	FEM; REM	Reported
Rambaldi	SB	FEM; REM	Reported
Mukhtar	SB; MR	FEM; REM	Reported
Siersma	ML; LG; MR	REM	Not Reported
Aher	SB	FEM	Reported

SB= Subgroup analysis; MR= Meta-regression; LG= Logistic regression; ML= Multi-level analysis;
TH= Threshold analysis; WT= Quality weight; CM= Cumulative Meta-analysis, GR= Graphical plotting
FEM= Fixed-Effects-Model; REM= Random-Effects-Model

2.4.7. Einfluss der methodischen Qualität von RCTs auf die Effektgröße

Aus 30 Studien konnten 134 empirische Vergleiche zwischen den Effektgrößen von RCTs-NQ und RCTs-HQ extrahiert und kombiniert werden. 36 Vergleiche verwendeten Scores zur Bewertung der Qualität von RCTs. Der Einfluss der Randomisierungsmethode wurde von 17 Vergleichen untersucht. Allocation Concealment und Verblindung wurden jeweils in 32 und 37 Vergleichen behandelt. In 12 Vergleichen stand die Berücksichtigung von Studienaustritten im Mittelpunkt.

2.4.7.1. Einfluss der Qualitäts-Scores

Der Jadad-Score war der am meisten verwendete Qualitäts-Score (n= 23; 64%). Bei der Studie von Potter waren RCTs-HQ diejenigen, die angemessenes Allocation Concealment und vollständiges Follow-Up aufwiesen [Potter, 1998]. Die Studie von Silva Filho benutzte 3 Scores (Jadad, Maastricht, Delphi) und eine Check-Liste (Cochrane) zur Bewertung der Qualität von RCTs [Silva Filho, 2005]. Sie definierte RCTs-HQ als Primärstudien, die 50% der Gesamtpunktzahl von mindestens 2 der verwendeten Scores erreichten. In den Studien von Nowak und Shang wurden RCTs-HQ so definiert, dass sie angemessene Randomisierung und angemessenes Allocation Concealment aufwiesen und zudem doppel-verblindet waren [Nowak, 2004; Shang, 2005]. Die Studie von Rambaldi bezeichnete RCTs als hochqualitativ, wenn sie angemessene Randomisierung, angemessenes Allocation Concealment und angemessene Berichterstattung über Studienaustritte zeigten und eine Doppel-Verblindung aufwiesen [Rambaldi, 2005]. In der Studie von Aher mussten RCTs-HQ angemessenes Allocation Concealment, Einfach-Verblindung und Endpunkt-Verblindung aufweisen [Aher, 2006].

36 Vergleiche stellten RCTs mit hohen und mit niedrigen Scores gegenüber. In 25 Vergleichen (69%) wurde kein Unterschied zwischen RCTs mit hohem oder niedrigem Qualitäts-Score gefunden. Während 10 Vergleiche eine Überschätzung des Effekts durch RCTs-NQ zeigten, fand 1 Vergleich eine Unterbewertung des Effekts durch RCTs-NQ.

Eine MA der Vergleiche über alle Scores hinweg ergibt eine Überschätzung der Behandlungswirksamkeit durch RCTs-NQ um 19% gegenüber RCTs-HQ (s. Tab. 14).
Eine MA der Vergleiche, die den Jadad-Score verwendeten, ergibt eine Überschätzung der Behandlungswirksamkeit durch RCTs-NQ gegenüber RCTs-HQ um 21% (s. Tab. 14).

Tab. 14 Einfluss der Unterschiede in Qualitäts-Scores

Study	Score	ROR	Impact
Caubet	Chalmers	1.23 (0.96 - 1.57)	No
Ortiz	Jadad	1.90 (0.61 - 5.95)	No
McAlister I	Jadad	1.19 (0.56 - 2.53)	No
McAlister II	Jadad	1.20 (0.71 - 2.05)	No
McAlister III	Jadad	0.57 (0.16 - 1.98)	No
Potter	Potter	0.85 (0.34 - 2.11)	No
Moher	Jadad	0.66 (0.52 - 0.83)	Overestimation
Jüni	Jadad	0.88 (0.60 - 1.28)	No
Linde	Jadad	0.56 (0.40 - 0.79)	Overestimation
Fergusson I	Jadad	0.81 (0.50 - 1.29)	No
Fergusson II	Jadad	5.06 (0.23 - 114.06)	No
Fergusson III	Jadad	1.71 (0.05 - 56.27)	No
Fergusson IV	Jadad	0.59 (0.26 - 1.33)	No
Fergusson V	Jadad	1.24 (0.16 - 9.91)	No
Fergusson VI	Jadad	1.33 (0.53 - 3.36)	No
Fergusson VII	Jadad	0.57 (0.25 - 1.29)	No
Kjaergard	Jadad	0.56 (0.33 - 0.98)	Overestimation
Nieuwenhoven I	Cook	1.66 (1.17 - 2.35)	Underestimation
Nieuwenhoven II	Cook	1.33 (0.70 - 2.52)	No
Gluud I	Jadad	1.25 (0.51 - 3.07)	No
Gluud II	Jadad	1.69 (0.87 - 3.31)	No
Nowak	Nowak	0.83 (0.66 - 1.03)	No
VillarI	Jadad	0.73 (0.57 - 0.92)	Overestimation
VillarI	Chalmers	0.72 (0.59 - 0.88)	Overestimation
VillarII	Jadad	0.61 (0.37 - 1.02)	No
VillarII	Chalmers	0.61 (0.37 - 1.02)	No
AbrahamI	Jadad	0.78 (0.67 - 0.91)	Overestimation
AbrahamII	Jadad	0.47 (0.37 - 0.60)	Overestimation
AbrahamIII	Jadad	0.94 (0.82 - 1.07)	No
Silva Filho	Silva Filho	1.09 (0.91 - 1.30)	No
Shang I	Shang	0.62 (0.43 - 0.90)	Overestimation
Shang II	Shang	0.61 (0.34 - 1.09)	No
Poeze	Chalmers	0.54 (0.35 - 0.82)	Overestimation
Rambaldi	Rambaldi	0.65 (0.30 - 1.43)	No
Mukhtar	Jadad	1.02 (0.89 - 1.18)	No
Aher	Aher	0.57 (0.45 - 0.73)	Overestimation
Summary (All scores, n= 36, $\chi 2 < 0.001$, $I^2 = 70\%$)		0.81 (0.74 - 0.89)	Overestimation
Summary (Jadad score, n= 23, $\chi 2 < 0.001$, $I^2 = 63\%$)		0.79 (0.71 - 0.88)	Overestimation

2.4.7.2. Einfluss der Randomisierungsmethode

17 Vergleiche kontrastierten RCTs, die eine angemessene Methode zur Generierung der Zufallsliste berichteten, mit RCTs, die unangemessene oder unklare Methoden darstellten. In 12 Vergleichen (71%) wurde kein Unterschied zwischen RCTs-NQ und RCTs-HQ

diesbezüglich gefunden. 5 Vergleiche zeigten eine Überschätzung des Effekts durch RCTs mit unangemessener Randomisierungsmethode. Eine MA der Vergleiche fand eine Überschätzung des Interventionseffektes durch RCTs ohne angemessene Randomisierungsmethode um 16%, im Vergleich mit RCTs mit angemessener Randomisierungsmethode (s. Tab. 15).

Tab. 15 Einfluss der Randomisierungsmethode

Study	ROR	Impact
Schulz	0.95 (0.81 - 1.12)	No
Moher	0.89 (0.67 - 1.20)	No
Linde	0.64 (0.43 - 0.94)	Overestimation
Kjaergard	0.49 (0.30 - 0.81)	Overestimation
Nieuwenhoven I	0.76 (0.54 - 0.98)	Overestimation
Nieuwenhoven II	0.85 (0.65 - 1.05)	No
Gluud I	1.02 (0.41 - 2.55)	No
Gluud II	1.24 (0.63 - 2.43)	No
Balk I	0.88 (0.71 - 1.05)	No
Balk II	0.88 (0.59 - 1.37)	No
Balk III	1.00 (0.68 - 1.59)	No
Balk IV	1.32 (0.93 - 1.89)	No
Villari I	0.79 (0.70 - 0.90)	Overestimation
Villari II	0.55 (0.37 - 0.82)	Overestimation
Shang I	0.67 (0.48 - 0.95)	Overestimation
Shang II	0.98 (0.65 - 1.46)	No
Siersma	0.87 (0.74 - 1.01)	No
Summary (n= 17, $\chi 2$< 0.133, I^2= 36%)	0.84 (0.78 – 0.91)	No

2.4.7.3. Einfluss des Allocation Concealment

32 Vergleiche beschäftigten sich mit dem Einfluss des Vorhandenseins oder Fehlens eines angemessenen Allocation Concealment. In 27 Vergleichen (84%) wurde diesbezüglich kein Unterschied zwischen RCTs-NQ und RCTs-HQ gefunden. 5 Vergleiche zeigten eine Überschätzung des Effekts durch RCTs ohne Verblindung der Randomisierung. Eine MA der Vergleiche fand eine Überbewertung der Interventionswirksamkeit durch RCTs ohne angemessenes Allocation Concealment um 11% im Vergleich mit RCTs mit angemessenem Allocation Concealment (s. Tab. 16).

Tab. 16 Einfluss des Allocation Concealment

Study	ROR	Impact
Schulz	0.70 (0.62 - 0.79)	Overestimation
D'Amico II	1.29 (0.86 - 1.95)	No
Moher	0.63 (0.45 - 0.88)	Overestimation
Jüni	1.12 (0.76 - 1.65)	No
Linde	0.84 (0.60 - 1.18)	No
Kjaergard	0.60 (0.31 - 1.15)	No
Nieuwenhoven I	0.73 (0.44 - 1.12)	No
Nieuwenhoven II	0.98 (0.73 - 1.23)	No
Wilkes	0.98 (0.73 - 1.32)	No
Gluud I	1.54 (0.60 - 4.00)	No
Gluud II	1.77 (0.88 - 3.58)	No
Balk I	0.88 (0.70 - 1.04)	No
Balk II	1.03 (0.70 - 1.47)	No
Balk III	1.11 (0.78 - 1.72)	No
Balk IV	1.37 (0.81 - 2.78)	No
Egger I	0.94 (0.76 - 1.16)	No
Egger II	0.44 (0.21 - 0.90)	Overestimation
Egger III	0.79 (0.67 - 0.94)	Overestimation
Egger IV	0.68 (0.52 - 0.89)	Overestimation
Panpanich	2.19 (0.83 - 5.81)	No
Villari I	0.84 (0.70 - 1.02)	No
Villari II	3.25 (0.93 - 11.34)	No
Shang I	0.78 (0.57 - 1.07)	No
Shang II	0.76 (0.48 - 1.16)	No
Roderick I	1.21 (0.74 - 1.97)	No
Roderick II	1.00 (0.66 - 1.52)	No
Roderick IV	0.69 (0.39 - 1.19)	No
Roderick V	0.81 (0.50 - 1.32)	No
Roderick VI	2.07 (0.93 - 4.61)	No
Rambaldi	0.65 (0.30 - 1.43)	No
Mukhtar	0.91 (0.81 - 1.02)	No
Siersma	1.04 (0.90 - 1.19)	No
Summary (n= 32, $\chi 2 < 0.001$, $I^2 = 53\%$)	0.89 (0.83 - 0.95)	Overestimation

2.4.7.4. Einfluss der Verblindung

37 Vergleiche untersuchten den Einfluss verschiedener Verblindungstypen. In 24 Vergleichen (65%) wurde diesbezüglich kein Unterschied zwischen RCTs-NQ und RCTs-HQ gefunden. Während 9 Vergleiche eine Überschätzung des Effekts durch RCTs mit Verblindung zeigten, fanden 4 Vergleiche eine Unterbewertung des Effekts durch RCTs ohne Verblindung. Eine MA der Vergleiche fand eine Überbewertung der Interventionswirksamkeit durch RCTs ohne Verblindung um 9% im Vergleich mit RCTs mit Verblindung [ROR= 0.91 (95%-KI: 0.84 - 0.99), n= 32, $\chi 2 < 0.001$, $I^2 = 91\%$)].

In der Studie von Poeze war der Verblindungstyp nicht klar [Poeze, 2005] und bei der Studie von Wilkes gab es verschiedene Verblindungstypen [Wilkes, 2001]. Beide Studien beziehen sich auf kritisch kranke Patienten und verwendeten objektive Endpunkte. Daher wurden diese Studien der Endpunkt-Verblindung zugeordnet.

In 43% der Vergleiche (n= 16) wurde mehr als ein Endpunkt verwendet und es wurden keine Angaben über ihre Objektivität gemacht. Daher ist es in den meisten Vergleichen nicht möglich, den Einfluss von Verblindung nach der Objektivität der Zielvariablen zu differenzieren. In 7 Vergleichen fungierte Gesamtmortalität als Endpunkt [Ioannidis I, 1997; D'Amico I, 1998 Nieuwenhoven II, 2001; Wilkes, 2001; Gluud I, 2001; Poeze, 2005; Rambaldi, 2005]. Eine MA dieser Vergleiche zeigte, dass die Verblindung keinen Einfluss auf den Endpunkt Gesamtmortalität aufweist [ROR= 1.03 (95%-KI: 0.80 – 1.33)].

2.4.7.4.1. Einfluss der Doppel-Verblindung

24 Vergleiche untersuchten die Unterschiede zwischen doppel-verblindeten und nicht-doppel-verblindeten RCTs in Bezug auf deren Effektgrößen. Das Fehlen der Doppel-Verblindung führte zur Überschätzung des Behandlungseffekts um 12% im Durchschnitt (s. Tab. 17).

Tab. 17 Einfluss der Doppel-Verblindung

Study	ROR	Impact
Schulz	0.83 (0.71 - 0.96)	Overestimation
Ioannidis I	0.59 (0.54 - 0.64)	Overestimation
Ioannidis II	1.19 (1.15 - 1.21)	Underestimation
DAmico I	1.43 (1.03 - 1.99)	Underestimation
Moher	1.11 (0.76 - 1.63)	No
Linde	0.26 (0.14 - 0.51)	Overestimation
Kjaergard	0.56 (0.33 - 0.98)	Overestimation
Balk I	0.91 (0.75 - 1.11)	No
Balk II	1.41 (0.89 - 2.13)	No
Balk III	0.95 (0.62 - 1.79)	No
Gluud I	1.25 (0.51 - 3.10)	No
Gluud II	1.58 (0.81 - 3.05)	No
Egger I	0.91 (0.39 - 2.17)	No
Egger II	0.88 (0.63 - 1.25)	No
Egger III	0.90 (0.61 - 1.33)	No
Egger IV	0.96 (0.66 - 1.39)	No

Tab. 17 (Forts.) Einfluss der Doppel-Verblindung

Egger V	0.47 (0.26 - 0.84)	Overestimation
Egger VI	0.97 (0.79 - 1.20)	No
Villari I	0.64 (0.46 - 0.90)	Overestimation
Villari II	0.55 (0.20 - 1.52)	No
Shang I	0.44 (0.22 - 0.87)	Overestimation
Shang II	0.63 (0.36 - 1.11)	No
Rambaldi	1.47 (0.44 - 4.98)	No
Siersma	1.09 (0.90 - 1.33)	No
Summary (n= 24, χ2< 0.001, I^2= 94%)	0.88 (0.80 – 0.97)	Overestimation

2.4.7.4.2. Einfluss der Endpunkt-Verblindung

13 Vergleiche kontrastierten RCTs mit Endpunkt-Verblindung mit RCTs ohne Endpunkt-Verblindung. Eine MA der Vergleiche fand keinen signifikanten Zusammenhang zwischen der unverblindeten Bewertung der Endpunkte und der Effektgröße von RCTs (s. Tab. 18).

Tab. 18 Einfluss der Endpunkt-Verblindung

Study	ROR	Impact
Siragusa I	2.63 (0.30 - 22.83)	No
Siragusa II	2.02 (0.67 - 6.13)	No
Jüni	0.65 (0.43 - 0.99)	Overestimation
Nieuwenhoven I	0.68 (0.45 - 0.91)	Overestimation
Nieuwenhoven II	1.03 (0.83 - 1.24)	No
Wilkes	1.60 (1.02 - 2.53)	Underestimation
Poeze	1.02 (0.55 - 1.88)	No
Roderick I	1.18 (0.73 - 1.90)	No
Roderick II	0.83 (0.51 - 1.35)	No
Roderick III	1.81 (0.47 - 6.99)	No
Roderick IV	1.96 (1.08 - 3.56)	Underestimation
Roderick V	1.24 (0.81 - 1.91)	No
Roderick VI	0.50 (0.22 - 1.16)	No
Summary (n= 13, χ2< 0.014, I^2= 56%)	1.02 (0.86 - 1.21)	No

2.4.7.5. Einfluss der Studienaustritte

12 Vergleiche widmeten sich der Berücksichtigung von Studienaustritten in RCTs. Allerdings wandten die Studien für die Angemessenheit dieser Qualitäts-Komponente unterschiedliche

Kriterien an (s. Tab. 19). Dennoch, stellte keiner der extrahierten Vergleiche einen Zusammenhang zwischen der Methode zur Berücksichtigung von Studienaustritten und den Effektgrößen der RCTs fest (s. Tab. 20).

Tab. 19 Definition der Qualitäts-Komponente: Berücksichtigung von Studienaustritten

Study	Definition of the domain attrition bias
Schulz	"…klinische Studien that reported, or gave the impression, that no exclusions had taken place…klinische Studien that reported having made exclusions."
Jüni	"…intention-to-treat analysis performed… intention-to-treat analysis not performed"
Linde	"…complete follow-up, or intention-to-treat analysis."
Kjaergard	"…(adequate [number and reasons for dropouts and withdrawals described] or inadequate [number and reasons for dropouts and withdrawals not described])."
Gluud	"The trial report stated use of intention to treat analysis…The trial report did not state or use intention to treat analysis"
Balk	"Reason for dropouts given."
Shang	"Analysis by intention to treat if the reported number of participants randomised and the number analysed were identical."
Siersma	"All randomized participants were included in the analysis in the group to which they originally were assigned ….Some participants were excluded from the analysis or not described."

Tab. 20 Einfluss angemessener Berücksichtigung von Studienaustritten

Study	ROR	Impact
Schulz	1.07 (0.94 - 1.21)	No
Jüni	1.37 (0.92 - 2.03)	No
Linde	1.23 (0.85 - 1.77)	No
Kjaergard	1.50 (0.80 - 2.78)	No
Gluud I	0.61 (0.18 - 2.06)	No
Gluud II	0.79 (0.35 - 1.81)	No
Balk I	1.06 (0.85 - 1.33)	No
Balk II	1.06 (0.66 - 1.82)	No
Balk IV	1.43 (0.86 - 2.33)	No
Shang I	1.25 (0.87 - 1.80)	No
Shang II	1.14 (0.78 - 1.66)	No
Siersma	0.92 (0.79 - 1.06)	No
Summary (n= 12, $\chi 2 < 0.886$, $I^2 = 0\%$)	1.06 (0.98 - 1.14)	No

2.4.7.6. Zusammenfassung des Einflusses der methodischen Qualität von RCTs auf die Effektgröße

Insgesamt wurden 134 Vergleiche zwischen den zusammengefassten Effektgrößen von RCTs-NQ und RCTs-HQ gefunden. Mit Ausnahme der Komponente Studienaustritte ergab sich kein konsistenter Zusammenhang zwischen der methodischen Qualität von RCTs und deren Effektgrößen (s. Tab. 21). In drei Vierteln aller Vergleiche wurde kein Einfluss der Studienqualität auf den Behandlungseffekt (n= 100) gefunden, über ein Fünftel der Vergleiche verzeichnete eine Überschätzung des Effekts durch Studien niedriger Qualität (n= 29) und lediglich in weniger als 3% der Vergleiche war eine Unterbewertung des Effekts durch RCTs niedriger Qualität zu sehen. Die MAs der Vergleiche ergeben präzise Schätzer für die Zusammenhänge zwischen der methodischen Qualität und den Effektgrößen von RCTs.

Allerdings gab es eine hohe Heterogenität zwischen den Vergleichen. Sie betrug über 50%, gemessen am Maß I^2, bei den MAs zum Einfluss von Scores, Allocation Concealment, Doppel-Verblindung und Endpunkt-Verblindung. Die Heterogenitäts-Maße waren bei den MAs zur Randomisierungsmethode und Berücksichtigung von Studienaustritten mittelmäßig bis niedrig. Niedrige Scores, inadäquate Randomisierungsmethode, inadäquates Allocation Concealment und fehlende Doppel-Verblindung führten im Durchschnitt zur Überschätzung des Interventionseffekts um 11% bis 19%. Es konnte kein Zusammenhang zwischen den Effektgrößen von RCTs und der Berücksichtigung von Studienaustritten oder Verblindung des Endpunkts gefunden werden.

Tab. 21 Einfluss verschiedener Biasarten

	Scores	Randomisierungs-methode	Allocation Concealment	Doppel-Verblindung	Endpunkt-Verblindung	Studien-austritte
Overestimation	10	5	5	7	2	0
Underestimation	1	0	0	2	2	0
No Difference	25	12	27	15	9	12
Summary	0.81 (0.74 - 0.89)	0.84 (0.78 - 0.91)	0.89 (0.83 - 0.95)	0.88 (0.80 - 0.97)	1.02 (0.86 - 1.21)	1.06 (0.98 - 1.14)

2.4.8. Confounding durch Erkrankung und Intervention

Die aus gemischten Medizingebieten geschätzten RORs zum Einfluss von Verblindung unterschieden sich nicht von RORs, die auf MAs aus demselben Erkrankungsfeld basierten [OR= 0.86 (95%-KI: 0.60 – 1.25)]. Diesbezüglich wurden ähnliche Ergebnisse bei den RORs

zum Einfluss von Allocation Concealment gefunden [OR= 0.92 (95%-KI: 0.71 – 1.20)]. RORs zum Einfluss von Verblindung, die sich auf mehrere MAs gründeten, variierten nicht von RORs, die aus einer MA berechnet wurden [OR= 0.96 (95%-KI: 0.76 – 1.22)]. Auch RORs zum Einfluss von Allocation Concealment, die sich auf mehrere MAs gründeten variierten nicht von RORs, die aus einer MA berechnet wurden [OR= 0.87 (95%-KI: 0.71 – 1.07)].

2.4.9. Confounding durch weitere Designmerkmale

21 Studien (54%) suchten nach einem Zusammenhang zwischen den Effektgrößen von RCTs und weiteren Qualitäts-Komponenten. Der Einfluss der Studiengröße und der Länge des Follow-Up wurde in jeweils 7 Studien überprüft (s. Tab. 22).

2.4.10. Untersuchung klinischer Heterogenität

Lediglich 14 Studien (36%) untersuchten interventionsbezogene Variationen, z.B. Typ, Dosis, Administrationsweg, Dauer, Ko-Intervention, zwischen den RCTs anhand von univariaten Verfahren. 5 weitere Studien diskutierten die Möglichkeit des Confounding des Zusammenhangs zwischen methodischer Qualität und Effektgröße durch interventions-bezogene Unterschiede. Mittels univariater Analyse überprüften 11 Studien (28%) Unterschiede bezüglich Patientenmerkmalen (z.B. Alter, Geschlecht, Schweregrad der Erkrankung, Dauer der Erkrankung). In 10 Studien wurde eine mögliche Verzerrung der Assoziation zwischen Qualität und Effektgröße durch Heterogenität der Patienten-charakteristika erörtert (s. Tab 22). Nur 3 Studien verwendeten multivariate Regression um gleichzeitig methodische und klinische Heterogenität zu kontrollieren [Villari, 2004; Shang, 2005; Mukhtar, 2006].

2.4.11. Berücksichtigung der Multiplizität

Nur 3 Studien [Gluud, 2001; Balk, 2002; Mukhtar, 2006] diskutierten Multiplizität als mögliche Quelle für falsch positive Zusammenhänge zwischen methodischer Qualität und Effektgröße von RCTs. Keine Studie adjustierte den Fehler erster Art, z.B. nach der Bonferroni-Methode, um der Multiplizität Rechnung zu tragen.

Tab. 22 Berücksichtigung weiterer Heterogenitätsquellen und der Multiplizität

Study	Confounding by Design Variables Number [Variables]	Confounding by Intervention Number [Variables]	Confounding by Patients' Variables Number [Variables]	Multiplicity
Schulz	None	Discussed	None	None
Khan I	None	None	None	None
Khan II	1 [Study Design (Parallel vs. Cross-over)]	None	None	None
Siragusa	None	None	None	None
Ioannidis	2 [Study Size; Duration of Follow-Up]	None	None	None
Caubet	1 [Controlling for Cross-overs]	None	Discussed	None
Ortiz	2 [Study Size; Duration of Follow-Up]	2 [Type; Dose]	None	None
McAlister	None	None	Discussed	None
D'Amico	None	1 [Route of Administration]	1 [Severity of Condition]	None
Potter	1 [Duration of Follow-Up]	4 [Type; Dose; Route of Administration; Duration]	5 [Age; Severity of Condition; Body Mass Index; Specialty Group (medical or surgical speciality); Underlying Disease (malignant or non-malignant)]	None
Moher	None	None	None	None
Jüni	None	None	Discussed	None
Linde	None	None	None	None
Verhagen I	None	Discussed	Discussed	None
Fergusson	1 [Type of Control (Placebo vs. Open-Label)]	None	None	None
McAlindon	2 [Study Size, Duration of Follow-Up]	Discussed	None	None
Kjaergard	1 [Study Size]	None	None	None
Nieuwenhoven	2 [Baseline Comparability; Patient Selection]	None	Discussed	None
Wilkes	2 [Study Design (Parallel vs. Cross-over), Type of End-Point (Primary or Not)]	Discussed	Discussed	None
Gluud	1 [Extended Follow-Up]	2 [Dose; Duration]	1 [Baseline Bilirubin]	Discussed
Verhagen II	None	None	None	None
Balk	24 [see original publication]	Discussed	Discussed	Discussed
Egger	None	None	Discussed	None
Als-Nielsen	None	4 [Type; Dose; Route of Administration; Duration]	2 [Duration of Condition; Severity of Condition]	None
Panpanich	None	None	2 [Age; Type of Infection]	None

Tab. 22 (Forts.) Berücksichtigung weiterer Heterogenitätsquellen und der Multiplizität

Study	Confounding by Design Variables Number [Variables]	Confounding by Intervention Number [Variables]	Confounding by Patients' Variables Number [Variables]	Multiplicity
Nowak	None	1 [Dose]	2 [Sex; Severity of Condition]	None
Wang	3 [Study Size; Trial Duration; Single-Center vs. Multicenter Trial]	2 [Type; Co-Interventions]	3 [Age; Severity of Condition; Effusion as inclusion or exclusion Criterion]	None
Villari	1 [Type of Control (Placebo vs. Active)]	3 [Type; Vaccine strains recommended; Matching between Vaccine and circulating strains]	1 [Age]	None
Abraham	3 [Rome-Score (3 points): Method of Outcome Assessment (Symptoms vs. Laboratory); Use of Validated Outcome Measurer; Amount of Improvement required to be considered as Treatment Responder]	None	None	None
Silva Filho	None	None	Discussed	None
Shang	2 [Study Size (log SE); Duration of Follow-Up]	1 [Type]	1 [Type of Indication]	None
Poeze	2 [Controlling for Cross-over; Mortality as End-Point (Primary or Not]	1 [Type]	Discussed	None
Roderick	1 [Venogram vs. Others for DVT Confirmation]	None	None	None
Tierney	None	None	None	None
Ni Mhurchu I	2 [Study Size; Duration of Follow-Up]	1 [Co-Interventions]	None	None
Rambaldi	None	1 [Type]	1 [Duration of Condition]	None
Mukhtar	None	None	2 [Sex; Baseline LDL]	Discussed
Siersma	2 [Power Calculation; Type of Control (None treatment vs. Placebo vs. Active)]	1 [Type]	None	None
Aher	1 [Use of Strict Guidelines for red blood cell transfusion]	2 [Dose; Co-Intervention]	None	None

2.5. Zusammenfassung und Diskussion

Die vorliegende systematische Übersicht identifiziert und bewertet 39 methodische Studien zum Vergleich von RCTs-HQ mit RCTs-NQ. Ein Fünftel der Studien basiert auf mehr als einer MA (Confounding durch Intervention) und lediglich in 5 Studien wurde der Einfluss von Bias nur aus gemischten Medizinfeldern (Confounding durch Erkrankung) berichtet. 40% der Studien verwendeten sowohl individuelle Komponenten als auch Scores zur Bewertung methodischer Qualität. Die meisten Studien definierten Kriterien zur Bewertung der Qualitäts-Komponenten und setzten a priori den Grenzwert zur Trennung von RCTs-HQ und RCTs-NQ bei der Verwendung von Qualitäts-Scores an. Etwa ein Drittel der Studien benutzten mehr als eine Methode zur Berücksichtigung methodischer Qualität von RCTs. In einem Drittel der Studien erfolgte die Synthese im FEM und im REM und in etwa einem Viertel der Studien wurde nicht über einen Heterogenitäts-Test berichtet.

Zum Zweck der MA konnten 134 empirische Vergleiche aus 30 Studien extrahiert werden. Betrachtet man die Vergleiche einzeln, findet man kein konsistentes Ergebnis bezüglich des Zusammenhangs zwischen der methodischen Qualität von RCTs und deren Effektgröße außer bei der Qualitäts-Komponente: die Berücksichtigung von Studienaustritten, bei der durchgehend kein Zusammenhang gefunden wurde. Im Durchschnitt überbewerteten RCTs mit niedrigen Qualitäts-Scores, mit unangemessener Randomisierungsmethode, mit unangemessenem Allocation Concealment und ohne jegliche Art der Verblindung die Behandlungswirksamkeit. Da die meisten Vergleiche nicht über die aus den RCTs extrahierten Endpunkte berichteten, konnte keine Differenzierung des Zusammenhangs zwischen Verblindung und Effektgröße nach der Objektivität des Endpunktes durchgeführt werden. Während das Fehlen der Doppel-Verblindung zur Überschätzung der Interventionswirksamkeit führte, zeigten die Effektgrößen von RCTs mit und ohne Verblindung der Endpunktmessung überraschenderweise keine Unterschiede.

Die meisten empirischen Studien zielten nicht darauf ab, die klinische Heterogenität zwischen den RCTs zu untersuchen. Lediglich ein Drittel der Studien untersuchte interventionsbezogene Variationen zwischen den RCTs und nur ein Viertel der Studien suchte nach patientenbezogenen Unterschieden. Simultane Kontrolle von methodischen und klinischen Heterogenitätsquellen durch multivariate Regression wurde nur in drei Studien verfolgt. Die Vernachlässigung klinischer Variationen zwischen den RCTs kann zu Confounding gefundener Zusammenhänge zwischen methodischer Qualität und Effektgröße führen.

Die vorliegende Arbeit weist mehrere Einschränkungen auf. Die Identifikation von meta-epidemiologischen Studien gilt als schwierig [Kunz, 1998, Kunz, 2002] und somit können relevante Studien durch die verwendete Suchstrategie unidentifiziert geblieben sein. Durch Kontakt mit Experten und Handrecherchen können möglicherweise weitere Studien gefunden werden. Zudem erfolgte die Auswahl der Studien durch eine Person, was zum Verfehlen relevanter Studien geführt haben könnte. Da die in den einzelnen empirischen Studien eingeschlossenen RCTs nicht identifizierbar waren, kann nicht ausgeschlossen werden, dass ein und dasselbe RCT mehrfach in den empirischen Studien eingeschlossen wurde. Die empirischen Studien bewerteten die berichtete Qualität von RCTs, d.h. bei berichteten unangemessenen Qualitäts-Komponenten oder fehlenden Angaben wurde von niedriger Qualität ausgegangen. Mehrere empirische Untersuchungen zeigten, dass die berichtete methodische Qualität von RCTs nicht mit der tatsächlichen Qualität übereinstimmt und dass Autoren von RCTs oft die Publikation angemessener Qualitäts-Komponenten versäumten [Hill, 2002; Devereaux, 2004; Soares, 2004]. Im Gegensatz dazu fanden andere Untersuchungen heraus, dass hinter schlechter Berichterstattung eher eine niedrige Qualität steht [Hadhazy, 1999; Pildal, 2004b; Manheimer, 2006].

Die vorliegende Arbeit ist die erste SR der empirischen Studien zum Einfluss von Bias in RCTs auf die Ergebnisse von MAs. Diese SR stellte im Durchschnitt eine mäßige Überschätzung des Behandlungseffekts durch RCTs mit niedriger methodischer Qualität fest. Allerdings berücksichtigt die große Mehrheit der eingeschlossenen Studien nicht die klinische Heterogenität zwischen den RCTs. Diese Tatsache kann zur Verzerrung des gefundenen Zusammenhangs zwischen der Studienqualität und der Effektgröße führen.

3. Meta-Analysen mit individuellen versus mit aggregierten Patientendaten: Eine systematische Review der empirischen Studien

3.1. Hintergrund

Meta-Analysen mit individuellen Patientendaten (MA-IPDs) bieten im Vergleich zu Meta-Analysen mit aggregierten Patientendaten (MA-APDs) mehr Möglichkeiten bei der Daten-Aufbereitung und bei der Untersuchung von klinischer Heterogenität. Dennoch setzen MA-IPDs die Bereitschaft für eine Datenbeteiligung und für eine Kollaboration voraus, beide sind mit zusätzlichem Organisations- und Personalaufwand verbunden sind (s. Abschnitt 1.2.3). MA-IPDs werden als „Goldstandard“ für die Durchführung von SR der RCTs bezeichnet [Stewart, 2002], obwohl diesbezüglich umfassende empirische Evidenz bis dato fehlt. Es gibt bislang keine systematische Review (SR) der Vergleiche zwischen MA-IPDs und MA-APDs.

Der vorliegende Abschnitt stellt eine SR der Literatur zum Vergleich von MA-IPDs und MA-APDs dar. Der Schwerpunkt der SR liegt in der Berücksichtigung von zufallsbedingter, methodischer und klinischer Heterogenität bei MA-IPDs und MA-APDs.

3.2. Zielsetzungen

Das Ziel dieser SR ist, die empirischen Studien, die MA-IPDs und MA-APDs vergleichen, systematisch zu suchen, zu bewerten und auszuwerten. Folgende Fragestellungen werden im Rahmen dieser SR untersucht:

- Unterscheiden sich die Punktschätzer der zusammengefassten Effektgrößen von MA-IPDs und MA-APDs?
- Unterscheiden sich die Konfidenz-Intervalle der zusammengefassten Effektgrößen von MA-IPDs und MA-APDs?
- Unterscheiden sich MA-IPDs, die Daten aus publizierten und nicht publizierten RCTs einschlossen, von den MA-APDs, die lediglich auf Daten aus publizierten RCTs zurückgriffen?
- Unterscheiden sich MA-IPDs, die alle randomisierten Patienten einschlossen und somit eine Intention-To-Treat-Analyse durchführten, von den MA-APDs, die einen Teil der randomisierten Patienten von der Auswertung ausschlossen?
- Unterscheiden sich MA-IPDs, die das Hazard Ratio als Effektmaße verwendeten, von den MA-APDs, die das OR als Effektmaße benutzten?

- Unterscheiden sich MA-IPDs mit längerem Follow-Up und MA-APDs ohne dieses?
- Wird die zufallsbedingte Variation bei der Synthese berücksichtigt?
- Wird die methodische Heterogenität zwischen den RCTs untersucht?
- Werden mögliche klinische Ursachen der Heterogenität zwischen den RCTs untersucht?

3.3. Methodik

3.3.1. Suchstrategie

Im „Cochrane Methodology Register“ (Ausgabe 2, 2006) wurde nach den Schlüsselwörtern „individual patient data”, einschließlich aller Teilmengen: “general methods”, “IPD vs other types of meta- analysis”, “IPD & non IPD in a meta- analysis“ gesucht. Medline wurde nach den Schlüsselwörtern “individual patient data” durchsucht und die Suche wurde auf den Publikationstyp „Meta-analysis“ begrenzt (März 2006). Zudem wurden 52 Kollaborative Gruppen für MA-IPDs (s. Tab. 23), die „Cochrane Database of IPD Reviews“ (http://www.ctu.mrc.ac.uk/cochrane/ipdmg/DBIPD.asp) und die Literaturverzeichnisse der in der SR eingeschlossenen Studien auf relevante Studien hin durchsucht. Zeitfenster- oder Sprachenbeschränkungen wurden bei den Suchen nicht angewendet.

MA-APDs bezieht sich auf MAs, die aggregierte Patientendaten verwenden, z.B. Effektgröße, durchschnittliches Alter, Frauenanteil. Die aggregierten Daten können aus den Publikationen extrahiert oder von den Forschern abgefragt werden. Da es nur eine begrenzte Anzahl von MA-IPDs gibt, zielte die Suchstrategie darauf ab, Publikationen zu MA-IPDs zu identifizieren und die Studien zum Vergleich von MA-IPDs und MA-APDs unter ihnen zu selektieren.

Tab.23 Kollaborationen für MA-IPDs (Eigene Recherche)

Cardio- and Cerebrovascular Diseases
1. Antithrombotic Trialists' Collaboration (previously called: Antiplatelet Trialists' Collaboration) 2. Dipyridamole in Stroke Collaboration (DISC) 3. Direct Thrombin Inhibitor Trialists' Collaborative Group 4. Fibrinolytic Therapy Trialists' (FTT) Collaborative Group 5. Blood Pressure Lowering Treatment Trialists' Collaboration 6. Blood pressure in Acute Stroke Collaboration (BASC) 7. ACE Inhibitors in Diabetic Nephropathy Trialist Group 8. Cholesterol Treatment Trialists' (CTT) Collaboration

9. Homocysteine Lowering Trialists' Collaboration
10. B-Vitamin Treatment Trialists' Collaboration
11. Stroke Unit Trialists' Collaboration
12. Early Supported Discharge Trialists (for stroke patients)
13. Outpatient Service Trialists (for stroke patients)
14. Coronary Artery Bypass Graft Surgery Trialists Collaboration
15. Carotid Endarterectomy Trialists' Collaboration
16. European Carotid Surgery Trialists' Collaboration

Cancer

1. Early Breast Cancer Trialists' Collaborative Group
2. Combined Hormone Agents Trialists' Group (for women with advanced breast cancer)
3. Neoadjuvant Chemotherapy for Cervical Cancer Meta-analysis Collaboration (NACCCMA)
4. The Ovarian Cancer Meta-Analysis Project
5. PORT Meta-analysis Trialists Group (Postoperative radiotherapy for non-small cell lung cancer)
6. Prophylactic Cranial Irradiation Overview Collaborative Group (for patients with small lung cancer)
7. Non-small Cell Lung Cancer Collaborative Group
8. Prostate Cancer Trialists' Collaborative Group
9. Advanced Bladder Cancer (ABC) Meta-analysis Collaboration
10. Colorectal Cancer Collaborative Group (also called: Colorectal Meta-analysis Collaboration)
11. Advanced Colorectal Cancer Meta-Analysis Project
12. Meta-Analysis Group in Cancer (for patients with colorectal cancer)
13. Meta-Analysis Group of the Japanese Society for Cancer of the Colon Rectum
14. Oesophageal Cancer Collaborative Group
15. Pancreatic Cancer Meta-analysis Group
16. International Hodgkin's Disease Collaborative Group
17. Sarcoma Meta-analysis Collaboration (SMAC)
18. Collaborative meta-analyses of Leukemia trials
19. Childhood Acute Lymphoblastic Leukaemia (ALL) Collaborative Group Chronic
20. Chronic Lymphocytic Leukemia (CLL) Trialists' Collaborative Group
21. Acute Myeloid Leukaemia (AML) Collaborative Group
22. Chronic Myeloid Leukemia Trialists' Collaborative Group
23. Myeloma Trialists' Collaborative Group
24. Glioma Meta-Analysis Trialists Group

Others

1. Artemether-Quinine Meta-analysis Study Group
2. HIV Trialists' Collaborative Group
3. Dementia Trialists' Collaboration
4. EU Hernia Trialists Collaboration
5. Digestive Tract Trialists' Collaborative Group
6. Recurrent Miscarriage Immunotherapy Trialists Group
7. Stem Cell Trialists' Collaborative Group
8. Dutch Collaborative Prostaglandin Trialists' Group
9. The Thrombolysis in Acute Stroke Pooling Project
10. The Prospective Pravastatin Pooling (PPP)
11. Zoladex Early Breast Cancer Research Association Trialists Group
12. Casodex Early Prostate Cancer Trialists' Group

3.3.2. Auswahl der Studien

Eingeschlossen wurden empirische Studien, die MA-IPDs und MA-APDs vergleichen. Die empirische Studien sollten folgende Einschlusskriterien erfüllen:

(1) Primärstudien sind RCTs
(2) quantitative Ergebnisse für MA-IPDs und MA-APDs werden dargestellt oder können von relevanten Publikationen extrahiert oder berechnet werden.

Ausgeschlossen wurden empirische Studien, die nur Vergleiche der Strukturen, der Organisation und der Kosten von MA-IPDs und MA-APDs präsentierten.

Es wurden keine Beschränkungen nach Art der Population, der Intervention oder der Endpunkte gesetzt.

Titel und Zusammenfassungen gefundener Studien wurden gesichtet. Bei Zweifel an der Relevanz eines Abstracts wurde der Volltext einbezogen. Relevante Studien, für die keine Volltext-Publikation gefunden wurde, wurden in die Übersicht einbezogen. Die Studien wurden von einer Person ausgewählt. Die Sichtung von Titeln und Abstracts und die Selektion von Studien wurde mit 3 Monaten Abstand von derselben Person wiederholt. Die Zahl der ausgeschlossenen Studien sowie die jeweiligen Gründe für den Ausschluss wurden dokumentiert (s. Abbildung 2).

3.3.3. Extraktion der Daten

Die Daten wurden anhand einer a priori entwickelten Standardform von einer Person extrahiert. Aus den eingeschlossenen Studien wurden Angaben für die folgende Punkte extrahiert:

- Allgemeine Studienmerkmale: Erster Autor, Publikationsjahr, medizinisches Spezialgebiet, Art der Population und der Intervention
- Datengrundlage von MA-IPDs und von MA-APDs: Zahl von RCTs, Zahl der Patienten
- Punktschätzer und Konfidenz-Intervalle der zusammengefassten Effektgrößen von MA-IPDs und MA-APDs
- Gründe für die Unterschiede zwischen den zusammengefassten Effektgrößen von MA-IPDs und MA-APDs

- Heterogenitäts-Test, Methode (nach RCT gewichtete Kombination) und Modell der Synthese (FEM, REM)
- Berücksichtigung der methodologischen Qualität von RCTs (Dimensionen, Scores, Methoden, Ergebnisse)
- Untersuchung der klinischen Heterogenität (Zahl der Kovariablen, Methoden, Ergebnisse)

3.3.4. Statistische Auswertung

Allgemeinmerkmale, Endpunkte und Datengrundlage empirischer MA-IPDs und MA-APDs wurden beschrieben. Verwendete Methoden und Modelle der Synthese sowie die Berichterstattung des Heterogenitäts-Tests wurden dargestellt. Die zusammengefassten Effektgrößen von MA-IPDs und MA-APDs mit identischer sowie mit unterschiedlicher Datengrundlage wurden verglichen. Zudem wurden die Unterschiede zwischen den zusammengefassten Effektgrößen von MA-IPDs und MA-APDs, die auf Publication-Bias, Patient-Exclusion-Bias, Effektmaße und längeres Follow-Up zurückzuführen sind, dargestellt. Bei allen Vergleichen der synthetischen Funktion von MA-IPDs und MA-APDs wurden sowohl die Punktschätzer als auch die Konfidenz-Intervalle der zusammengefassten Effektgrößen kontrastiert. Alle statitische Auswertungen wurden mit der freien Software R (Version 2.2.0) durchgeführt.

Es ist nicht zulässig, die Unterschiede zwischen den einzelnen MA-IPDs und MA-APDs statistisch als unabhängige Effektgrößen zu testen, da beide MAs aus gemeinsamer Datengrundlage geschätzt wurden. Daher wurden die Punktschätzer und die Konfidenz-Intervalle der zusammengefassten Effektgrößen von MA-IPDs und MA-APDs tabelliert und beschrieben. Der t-Test für gepaarte Stichproben wurde benutzt, um die Unterschiede zwischen den Punktschätzern von MA-IPDs und MA-APDs zu testen. Bei diesem Test wurden MA-IPDs und MA-APDs paarweise verbunden und als zwei Messungen an denselben Untersuchungseinheiten (RCTs) betrachtet. Ein zweiseitiger Test mit einem Signifikanzniveau von 5% wurde benutzt. Um die analytische Funktion von MA-IPDs und von MA-APDs zu vergleichen, wurden die Berücksichtigung methodischer Qualität und die Untersuchung klinischer Heterogenität in beiden Arten der MAs dargestellt.

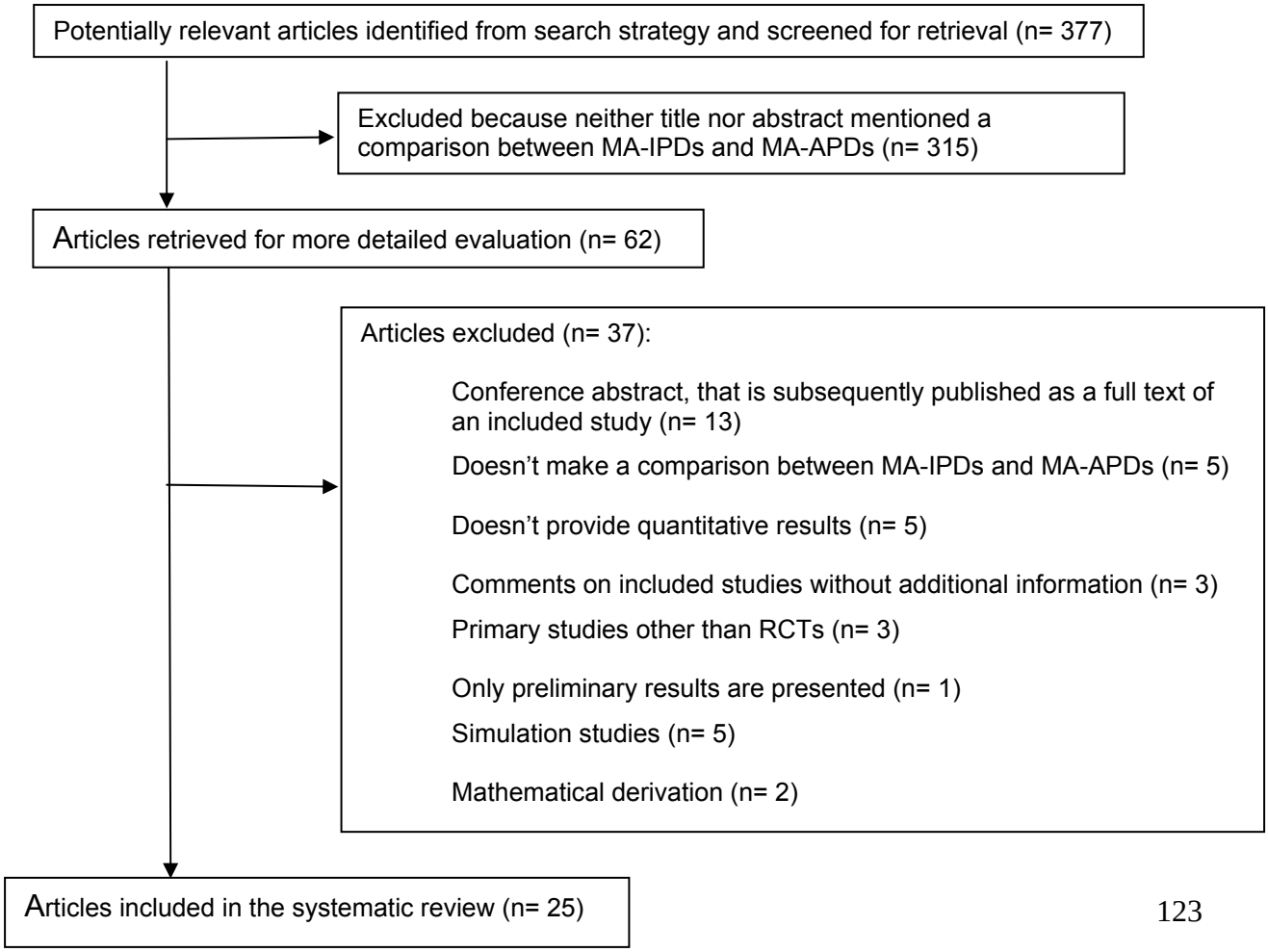
3.4. Ergebnisse

3.4.1. Ergebnisse der Suchen

Die Suchen ergaben 377 Zitationen. Nach der Sichtung von Titeln und Zusammenfassungen wurden 62 als möglicherweise relevant eingestuft. Für 42 von ihnen wurden Volltext-Publikationen beschafft. Für die übrigen 20 Zitationen gab es nur Zusammenfassungen. Medline- und Google-Suchen nach späteren Publikationen fanden Volltext für 13 der 20 Zitationen.

25 empirische Studien, einschließlich einer nicht publizierten empirischen Studie des Verfassers, wurden in die vorliegende SR eingeschlossen. Für 2 empirische Studien wurden nur Zusammenfassungen gefunden [Franzosi, 1997b; Spooner, 1998]. Diese wurden nicht ausgeschlossen. 37 Zitationen wurden ausgeschlossen, die Gründe dafür sind in Abbildung 2 dargelegt. Ein Literaturverzeichnis für alle ausgeschlossenen Studien ist im Anhang 2 zu finden.

Abb. 2 Flussdiagramm zur Auswahl der Studien zum Vergleich von MA-IPDs und MA-APDs



3.4.2. Allgemeine Merkmale eingeschlossener Studien

Es wurden 25 empirische Studien, die mehrere Medizingebiete und Gesundheits-technologien abdecken, in diese SR eingeschlossen (s. Tab. 24). Mehr als ein Drittel der Studien kamen aus der Onkologie (n= 9). Dies war erwartet, da auf diesem Gebiet die meisten MA-IPDs durchgeführt wurden. 4 empirische Studien stellten keine zusammengefassten Effektgrößen von MA-IPD und MA-APD dar [Schmid, 1999; Berlin, 2002; Schmid, 2004; Smith, 2005]. Sie verglichen beide Arten von MA lediglich in Bezug auf ihre analytische Funktion. 70 Vergleiche zwischen den Effektgrößen von MA-IPDs und MA-APDs wurden aus 21 empirischen Studien extrahiert. Fortin stellt 2 Vergleiche dar (2 Endpunkte: Zahl empfindlicher Gelenke und Dauer der Morgensteifheit) [Fortin, 1995], D'Amico 4 (2 Interventionen: topische plus systemische Antibiotikaprophylaxe vs. Kontrolle und nur topische Antibiotikaprophylaxe vs. Kontrolle; je 2 Endpunkte: Mortalität und Infektion der Atemwege) [D'Amico, 1998], Duchateau 2 (2 Endpunkte: Überleben nach 2 Jahren und Überleben nach 5 Jahren) [Duchateau, 2001], McCormack 2 (2 Endpunkte: Rezidiv bei Leistenbruch und persistierende Schmerzen) [McCormack, 2004] (s. Tab. 25). Tierney und Michiels präsentierten jeweils 14 und 13 Vergleiche für diverse onkologische Erkrankungen [Tierney, 2005, Michiels, 2005] (s. Tab. 26). 4 empirische Studien verwendeten APDs, die nicht aus den Publikationen von Primärstudien extrahiert wurden, sondern welche aus der Kumulation von IPDs auf RCT-Ebene resultierten (Pseudo-MA-APDs) [Schmid, 2004; Smith, 2005; Tierney, 2005; Michiels, 2005]. Eine empirische Studie wurde im Rahmen dieser Dissertation durchgeführt (s. Abschnitt 4) [Mukhtar, 2006].

Tab. 24 Allgemeine Merkmale der eingeschlossenen empirischen Studien

Study	Year	Medical Field	Population	Intervention
Stewart	1993	Oncology	Ovarian Cancer	Chemotherapy
Pignon	1993	Oncology	Small-Cell-Lung Cancer	Radiotherapy
Fortin	1995	Rheumatology	Rheumatoid Arthritis	Dietary Supplementation
Jeng	1995	Immunology	Recurrent Miscarriage	Immunotherapy
Le Chevalier	1996	Oncology	Non-Small-Cell-Lung-Cancer	Chemotherapy
Franzosi	1997	Cardiovascular Diseases	Myocardial Infarction	Drugs
Clarke	1998	Oncology	Early Breast Cancer	Chemotherapy
D'Amico	1998	Infectious Diseases	Critically ill Adult Patients	Drugs
Szczech	1998	Immunology	End Stage Renal Disease	Immunotherapy
Spooner	1998	Chest Medicine	Exercise-induced Bronchoconstriction	Drugs
Ioannidis	1999	Infectious Diseases	HIV infected Patients	Drugs
Schmid I	1999	Cardiovascular Diseases	Myocardial Infarction	Drugs
Williamson	2000	Neurology	Epilepsy	Drugs

Tab. 24 (Forts.) Allgemeine Merkmale der eingeschlossenen empirischen Studien

Study	Year	Medical Field	Population	Intervention
Fritz	2000	Oncology	Multiple Myeloma	Chemotherapy
Saillourglenisson	2000	Infectious Diseases	HIV infected Patients	Drugs
Duchateau	2001	Oncology	Advanced Head and Neck Cancer	Chemotherapy
Tudur	2001	Oncology	Metastatic Colorectal Cancer	Chemotherapy
Berlin	2002	Immunology	End Stage Renal Disease	Immunotherapy
McCormack	2004	Surgery	Inguinal Hernia	Surgery
Teramukai	2004	Surgery	Non-Small-Lung Cancer	Chemotherapy
Schmid II	2004	Nephrology	Non-Diabetic Renal Disease	Drugs
Smith	2005	Neurology	Epilepsy	Drugs
Tierney	2005	Oncology	Diverse Cancers	Chemotherapy, Radiotherapy
Michiels	2005	Oncology	Diverse Cancers	Chemotherapy, Radiotherapy
Mukhtar	2006	Cardiovascular Diseases	Cardiovascular Diseases	Drugs

3.4.3. Endpunkte und Datengrundlagen von MA-IPDs und MA-APDs

Bei drei Vierteln aller Endpunkte wurde Mortalität oder Überleben (Zeit bis zum Tod) verwendet (n= 42). Lediglich ein Vergleich weist ein erwünschtes Ereignis als Endpunkt auf [Jeng, 1995a]. Sowohl in MA-IPDs als auch in MA-APDs treten bei 7 Studien das Odds-Ratio (OR), bei 3 das Relative Risiko (RR) und bei 3 das Hazard-Ratio (HR) als Effektmaße auf. In 5 Studien wurden bei der MA-IPD das HR und bei der MA-APD das OR berechnet. Bei einer Studie wurde das HR (MA-IPDs) dem RR (MA-APDs) gegenübergestellt. 2 Studien weisen stetige Endpunkte und dementsprechend Risiko-Differenzen als Effektmaße auf.

Unter Ausschluss der Vergleiche zwischen MA-IPD und Pseudo-MA-APD, da letztere „künstlich“ aus ersteren erzeugt wurden (s. Tab. 26), war die Datengrundlage für 11 Vergleiche zwischen MA-IPD und MA-APD identisch, für 13 unterschiedlich (unterschiedliche RCTs) und für 3 vergleichbar (identische RCTs, aber unterschiedliche Zahl der Patienten) (s. Tab. 25). Die Studien von Szczech [Szczech, 1998] und Berlin [Berlin, 2002], von Williamson [Williamson, 2000] und Smith [Smith, 2005], von Le Chevalier [Le Chevalier, 1996] und ein Vergleich von Michiels (Lung2) [Michiels, 2005] sowie die Studien von Pignon [Pignon, 1993] und ein Vergleich von Michiels (Lung3) [Michiels, 2005] basierten auf denselben MAs, verfolgten aber unterschiedliche Ziele. Mit Ausnahme von drei Vergleichen (Fortin I, Fortin II, Spooner), bei denen eine stetige Zielvariable benutzt wurde, verwendeten alle Vergleiche binäre Endpunkte [Fortin, 1995; Spooner, 1998].

Tab. 25 Endpunkte und Datengrundlagen von MA-IPDs und MA-APDs

Study	Outcome	Data Set RCTs (Patients)	Data Set RCTs (Patients)	Comparability of the Data Set
		MA-IPDs	MA-APDs	
Stewart	Survival, Mortality	7 (835)	8 (788)	Different
Pignon	Survival, Mortality	13 (2103)	11 (1911)[¶]	Different
Fortin I	Tender Joints Counts	10 (395)	10 (368)	Comparable
Fortin II	Duration of Morning Stiffness	10 (395)	10 (368)	Comparable
Jeng	Relative Live-Birth-Ratio	4 (239)	4 (202)	Comparable
Le Chevalier	Survival, Mortality	11 (1190)	8 (712)	Different
Franzosi	Mortality	4 (98496)	4 (98496)	Identical
Clarke	Mortality	5 (1145)	5 (1145)	Identical
D'Amico[a] I	Mortality	12 (2640)	16 (3581)	Different
D'Amico[b] II	Mortality	12 (2160)	17 (2543)	Different
D'Amico[a] III	Respiratory Tract Infection	12 (2641)	15 (2883)	Different
D'Amico[b] IV	Respiratory Tract Infection	12 (2060)	15 (2377)	Different
Szczech	Allograft Failure at 2 Years	5 (628)	7 (794)	Different
Spooner	Forced Expiratory Volume in 1 second	9 (112)	9 (112)	Identical
Ioannidis	Mortality	8 (1792)	8 (1792)	Identical
Schmid I	Mortality	9 (58600)	56 (NR)	Different
Williamson	Withdrawal of Treatment	5 (705)	5 (705)	Identical
Fritz	Response Rate	12 (2469)	17 (2333)	Different
Saillourglenisson	Survival, Mortality	10 (3115)	10 (3115)	Identical
Duchateau[c] I	Survival, Mortality	NR (5049*)	NR (5049*)	Identical
Duchateau[d] II	Survival, Mortality	NR (3937*)	NR (3937*)	Identical
Tudur	Survival	7 (866)	9 (1114)	Different
Berlin	Allograft Failure at 2 Years	5 (628)	5 (628**)	Identical
McCormack[e] I	Hernia Recurrence	37 (NR)	27 (NR)	Different
McCormack[f] II	Persisting Pain	20 (NR)	3 (NR)	Different
Teramukai	Survival, Mortality	10 (1355)	10 (1355)	Identical
Mukhtar	Mortality	14 (90056)	14 (90056)	Identical

a= Interventions: topical plus systematic antibiotic vs. control, b= Interventions: topical antibiotic vs. control
c= Outcome: survival, mortality at 5 years, d= Outcome: survival, mortality at 2 years
e= Outcome: hernia recurrence, f= Outcome: persisting pain
[¶]= Non-randomised clinical trials were also included.
*= number of events, **= IPD aggregated at RCT level
NR= not reported

Tab. 26 Endpunkte und Datengrundlagen von MA-IPDs und Pseudo-MA-APDs

Study	Outcome	Data Set RCTs (Patients)	Data Set RCTs (Patients)	Comparability of the Data Set
Schmid II	Glomerular Filtration Rate	11 (1860)	11 (1860)	Identical
Smith	Remission from seizures	5 (1225)	5 (1225)	Identical
Tierney Bladder	Survival	4 (479)	4 (NR)	Comparable
Tierney Brain	Survival	12 (3004)	12 (NR)	Comparable
Tierney Lung1	Survival	8 (1394)	8 (NR)	Comparable
Tierney Lung2	Survival	6 (668)	6 (NR)	Comparable
Tierney Lung3	Survival	12 (1780)	12 (NR)	Comparable
Tierney Lung4	Survival	8 (777)	8 (NR)	Comparable
Tierney Lung5	Survival	9 (2128)	9 (NR)	Comparable
Tierney Ovary1	Survival	19 (3146)	19 (NR)	Comparable
Tierney Ovary2	Survival	11 (1329)	11 (NR)	Comparable
Tierney Ovary3	Survival	9 (1705)	9 (NR)	Comparable
Tierney Ovary4	Survival	9 (1095)	9 (NR)	Comparable
Tierney Ovary5	Survival	12 (2046)	12 (NR)	Comparable
Tierney Oesophagus	Survival	5 (1147)	5 (NR)	Comparable
Tierney Sarcoma	Survival	14 (1568)	14 (NR)	Comparable
Michiels Lung1	Survival, Mortality	7 (987)	7 (987)	Identical
Michiels Lung2	Survival, Mortality	11 (1190)	11 (1190)	Identical
Michiels Lung3	Survival, Mortality	13 (2536)	13 (2536)	Identical
Michiels Lung4	Survival, Mortality	22 (3033)	22 (3033)	Identical
Michiels Lung5	Survival, Mortality	9 (2128)	9 (2128)	Identical
Michiels Colorectal1	Survival, Mortality	7 (1216)	7 (1216)	Identical
Michiels Colorectal2	Survival, Mortality	9 (1380)	9 (1380)	Identical
Michiels Colorectal3	Survival, Mortality	10 (1544)	10 (1544)	Identical
Michiels Colorectal4	Survival, Mortality	6 (1031)	6 (1031)	Identical
Michiels Colorectal5	Survival, Mortality	8 (1168)	8 (1168)	Identical
Michiels Colorectal6	Survival, Mortality	5 (494)	5 (494)	Identical
Michiels Oesophagus	Survival, Mortality	6 (1147)	6 (1147)	Identical
Michiels Glioma	Survival, Mortality	12 (3004)	12 (3004)	Identical

3.4.4. Vergleich der Synthetischen Funktion von MA-IPDs und MA-APDs

3.4.4.1. Methode und Modell der Synthese bei MA-IPDs und MA-APDs

Alle eingeschlossenen MA-IPDs und MA-APDs führten eine nach RCT gewichtete Synthese durch. Dies ist von besonderer Bedeutung für MA-IPDs, da eine simultane Kombination der

IPDs in einem Datensatz ohne Berücksichtigung der RCTs, aus denen sie stammen, die Heterogenität zwischen den RCTs ausblendet.

Mit Ausnahme von 2 MA-IPDs wurde das Intention-To-Treat-Prinzip bei der Synthese beachtet. Bei Jeng wurde das Prinzip verletzt und bei D'Amico konnte das Prinzip nur bei einem Teil der RCTs befolgt werden [Jeng, 1995a; D'Amico, 1998]. Die Synthese erfolgte bei 3 Paaren von MA-IPDs und MA-APDs sowohl im FEM als auch im REM [Jeng, 1995a; Ioannidis, 1999; Smith, 2005] und bei 3 Paaren im REM [Fortin, 1995; Spooner, 1998; Schmid, 2004]. In den Studien von Szczech, Saillourglenisson und Mukhtar wurden die RCTs bei MA-IPDs im FEM und bei MA-APDs im FEM und REM kombiniert [Szczech, 1998; Saillourglenisson, 2000; Mukhtar, 2006]. In einer Studie wurde MA-IPD in FEM und MA-APD in REM durchgeführt [Schmid, 1999]. Bei allen weiteren empirischen Studien erfolgte die Synthese in beiden Arten von MA im FEM (n= 15).

3.4.4.2. Zufallsbedingte Heterogenität bei MA-IPDs und MA-APDs

12 MA-IPDs [Stewart, 1993; Jeng, 1995a; Le Chevalier, 1996; Franzosi, 1997b; Williamson, 2000; Fritz, 2000; Saillourglenisson, 2000; McCormack, 2004; Teramukai, 2004; Smith, 2005; Tierney, 2005; Michiels, 2005] und 13 MA-APDs [Stewart, 1993; Fortin, 1995; Jeng, 1995a; Franzosi, 1997b; Ioannidis, 1999; Williamson, 2000; Fritz, 2000; Saillourglenisson, 2000; Teramukai, 2004; Smith, 2005; Tierney, 2005; Michiels, 2005; Mukhtar, 2006] stellten die Ergebnisse eines Heterogenitäts-Tests oder eines Heterogenitäts-Maßes dar. Da bei fast der Hälfte der Studien jedoch keine diesbezüglichen Angaben präsentiert wurden, ist es unmöglich, die Berücksichtigung der statistischen Heterogenität bei der Auswahl des Synthesemodells (z.B. Synthese im REM beim p-Wert des Cochran-Testes < 0.1) in MA-IPDs und MA-APDs zu vergleichen.

3.4.4.3. Effektgrößen von MA-IPDs und MA-APDs mit identischen Datengrundlagen

10 Vergleiche zwischen MA-IPDs und MA-APDs basierten auf identischen Datengrundlagen. In 4 Vergleichen wurde die zusammengefasste Effektgröße durch MA-APDs und in 3 Vergleichen durch MA-IPDs überschätzt. Bei 3 Vergleichen war kein Unterschied zwischen den Punktschätzern von MA-IPDs und MA-APDs zu verzeichnen (s. Tab. 27). Während MA-APDs ein breiteres Konfidenz-Intervall der zusammengefassten Effektgröße in 5 Vergleichen

aufwiesen, zeigten MA-IPDs ein breiteres Konfidenz-Intervall in 3 Vergleichen. In 2 Vergleichen waren die Konfidenz-Intervalle von MA-IPDs und MA-APDs identisch.

Tab. 27 Effektgrößen von MA-IPDs und MA-APDs mit identischen Datengrundlagen

Study	MA-IPDs			MA-APDs		
	Measure	Estimate*	95% CI¶	Measure	Estimate*	95% CI¶
Franzosi	OR	*0.93*	*0.89 – 0.98*	OR	*0.93*	*0.89 – 0.98*
Clarke	OR	**0.72**	0.61– 0.86	OR	0.81	**0.63 – 1.03**
Spooner	WMD	*16.00*	**13.10 – 19.90**	WMD	*16.00*	12.70 – 19.20
Ioannidis	HR	0.79	0.66 – 0.93	OR	**0.75**	**0.57 – 1.00**
Williamson	HR	**1.02**	**0.67 – 1.56**	HR	1.03	0.79 – 1.33
Saillourglenisson	HR	1.12	0.99 – 1.27	OR	**1.10**	**0.93 – 1.29**
Duchateau I	HR	0.88	0.83 – 0.93	OR	**0.83**	**0.76 – 0.91**
Duchateau II	HR	0.88	0.83 – 0.94	OR	**0.87**	**0.79 – 0.95**
Teramukai	HR	**0.80**	**0.69 – 0.93**	RR[1]	0.86	0.77 – 0.95
Mukhtar	RR	*0.88*	*0.84 – 0.91*	RR	*0.88*	*0.84 – 0.91*

* Larger effect size is bolded.
¶ Larger confidence interval is bolded.
Identical effect sizes and identical confidence intervals are italic.
OR= odds ratio, RR= relative risk, HR= hazard ratio, WMD= weighted mean difference, CI= confidence interval
[1]= Calculated by the author (Mukhtar) by the Mantel-Haenszel Method *

3.4.4.4. Effektgrößen von MA-IPDs und MA-APDs mit unterschiedlichen Datengrundlagen

15 Vergleiche zwischen MA-IPDs und MA-APDs basierten auf nicht-identischen Datengrundlagen (vergleichbar oder unterschiedlich). In 9 Vergleichen wurde die zusammengefasste Effektgröße durch MA-APDs und in 5 Vergleichen durch MA-IPDs überschätzt. Bei lediglich einem Vergleich war kein Unterschied zwischen den Punktschätzern von MA-IPDs und MA-APDs zu verzeichnen (s. Tab. 28). Während MA-APDs ein breiteres Konfidenz-Intervall der zusammengefassten Effektgröße in 7 Vergleichen aufwiesen, zeigten MA-IPDs ein breiteres Konfidenz-Intervall in 8 Vergleichen.

Tab. 28 Effektgrößen von MA-IPDs und MA-APDs mit unterschiedlichen Datengrundlagen

Study	MA-IPDs			MA-APDs		
	Measure	Estimate	95% CI	Measure	Estimate	95% CI
Stewart	OR	**0.70**	0.51 – 0.94	OR	0.71	**0.52 – 0.96**
Pignon	OR	**0.64**	0.51 – 0.79	OR	0.65	**0.53 – 0.83**
Fortin I	RD	-2.30	**-3.53 – -1.07**	RD	**- 2.90**	-3.80 – -2.10
Fortin II	RD	-18.30	-33.59 – -3.01	RD	**-25.90**	**-44.30 – -7.50**
Jeng	RR	1.18	0.98 – 1.42	RR	**1.38**	**0.89 – 1.87**

Tab. 28 (Forts.) Effektgrößen von MA-IPDs und MA-APDs mit unterschiedlichen Datengrundlagen

Study	MA-IPDs			MA-APDs		
	Measure	Estimate	95% CI	Measure	Estimate	95% CI
Le Chevalier	HR[2]	0.84	0.74 – 0.95	OR	**0.44**	**0.32 – 0.59**
D'Amico I	OR	**0.79**	**0.65 – 0.97**	OR	0.80	0.69 – 0.93
D'Amico II	OR	1.02	**0.81 – 1.30**	OR	**1.01**	0.84 – 1.22
D'Amico III	OR	0.40	**0.33 – 0.49**	OR	**0.35**	0.29 – 0.41
D'Amico IV	OR	0.61	**0.49 – 0.75**	OR	**0.56**	0.46 – 0.68
Szczech	RR	*0.69*	**0.47 – 1.01**	RR	*0.69*	0.49 – 0.97
Fritz	OR	0.83	**0.71 – 0.98**	OR	**0.77**	0.65 – 0.91
Tudur	HR	**0.65**	**0.56 – 0.76**	HR	0.74	0.65 – 0.84
McCormack I	OR	0.81	0.61 – 1.08	OR	**0.76**	**0.55 – 1.04**
McCormack II	OR	**0.54**	0.46 – 0.64	OR	2.03	**1.03 – 4.01**

* Larger effect size is bolded.
¶ Larger confidence interval is bolded.
Identical effect sizes and identical confidence intervals are italic.
OR= odds ratio, RR= relative risk, HR= hazard ratio, RD= rate difference, CI= confidence interval
[2] Calculated by the author (Mukhtar) by the Peto Method

3.4.5. Einfluss von Publication-Bias

Es wurden 5 Vergleiche extrahiert zwischen MA-IPDs, die Daten aus publizierten und nicht publizierten RCTs einschlossen, und MA-APDs, die lediglich auf Daten aus publizierten RCTs zurückgriffen. In 3 Vergleichen wurde die zusammengefasste Effektgröße durch MA-APDs und in 2 Vergleichen durch MA-IPDs überschätzt (s. Tab. 29). Während MA-APDs ein breiteres Konfidenz-Intervall der zusammengefassten Effektgröße in 4 Vergleichen aufwiesen, zeigten MA-IPDs ein breiteres Konfidenz-Intervall in lediglich einem Vergleich.

Tab.29 Einfluss von Publication-Bias

Study	Based on "All" RCTs			Based on Published RCTs		
	Measure	Estimate	95% CI	Measure	Estimate	95% CI
Stewart[‡]	OR[‡]	0.75	0.59 – 0.96	OR	**0.71**	**0.52 – 0.96**
Jeng	RR	1.12	0.97 – 1.31	RR	**1.38**	**0.89 – 1.87**
Clarke	OR	**0.76**	0.65 – 0.88	OR	0.81	**0.63 – 1.03**
Ioannidis	OR	0.64	**0.47 – 0.86**	OR	**0.43**	0.28 – 0.66
Duchateau	HR	**0.88**	0.84 – 0.92	HR	0.89	**0.84 – 0.94**

* Larger effect size is bolded. ¶ Larger confidence interval is bolded.
Identical effect sizes and identical confidence intervals are italic.
OR= odds ratio, RR= relative risk, HR= hazard ratio, CI= confidence interval
[‡] One published trial was not included in the IPDs

3.4.6. Einfluss von Patient-Exclusion-Bias

Es wurden 19 Vergleiche extrahiert zwischen MA-IPDs, die alle randomisierten Patienten einschlossen und somit eine Intention-To-Treat-Analyse durchführten, und MA-APDs, die einen Teil der randomisierten Patienten von der Auswertung ausschlossen. In 14 Vergleichen wurde die zusammengefasste Effektgröße durch MA-APDs und in 2 Vergleichen durch MA-IPDs überschätzt. Bei 3 Vergleichen war kein Unterschied zwischen den Punktschätzern von MA-IPDs und MA-APDs zu verzeichnen (s. Tab. 30). Während MA-APDs ein breiteres Konfidenz-Intervall der zusammengefassten Effektgröße in 10 Vergleichen aufwiesen, zeigten MA-IPDs ein breiteres Konfidenz-Intervall in 4 Vergleichen. In 5 Vergleichen waren die Konfidenz-Intervalle von MA-IPDs und MA-APDs identisch.

Tab.30 Einfluss von Patient-Exclusion-Bias

Study	Based on "All" Patients			Based on "Included" Patients		
	Measure	Estimate	95% CI	Measure	Estimate	95% CI
Stewart	OR	0.79	**0.62 – 1.00**	OR	**0.75**	0.59 – 0.96
Fortin I	RD	-2.3	**-3.53 – -1.07**	RD	**- 2.9**	-3.8 – -2.1
Fortin II	RD	-18.3	-33.59 – -3.01	RD	**-25.9**	**-44.3 – -7.5**
Jeng	RR	1.18	0.98 – 1.42	RR	**1.38**	**0.89 – 1.87**
Duchateau	HR	0.90	**0.85 – 0.94**	HR	**0.88**	0.84 – 0.92
Tierney Bladder	HR	1.02	0.81 – 1.28	HR	**0.99**	**0.78 – 1.26**
Tierney Brain	HR	0.85	*0.78 – 0.92*	HR	**0.84**	*0.77 – 0.91*
Tierney Lung1	HR	**0.87**	0.74 – 1.02	HR	0.88	**0.75 – 1.04**
Tierney Lung2	HR	0.94	*0.79 – 1.11*	HR	**0.93**	*0.79 – 1.11*
Tierney Lung3	HR	*0.87*	0.79 – 0.96	HR	*0.87*	**0.78 – 0.96**
Tierney Lung4	HR	*0.73*	*0.63 – 0.86*	HR	*0.73*	*0.62 – 0.85*
Tierney Lung5	HR	1.21	1.08 – 1.13	HR	**1.19**	**1.07 – 1.33**
Tierney Ovary1	HR	0.98	0.91 – 1.06	HR	**0.97**	**0.89 – 1.05**
Tierney Ovary2	HR	*0.93*	**0.83 – 1.05**	HR	*0.93*	0.83 – 1.04
Tierney Ovary3	HR	0.88	*0.79 – 0.98*	HR	**0.87**	*0.78 – 0.97*
Tierney Ovary4	HR	0.91	*0.80 – 1.05*	HR	**0.90**	*0.79 – 1.04*
Tierney Ovary5	HR	1.02	0.93 – 1.12	HR	**1.00**	**0.91 – 1.11**
Tierney Oesophagus	HR	**0.89**	0.78 – 1.01	HR	0.90	**0.79 – 1.03**
Tierney Sarcoma	HR	0.90	0.77 – 1.04	HR	**0.85**	**0.72 – 1.00**

* Larger effect size is bolded.
¶ Larger confidence interval is bolded.
Identical effect sizes and identical confidence intervals are italic.
OR= odds ratio, RR= relative risk, HR= hazard ratio, RD= rate difference, CI= confidence interval

3.4.7. Einfluss der Effektmaße

Es wurden 16 Vergleiche extrahiert zwischen MA-IPDs, die das Hazard Ratio als Effektmaße verwendeten, und MA-APDs, die das OR als Effektmaße benutzten. In 13 Vergleichen wurde die zusammengefasste Effektgröße durch MA-APDs und in 3 Vergleichen durch MA-IPDs überschätzt (s. Tab. 31). Bei allen Vergleichen zeigten die MA-APDs ein breiteres Konfidenz-Intervall als die MA-IPDs.

Tab. 31 Einfluss der Effektmaße

Study	Hazard Ratio		OR	
	Estimate	95% CI	Estimate	95% CI
Stewart	0.88	0.77 – 1.00	**0.79**	**0.62 – 1.00**
Pignon	0.83	0.76 – 0.92	**0.64**	**0.51 – 0.79**
Duchateau	0.88	0.83 – 0.94	**0.87**	**0.79 – 0.95**
Michiels Lung1	0.84	0.73 – 0.97	**0.83**	**0.65 – 1.08**
Michiels Lung2	0.84	0.74 – 0.95	**0.70**	**0.53 – 0.93**
Michiels Lung3	0.86	0.79 – 0.94	**0.84**	**0.71 – 0.99**
Michiels Lung4	0.90	0.83 – 0.97	**0.87**	**0.75 – 1.00**
Michiels Lung5	**1.21**	1.08 – 1.34	1.29	**1.06 – 1.58**
Michiels Colorectal1	0.88	0.78 – 0.99	**0.84**	**0.67 – 1.06**
Michiels Colorectal2	0.97	0.86 – 1.09	**0.84**	**0.68 – 1.05**
Michiels Colorectal3	**0.98**	0.88 – 1.09	0.99	**0.81 – 1.22**
Michiels Colorectal4	**1.16**	1.01 – 1.33	1.27	**0.99 – 1.62**
Michiels Colorectal5	0.87	0.77 – 0.98	**0.77**	**0.60 – 0.97**
Michiels Colorectal6	0.74	0.61 – 0.89	**0.47**	**0.33 – 0.67**
Michiels Oesophagus	0.89	0.78 – 1.01	**0.86**	**0.68 – 1.10**
Michiels Glioma	0.85	0.78 – 0.92	**0.76**	**0.66 – 0.88**

* Larger effect size is bolded.
¶ Larger confidence interval is bolded.
Identical effect sizes and identical confidence intervals are italic.
CI= confidence interval

3.4.8. Einfluss von längerem Follow-Up

5 Vergleiche berichteten über den Unterschied zwischen MA-IPDs mit längerem Follow-Up und MA-APDs ohne dieses [Stewart, 1993; Clarke, 1998b; Szczech, 1998; Ioannidis, 1999; Duchateu, 2001]. Allerdings wurde das Konfidenz-Intervall von MA-IPDs mit längerem Follow-Up nur in 2 Vergleichen dargestellt [Ioannidis, 1999; Duchateu, 2001]. Eine Überschätzung des Interventionseffekts durch MA-APDs mit kürzerem Follow-Up wurde bei 3 Studien beobachtet [Stewart, 1993; Szczech, 1998; Ioannidis, 1999]. Während Clarke das Gegenteil (Unterschätzung des Effekts durch MA-APDs mit kürzerem Follow-Up) heraus

fand [Clarke, 1998b], beobachtete Duchateu keinen Unterschied zwischen MA-IPDs und MA-APDs [Duchateu, 2001].

3.4.9. Zusammenfassung des Vergleichs der synthetischen Funktion von MA-IPDs und MA-APDs

Insgesamt wurden 70 Vergleiche zwischen den zusammengefassten Effektgrößen von MA-IPDs und MA-APDs analysiert (s. Tab. 32). Weder in Bezug auf die Punktschätzer noch in Bezug auf die Konfidenz-Intervalle der untersuchten Vergleiche ergab sich ein konsistentes Ergebnis. In zwei Dritteln aller Vergleiche wurde ein größerer Behandlungseffekt (n= 46) oder ein breiteres Konfidenz-Intervall (n= 42) bei MA-APDs gefunden. Patient-Exlusion-Bias und die Verwendung unterschiedlicher Effektmaße resultierten oft in überschätzten und weniger präzisen zusammengefassten Effektgrößen von MA-APDs. Der Einfluss von Publication-Bias auf die zusammengefasste Effektgröße von MA-APDs war nicht konsistent.

In Anlehnung an ähnliche emprische Vergleiche, die zwei Ergebnisse aus zwei verbundenen Stichproben beziehen, z.B. ein Vergleich der Ergebnisse von einer „alten“ und einer aktualisierten MA zu derselben Fragestellung [Shojania, 2007], wurde *post hoc* ein relativer Unterschied zwischen den Punktschätzern einer MA-IPD und einer MA-APD (Differenz der Punktschätzer dividiert durch den Punktschätzer der MA-IPD), der mindestens 50% betrug, als ein relevanter Unterschied betrachtet. Allerdings war ein solcher Unterschied nur bei einem Vergleich (McCormack II) [McCormack, 2004] zu finden.

Bemerkenswert ist die Tatsache, dass eine Überschätzung des Effekts und eine Reduzierung der Präzision nicht immer gleichzeitig bei MA-APDs zu beobachten waren. Bei 9 Vergleichen wiesen MA-APDs einen größeren Effekt, aber ein kleineres Konfidenz-Intervall auf und bei 11 Vergleichen zeigten sie das Gegenteil. Nach Ausschluss von 4 Vergleichen mit stetigen oder erwünschten Endpunkten (Fortin I, Fortin II, Jeng, Spooner) [Fortin, 1995; Jeng, 1995a; Spooner, 1998], wurde ein Paired-t-Test für die Vergleiche von MA-IPDs und MA-APDs durchgeführt. Bei Endpunkten, für die mehr als ein Vergleich gezogen wurde (z.B. erster Vergleich: zwischen der MA-IPD mit publizierten und nicht publizierten RCTs und der MA-APD mit lediglich publizierten RCTs, zweiter Vergleich: zwischen der MA-IPD mit allen randomisierten Patienten und der MA-APD mit lediglich einem Teil der randomisierten Patienten), wurde für jeden Endpunkt nur ein Vergleich in das Testverfahren einbezogen. Der Test zeigte keinen signifikanten Unterschied zwischen den Paaren von MA-IPDs und MA-APDs ($t= 0.1358$, $df = 47$, p-Wert = 0.89).

Tab. 32 Überschätzung der Punktschätzer und Reduzierung der Präzision

	Identical Dataset	Different Dataset	Publication Bias	Patient Exclusion Bias	Effect Measure	Follow-Up
Point Estimates						
Overestimation by MA-APDs	4	9	3	14	13	3
Overestimation by MA-IPDs	3	5	2	2	3	1
No Difference	3	1	0	3	0	1
Confidence Intervals						
Wider CI of MA-APDs	5	7	4	10	16	-
Wider CI of MA-IPDs	3	8	1	4	0	-
No Difference	2	0	0	5	0	-

3.4.10. Vergleich der analytischen Funktion von MA-IPDs und MA-APDs

3.4.10.1. Berücksichtigung methodischer Qualität

In 13 Vergleichen wurde die methodische Heterogenität berücksichtigt, wobei Subgruppen-Analyse als die häufigste Methode eingesetzt wurde (s. Tab. 33). Lediglich bei einem Vergleich wurden Unterschiede zwischen RCTs mit hoher und solchen mit niedriger methodischer Qualität gefunden (D'Amico I). D'Amico fand eine Unterschätzung des Effekts durch RCTs ohne Doppel-Verblindung [D'Amico, 1998].

3.4.10.2. Untersuchung klinischer Heterogenität

Die klinische Heterogenität wurde bei 13 Vergleichen nur anhand von MA-IPD und bei 1 Vergleich nur mittels MA-APD untersucht. Lediglich bei 9 Vergleichen wurde sie durch beide Arten von MA geprüft (s. Tab. 34). In einer Studie wurde eine Kovariable durch MA-IPD und MA-APD (Zeit bis zur Behandlung), eine nur durch MA-IPD (Alter) und eine lediglich durch MA-APD (APSAC „Anisoylated Plasminogen Streptokinase Activator Complex„, was ein Thrombolytikum darstellte) als signifikant identifiziert [Schmid, 1999]. Allerdings wurde in der MA-IPD nur eine RCT einbezogen, in der APSAC verwendet wurde. Die Zahl von CD4-T-Lymphozyten wurde in einem weiteren Vergleich als signifikanter Effektmodifikator mittels MA-APD, aber nicht durch MA-IPD gefunden [Saillourglenisson, 2000]. Der Typ der Prophylaxe wurde in diesem Vergleich nur von MA-IPD untersucht und als ein weiterer signifikanter Effektmodifikator ermittelt. Allerdings führte der Ausschluss einer RCT dazu, dass keine der geprüften Kovariablen in beiden Arten von MA weiter einen signifikanten Einfluss auf den Interventionseffekt zeigte. Berlin fand bei MA-IPD aus 6 untersuchten

Kovariablen einen Unterschied bezüglich eines Patientencharakeristikums [Berlin, 2002]. Dies konnte nicht durch MA-APDs identifiziert werden. Bei Teramkuai fanden weder MA-IPDs noch MA-APDs einen Unterschied bezüglich der untersuchten Kovariablen, wobei nach dem Ausschluss von 2 RCTs das Ergebnis der IPD-Regression robust blieb und die APD-Regression die Richtung änderte [Teramkuai, 2004]. Der Vergleich von McCormack bezüglich des Endpunkts Leistenbruch-Rezidiv zeigte keinen Unterschied in der Gesamtpopulation und in Subgruppen der Patienten [McCormack, 2004]. Allerdings wiesen MA-IPD und MA-APD bei McCormack qualitative Unterschiede (Richtungsunterschied) bezüglich des Endpunkts persistierende Schmerzen sowohl im Gesamtkollektiv als auch in den klinischen Subgruppen auf. Während Unterschiede zwischen den Patienten bezüglich der Kovariablen Harneiweiß in einer weiteren MA-IPD gefunden wurde, konnte diese nicht von der MA-APD identifiziert werden [Schmid, 2004]. In der Studie von Smith wurde eine signifikante Kovariable durch MA-IPD und MA-APD (Alter) und eine signifikante Kovariable (Zeit von epileptischem Anfall bis zur Randomisierung) nur anhand von MA-APD ermittelt [Smith, 2005].

3.5. Zusammenfassung und Diskussion

Die vorliegende SR identifiziert und bewertet 70 Vergleiche zwischen den Effektgrößen von MA-IPDs und MA-APDs aus 25 empirischen Studien. Bei zwei Dritteln der Vergleiche wurde eine Tendenz zur Überschätzung der Effektgröße und zur Reduzierung der Präzision bei MA-APDs im Vergleich zu MA-IPDs beobachtet. Diese Tendenz war oft auf den Einfluss der Patient-Exclusion-Bias und der Verwendung unterschiedlicher Effektmaße, Odds Ratio bei MA-APDs und Hazard Ratio bei MA-IPDs, zurückzuführen. Allerdings waren die relativen Unterschiede zwischen den Punktschätzern von MA-IPDs und MA-APDs in allen Vergleichen, mit einer Ausnahme [McCormack, 2004], kleiner als 50%. Bei diesem Vergleich war die Diskrepanz zwischen den Datengrundlagen für MA-IPD (RCT= 20) und für MA-APD (RCTs=3) groß. Allerdings ergab der Paired-t-Test keinen signifikanten Unterschied zwischen den beiden Arten von MA. Dies kann mit der geringen Zahl der Vergleiche oder den geringen Unterschieden zwischen den Punktschätzern von MA-IPDs und MA-APDs zusammenhängen. Eine mathematische Ableitung wies darauf hin, dass unter Annahme der Homogenität von RCTs, die zusammengefassten Effektgrößen von MA-IPDs und MA-APDs sich nicht unterscheiden [Olkin, 1998]. Mathew und Nordstrom leiten mathematisch auch keinen Unterschied her, wenn Heterogenität zwischen primären Studien besteht [Mathew, 1999].

Da in fast der Hälfte der Studien keine Ergebnisse des Heterogenitäts-Tests berichtet werden, können keine Aussagen über die Berücksichtigung zufallsbedingter Heterogenität bei der Synthese gemacht werden.

Die meisten empirischen Studien zielen nicht darauf ab, MA-IPDs und MA-APDs bezüglich der Untersuchung von Heterogenitätsursachen zu vergleichen. Die methodische Qualität wurde bei einem Viertel der Studien in beiden Arten von MAs berücksichtigt, wobei sie keinen konsistenten Einfluss auf die Ergebnisse der MAs zeigte. Die klinische Heterogenität wurde bei einem Drittel der Studien anhand beider Arten von MAs untersucht. Dabei sind diesbezüglich keine konsistenten Unterschiede zwischen MA-IPDs und MA-APDs zu verzeichnen.

Die vorliegende Arbeit ist die erste abgeschlossene SR der empirischen Vergleiche von MA-IPDs und MA-APDs. Clarke und Stewart legten ähnliche Vorhaben vor, die Vergleiche zwischen MA-IPDs und MA-APDs publizierter RCTs systematisch zu bewerten [Clarke, 1997]. Für diese bislang nicht abgeschlossene Cochrane SR liegt lediglich ein Protokoll vor, in dem über nur 5 empirische Vergleiche, ohne quantitative Ergebnisse zu nennen, berichtet wurde [Clarke, 1997]. Eine SR schloss 35 MA-IPDs und 37 MA-APDs ein, die für das klinische Spezialgebiet, den Interventionstyp und die Effektmaße gepaart (matched) wurden [Koopman, 2007]. Diese SR zeigte, dass MA-IPDs und MA-APDs eine ähnliche Anzahl von Subgruppen-Analysen durchführten und dass bei über zwei Dritteln der MA-IPDs kein Interaktions-Test für Subgruppenunterschiede unternommen wurde [Koopman, 2007]. Leider stellte diese SR keine Vergleiche der Ergebnisse von Subgruppen-Analysen anhand von MA-IPD und MA-APD dar.

Die vorliegende SR weist mehrere Einschränkungen auf. Die Identifikation von MA-IPDs gilt als schwierig [Pignon, 2001; Flanagin, 2002] und somit können relevante Studien durch die verwendete Suchstrategie unidentifiziert geblieben sein. Durch Kontakt mit Experten und Handsuche können möglicherweise weitere Vergleiche gefunden werden. Eine weitere Limitation der SR ist, dass sie von einer einzigen Person durchgeführt wurde. Allerdings wurde die Selektion der Studien nach 3 Monaten Abstand von derselben Person wiederholt.

Die Ergebnisse dieser SR deuten darauf hin, dass MA-IPDs und MA-APDs bezüglich der synthetischen Funktion von MA keine bedeutsamen Unterschiede aufweisen. Anhand der vorhandenen Evidenz kann keine Aussage über die Unterschiede zwischen MA-IPDs und MA-APDs bezüglich der analytischen Funktion von MA gemacht werden.

Tab. 33 Berücksichtigung methodischer Qualität bei MA-IPDs und MA-APDs

Study	Dimensions [Method]		Ergebnisse	
	MA-IPD	MA-APD	MA-IPD	MA-APD
Fortin I	Chalmers-Score [DS], DoB [IN]	Chalmers-Score [DS], DoB [IN]	RCTs weisen ähnliche Scores auf, DoB als Einschlusskriterium	RCTs weisen ähnliche Scores auf, DoB als Einschlusskriterium
Fortin II	Chalmers-Score [DS], DoB [IN]	Chalmers-Score [DS], DoB [IN]	RCTs weisen ähnliche Scores auf, DoB als Einschlusskriterium	RCTs weisen ähnliche Scores auf, DoB als Einschlusskriterium
Le Chevalier	AllCon [IN]	Chalmers-Score [DS]	AllCon als Einschlusskriterium	RCTs weisen ähnliche Scores auf
D'Amico I	AllCon [SB], DoB [SB]	AllCon [SB], DoB [SB]	kein Unterschied bzgl. AllCon, RCTs mit DoB überschätzen den Effekt	kein Unterschied bzgl. AllCon, RCTs mit DoB überschätzen den Effekt
D'Amico IV	AllCon [SB], DoB [SB]	AllCon [SB], DoB [SB]	kein Unterschied bzgl. AllCon, kein Unterschied bzgl. DoB	kein Unterschied bzgl. AllCon, kein Unterschied bzgl. DoB
Szczech	Keine	Cho-Score [SA]	Keine	Kein Unterschied bzgl. Score
Schmid I	Keine	DuR [HBR]	Keine	Kein Unterschied bzgl. DuR
Fritz	Trial Size [SB]	Keine	Kleine RCTs überschätzen den Effekt	Keine
Saillourglenisson	Keine	DuR [MR]. ContG [MR]	Keine	kein Unterschied bzgl. DuR kein Unterschied bzgl. ContG
McCormack I	AllCon [TH]	AllCon [TH]	kein Unterschied bzgl. AllCon	Kein Unterschied bzgl. AllCon
McCormack II	AllCon [TH]	AllCon [TH]	kein Unterschied bzgl. AllCon	Kein Unterschied bzgl. AllCon
Schmid II	SnB [HBR], DuR [HBR], ContG [HBR]	SnB [HBR], DuR [HBR], ContG [HBR]	Kein Unterschied bzgl. SnB Kein Unterschied bzgl. DuR Kein Unterschied bzgl. ContG	Kein Unterschied bzgl. SnB Kein Unterschied bzgl. DuR Kein Unterschied bzgl. ContG
Mukhtar	Keine	Jadad-Score [SB, MR], AllCon [SB, MR], ITT [SB, MR]	Keine	kein Unterschied bzgl. Score, kein Unterschied bzgl. AllCon, kein Unterschied bzgl. ITT-Analyse

NR= not reported, DS= Description, IN= Inclusion criterion, SA= Sensitivity analysis, SB= Subgroup analysis, TH= Quality as a threshold, MR= Meta-regression, HBR= Hierarchical Bayesian Regression Rand= Randomization, AllCon= appropriate Allocation Concealment, DoB= Double blinding, SnB= Single blinding, ITT= Intention-to-treat analysis, DuR= Duration of Follow-Up, ContG= Type of Control Group

Tab. 34 Untersuchung klinischer Heterogenität bei MA-IPDs und MA-APDs

Study	Nr. of Covariates		Method of Analysis [significant covariates]	
	MA-IPD	MA-APD	MA-IPD	MA-APD
Fortin I	7	None	LR [Sex]	-
Fortin II	7	None	LR [Sex]	-
Jeng	2	None	LR [Adjusted RR= 1.17 (1.01-1.36) vs. Unadjusted RR= 1.18 (0.98-1.42)]	-
Le Chevalier	6	None	SB [None]	-
Franzosi	12	None	SB [Age, Heart rate, Location of MI]	-
D'Amico I	2	None	SB [None]	-
D'Amico II	2	None	SB [None]	-
D'Amico III	2	None	SB [None]	-
D'Amico IV	2	None	SB [None]	-
Szczech	7	None	LR [Adjusted RR= 1.13 (0.72-1.78) vs. Unadjusted RR= 0.69 (0.47-1.01)], SB [Presensitization]	-
Ioannidis	6	None	CR [Adjusted HR= 0.78 (0.65-0.93) vs. Unadjusted HR= 0.79 (0.66-0.93)]	-
Schmid I	8	11	SB† [Age, Time to treat]	HBR [Time to treat + Use of APSAC]
Williamson	8	None	CR [NR]	-
Fritz	15	None	SB [B2-Microglobulin]	-
Saillourglenisson	4	6	SB [Type of Prophylaxis]	MR [Baseline CD4 count¶]
Tudur	1	1	SB [None]	SB [None]
Berlin	6	7	GR [Presensitization]	GR, HBR [None]
McCormack I	2	2	SB [TAPP vs. Non-Mesh: OR= 0.45 (0.28-0.72)]	SB [TAPP vs. Non-Mesh: OR= 0.56 (0.33-0.93)]
McCormack II	2	2	SB [TAPP vs. Mesh: OR= 0.59 (0.43-0.83), TAPP vs. Non-Mesh: OR= 0.35 (0.24-0.50)]	SB [TAPP vs. Mesh: OR= 2.21 (0.63-7.74), TAPP vs. Non-Mesh: OR= 2.06 (1.00-4.27)]
Teramukai	1	1	CR [None]	MR [None]
Schmid II	6	6	HBR [Baseline urine protein]	HBR [None]
Smith	5	5	CR [Age]	MR [Age + time from first ever seizure to randomization)]
Mukhtar	None	3	-	SB [Proportion of women], MR [Proportion of women]

NR= not reported, LR= Linear regression, SB= Subgroup analysis, CR= Cox regression, GR= Logistic regression, MR= Meta-regression, HBR= Hierarchical Bayesian Regression

† Only one RCT in this MA-IPD used APSAC, ¶ Type of Prophylaxis wasn't included in the meta-regression

4. Berücksichtigung von zufallsbedingter, methodischer und klinischer Heterogenität in Meta-Analyse: Meta-Analyse zu Statinen als Fallstudie

4.1. Hintergrund

Die Unterschiede zwischen den Effektgrößen der in einer MA einbezogenen RCTs zu einer standardisierten medizinischen Intervention können durch Zufall, Variation methodischer Qualität der RCTs (methodische Heterogenität) oder unterschiedliche Charakteristika der in den Studien eingeschlossenen Patienten (klinische Heterogenität) verursacht werden. Das Ausmaß der Heterogenität in einer MA, das über den Zufall hinaus geht, kann anhand von I^2 nach Higgins und Thompson geschätzt werden [Higgins, 2003]. Erst unter Berücksichtigung der methodischen Variationen zwischen den Primärstudien kann die klinische Heterogenität in der MA untersucht werden. Durch die Berücksichtigung heterogener Behandlungseffekte in Subgruppen des Patientenkollektivs kann festgestellt werden, wer von der Intervention nicht, weniger oder mehr profitiert und wie der durchschnittliche Behandlungseffekt in der Gesamtpopulation ausfällt.

Sowohl bei MA-IPDs als auch bei MA-APDs soll Heterogenität zwischen den Primärstudien berücksichtigt werden. Während MA-IPDs einen großen Vorteil bei der Untersuchung klinischer Heterogenität anbieten, unterschätzen MA-APDs, aufgrund des Verlustes von Information durch die Aggregation, die Patientenvariationen bezüglich. der Interventionseffekte [Berlin, 2002]. Gleichwohl lässt sich aber die methodische Heterogenität mit MA-APDs und MA-IPDs prüfen (s. Abschnitt 1.2.3). MA-IPDs werden selten durchgeführt, da sie mit hohen Kosten und Aufwand verbunden sind und eine selten zu erzielende internationale Kooperation der für die Primärstudien Verantwortlichen voraussetzen.

Im Folgenden wurden die zusammengefassten Effektgrößen einer prospektiven MA-IPD [Baigent, 2005] und einer vom Verfasser durchgeführten MA-APD zu Statinen verglichen. Weiterhin wurde der Einfluss der methodischen und klinischen Heterogenität von Primärstudien auf die Ergebnisse von MA-APD untersucht.

4.2. Gegenstand der Fallstudie

4.2.1. MA-IPD zu Statinen

Die MA-IPD zu Statinen ist ein kollaboratives Projekt der leitenden Prüfärzte aller großen RCTs (≥ 1 000 Probanden) mit einer Statinbehandlung von mindestens zwei Jahren. Die Identifizierung von RCTs erfolgt durch eine systematische Suche in mehreren Quellen. Es handelt sich um eine prospektive MA-IPD, d.h. es wurden nur RCTs eingeschlossen, deren Ergebnisse erst nach der Anfertigung des Protokolls der MA-IPD veröffentlicht wurden (Nov. 1994). Dies verringert die Selektion von Primärstudien nach deren Ergebnissen und gewährleistet eine a priori-Definierung von Fragestellungen und Hypothesen. Die MA-IPD zu Statinen schließt Daten von 90 056 Probanden aus 14 RCTs ein. Daten aus einer weiteren einschlussberechtigten RCT [Athyros, 2002] fehlten. Die Synthese ist nach Primärstudie und Follw-Up-Jahr stratifiziert, d.h. die Daten werden nicht behandelt, als ob sie einer Mega-RCT entstammen. Für die Synthese im FEM wurde die Mantel-Haenszel-Methode verwendet. Die Auswertung erfolgte nach dem „Intention-To-Treat-Prinzip“. A priori wurden primäre (Gesamtmortalität, Mortalität aufgrund von koronarer Herzkrankheit „KHK-Mortalität“ und Mortalität nicht aufgrund von koronarer Herzkrankheit in der Gesamtpopulation) und sekundäre (u.a. KHK-Mortalität und koronare Ereignisse in Patientensubgruppen) Endpunkte bestimmt.

4.2.2. MA-APD zu Statinen

Die MA-APD zu Statinen ist eine MA mit aggregierten Patientendaten der in der MA-IPD eingeschlossenen RCTs. Eine in der MA-IPD einbezogene Primärstudie mit 2x2-faktoriellem Design wurde von der MA-APD ausgeschlossen, da sie lediglich zwei Dosierungen von Lovastatin vergleicht [Post-CABG, 1997]. Aggregierte Daten könnten aus der zum Einschluss in der MA-IPD vorgesehenen GREACE-Studie gewonnen werden [Athyros, 2002], allerdings weist diese Studie erhebliche Mängel auf. Anders als die Patienten in der Kontrollgruppe, erhielten die Patienten in der Statingruppe zusätzlich ein strukturiertes stationäres Programm zur Erreichung einer Zielkonzentration von Low-Density-Lipoprotein (LDL). Da die Ergebnisse dieser Studie verzerrt sein könnten, wurde sie nicht in die MA-IPD einbezogen. Daten zu dieser Studie fehlten bei der MA-IPD.

4.3. Zielsetzungen

- Vergleich von MA-IPD und MA-APD bezüglich der zusammengefassten Effektgröße
- Vergleich von MA-APD der Primärstudien mit hoher methodischer Qualität und MA-APD der Primärstudien mit niedriger Qualität bezüglich der zusammengefassten Effektgröße
- Vergleich von MA-APD der Primärstudien mit hohem Frauenanteil und MA-APD der Primärstudien mit niedrigem Frauenanteil bezüglich der zusammengefassten Effektgröße
- Vergleich von MA-APD der Primärstudien mit hoher durchschnittlicher Konzentration von Low-Density-Lipoprotein am Anfang der Studie (Basis-LDL-K) und MA-APD der Primärstudien mit niedriger Basis-LDL-K bezüglich der zusammengefassten Effektgröße
- Untersuchung des potenziellen Einflusses der methodischen Qualität, des Frauenanteils und der Basis-LDL-K auf die zusammengefasste Effektgröße durch Meta-Regression
- Vergleich der Synthese nach dem FEM und nach dem REM

4.4. Methodik

Um die unter Abschnitt 4.3 aufgeführten Vergleiche durchzuführen, wurde die Gesamtmortalität als Zielgröße gewählt. Dies hängt damit zusammen, dass Daten zur Gesamtmortalität in allen Primärstudien verfügbar sind und dass es sich um einen objektiven „harten“ Endpunkt handelt. Eine wesentliche Einschränkung von MA-APD ist die fehlende oder nicht standardisierte Berichterstattung über wichtige Endpunkte und potenzielle Modifikatoren des Behandlungseffekts in der Publikation von Primärstudien, was die Untersuchung von Heterogenitätsquellen mittels MA-APD sehr erschwert.

Zwei statistische Ansätze zur Untersuchung von Heterogenitätsquellen in der MA-APD wurden verwendet, nämlich: Subgruppen-Analyse und Meta-Regression. Aufgrund der niedrigen Power der Subgruppen-Analyse und der Meta-Regression bei MA-APD mit kleiner Anzahl von Primärstudien wurde zur Testung der Subgruppenunterschiede und der geschätzten Regressions-Koeffizienten ein Signifikanzniveau von 10% gewählt. Alle Auswertungen wurden anhand des freien Software-Pakets „R“ (Version 2.2.0) durchgeführt.

Die zu untersuchenden Ursachen möglicher Heterogenität wurden a priori definiert und ihre Auswahl wurde begründet.

4.4.1. Berücksichtigung zufallsbedingter Heterogenität

Die Überprüfung der Zufallsrolle bei der Heterogenität in MA erfolgt durch den Cochran-Test (Q-Test). Zur Quantifzierung der Heterogenität wurde das Heterogenitäts-Maß I^2 verwendet [Higgins, 2002a]. Da die Haupteinschränkung des Heterogenitäts-Tests seine geringe statistische Power ist, d.h. Heterogenität kann noch bestehen, auch wenn der statistische Test nicht signifikant ist, wurde zunächst die von Fleiss und Bailey empfohlene Anhebung des Signifikanzniveaus dieses Tests von 0,05 auf 0,10 übernommen [Fleiss, 1986; Bailey, 1987]. Damit soll das Risiko von falsch negativen Ergebnissen verringert werden. Zur Schätzung der Inter-Studien-Varianz für das REM wurde das am meisten verwendete Verfahren nach DerSimonian und Laird benutzt [DerSimonian, 1996; Sutton, 2008]. Zur Berücksichtigung der zufallsbedingten Heterogenität wurde das REM als Synthese-Modell verwendet. Zwecks der Sensitivitäts-Analyse erfolgte die Synthese auch im FEM nach der Mantel-Haenszel-Methode (s. Abschnitt 1.2.4.2).

4.4.2. Berücksichtigung methodischer Heterogenität

Anhand des Jadad-Scores (0-5 Punkte) wurden die berichtete Randomisierung, Verblindung und Studienaustritte für jede Primärstudie zusammen erfasst (s. Tab. 35). Der nach Jadad genannte Score gilt als einer der am meisten verwendeten Qualitätsbewertungsinstrumente von RCTs und es wurde seine Validität und Reliabilität untersucht [Jadad, 1996b; Moja, 2005]. Eine SR der SRs, die in der Cochrane Library oder in „peer-reviewed“ medizinischen Zeitschriften veröffentlicht wurden, ergab, dass der Jadad-Score bei 11,7% der zwischen 1995 und 2005 veröffentlichten SRs (n = 965) als Instrument der Bewertung methodischer Qualität fungiert [Moja, 2005].

Bewertet wurden weiterhin die Qualitäts-Komponenten: Allocation Concealment (=Verblindung der Randomisierung) und Intention-To-Treat-Analyse (ITT-Analyse), da diese nicht im Jadad-Score enthalten sind. Die empirische Arbeit von Egger und Kollegen fand heraus, dass RCTs mit unangemessenem Allocation Concealment eine durchschnittliche Überschätzung des Behandlungseffekts von 21% aufweisen [Egger, 2003]. Schulz und

Kollegen zeigten sogar eine Überschätzung von 41% [Schulz, 1995]. Die empirischen Studien von Hollis und Campbell sowie von Ruiz-Canela und Kollegen fanden heraus, dass weniger als die Hälfte der untersuchten RCTs nicht über eine ITT-Analyse berichteten [Hollis, 1999; Ruiz-Canela, 2000]. Weiterhin detektierten Kruse und Kollegen, dass bei 100 RCTs mit berichteter ITT-Analyse weniger als die Hälfte tatsächlich dementsprechend auswerteten [Kruse, 2002].

Da die methodische Qualität von einer Person bewertet wurde, wurde sie zweimal in einem zeitlichen Abstand von drei Monaten durchgeführt (Oktober 2005 und Januar 2006).

Tab. 35 Berechnung des Jadad-Scores [Quelle: http://www.naturalstandards.com; nach: Jadad, 1996b]

Item	**Score**
Was the study described as randomized (this includes words such as randomly, random, and randomization)?	0/1
Was the method used to generate the sequence of randomization described and appropriate (table of random numbers, computer-generated, etc)?	0/1
Was the study described as double blind?	0/1
Was the method of double blinding described and appropriate (identical placebo, active placebo, dummy, etc)?	0/1
Was there a description of withdrawals and dropouts (number and reasons)?	0/1
Deduct one point if the method used to generate the sequence of randomization was described and it was inappropriate (patients were allocated alternately, or according to date of birth, hospital number, etc).	0/-1
Deduct one point if the study was described as double blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy).	0/-1

4.4.3. Untersuchung klinischer Heterogenität

Ein positiver Zusammenhang zwischen den Serum-Werten von Low-Density-Lipoprotein (LDL) und dem Risiko für Koronare Herzkrankheit (KHK) wurde in mehreren beobachteten Studien gefunden [MRFIT, 1982; Anderson, 1987; Assmann, 1987]. Mehrere RCTs haben gezeigt, dass HMG-CoA-Reduktase-Hemmer (Statine) die Wiederholung eines kardiovaskulären Ereignisses bei Patienten mit manifesten KHK und den erstmaligen Auftritt eines solchen Ereignisses bei Probanden mit erhöhtem Risiko für KHK verringern [LaRosa, 1999; Pignone, 2000]. Statine wirken hauptsächlich durch die Senkung der LDL [Law, 2003]. Daher ist zu erwarten, dass Patienten mit hoher Basis-LDL-K mehr als die mit niedriger Basis-LDL-K von der Statintherapie profitieren. Bei einer RCT wurden die Teilnehmer vor der Randomisierung bezüglich der Basis-LDL-K stratifiziert, aber in der Publikation erschienen lediglich die Ergebnisse für das Stratum niedriger Basis-LDL-K (> 4 mmol/L) [Pitt, 1995].

Während eine nach der Randomisierung erfolgten Stratifizierung der Teilnehmer bezüglich der Basis-LDL-K in der CARDS-Studie keine differentielle Wirksamkeit heraus fand [Colhoun, 2004], zeigte eine ähnliche Auswertung der WOSCOPS-Studie einen nicht signifikanten höheren Nutzen für Probanden mit niedriger Basis-LDL-K (< 4,9 mmol/L) [Shepherd, 1995]. Ob die Wirksamkeit von Statinen bei unterschiedlicher Basis-LDL-K sich unterscheidet, wird im Rahmen der vorliegenden MA-APD untersucht.

Im Allgemeinen sind Frauen unterrepräsentiert in RCTs zu Statinen. Eine SR der RCTs zu Statinen, die zwischen 1990 und 2002 veröffentlicht wurden, fand heraus, dass 8 von den 50 identifizierten Studien Frauen ausschlossen. Der Median des Frauenanteils in diesen Studien betrug lediglich 19% (Interquartil-Bereich 12-30%) [Bartlett, 2005]. Zwei SRs zu Cholesterin senkenden Interventionen fanden keine signifikante Reduzierung der Gesamtmortalität bei Frauen [Grady, 2003; Walsh, 2004]. Anhand der vorliegenden MA-APD wird der Einfluss des Frauenanteils in den Primärstudien auf die Gesamtmortalität untersucht.

4.4.4. Extraktion der Daten

Aus den Publikationen der Primärstudien und aus ihren Studienplänen, falls letztere vorhanden sind, wurden Angaben zu folgenden Variablen aus jeder Studie extrahiert:

- Publikationsjahr
- Anzahl der Studienteilnehmer
- Dauer des Follow-Ups
- Probandenkollektive in Bezug auf KHK-Vorgeschichte
- Typ der Kontrollgruppe
- Statintyp
- Gesamtmortalität in der Statin- und in der Kontrollgruppe
- Frauenanteil (in Prozent)
- Basis-LDL-K (in mmol/L)
- Angaben zu Randomisierung, Allocation Concealment, Verblindung, Studienaustritten, Intention-To-Treat-Analyse (ITT-Analyse).

4.4.5. Subgruppen-Analysen

Zur Untersuchung des Einflusses methodischer Qualität auf die zusammengefasste Effektgröße wurden die Primärstudien nach dem Median ihres erzielten Jadad-Scores (4 Punkte), nach der Angemessenheit ihres Allocation Concealment und nach der Befolgung des ITT-Prinzips jeweils in zwei Gruppen geteilt. Demnach wurden folgenden Subgruppen definiert:

RCTs mit hoher Qualität (Jadad-Score ≥ 4)
RCTs mit niedriger Qualität (Jadad-Score < 4)

RCTs mit hoher Qualität (angemessens Allocation Concealment)
RCTs mit niedriger Qualität (unangemessenes Allocation Concealment)

RCTs mit hoher Qualität (ITT-Analyse)
RCTs mit niedriger Qualität (Per-Protokoll-Analyse)

Für die RCTs mit hoher Qualität (RCTs-HQ) und die RCTs mit niedriger Qualität (RCTs-NQ) wurde jeweils eine MA-APD durchgeführt. Der Interaktions-Test wurde verwendet, um die Unterschiede zwischen der MA-APD für RCTs-HQ und der MA-APD für RCTs-NQ bezüglich der zusammengefassten Gesamtmortalität zu testen [Altman, 2003].

Zwecks der Ermittlung des Einflusses von Patientencharakteristika auf die zusammengefasste Effektgröße wurden die Primärstudien nach dem Median ihrer Basis-LDL-K (3.8 mmol/L) und ihres Frauenanteils (18,61%) jeweils zwei Gruppen zugewiesen. Demnach wurden folgende Subgruppen bestimmt:

RCTs mit hohem Frauenanteil (RCTs-HF; Frauenanteil $\geq$ 18,61%)
RCTs mit niedrigem Frauenanteil (RCTs-NF; Frauenanteil $<$ 18,61%)

RCTs mit hoher Basis-LDL-K (RCTs-HL; Basis-LDL-K $\geq$ 3,8 mmol/L)
RCTs mit niedriger Basis-LDL-K (RCTs-NL; Basis-LDL-K $<$ 3,8 mmol/L)

Eine MA-APD wurde für jede Subgruppe durchgeführt und die Variation der zusammengefassten Effektgröße zwischen RCTs-HF und RCTs-NF bzw. RCTs-HL und RCTs-NL mit dem Interaktions-Test geprüft.

4.4.6. Meta-Regressionen

Zwecks der gleichzeitigen Berücksichtigung methodischer und klinischer Heterogenitätsquellen wurde der potenzielle Einfluss von Studienqualität, Frauenanteil und Basis-LDL-K auf die Gesamtmortalität in den Primärstudien anhand von drei Meta-Regressions-Modellen untersucht:

Modell 1: $\text{Log}(RR_i) = \beta_0 + \beta_1 * \text{Jadad} + \beta_2 * \text{Frauenanteil}_i + \beta_3 * \text{Basis-LDL-K}_i$

Modell 2: $\text{Log}(RR_i) = \beta_0 + \beta_1 * \text{Allocation Concealment} + \beta_2 * \text{Frauenanteil}_i + \beta_3 * \text{Basis-LDL-K}_i$

Modell 3: $\text{Log}(RR_i) = \beta_0 + \beta_1 * \text{ITT-Analyse} + \beta_2 * \text{Frauenanteil}_i + \beta_3 * \text{Basis-LDL-K}_i$

Wobei RR_i das Relative Risiko zur Gesamtmortalität in Studie i (i = 1, 2, ..., k; k = 14), β_1 den Regressionskoeffizienten zur Studienqualität, β_2 den Regressionskoeffizienten zum Frauenanteil (von 0,00 bis 1,00) und β_3 den Regressionskoeffizienten zur Basis-LDL-K darstellen. Das verwendete generalisierte lineare Modell weist eine Normalverteilung mit einer „Log-Link-Funktion“ auf. Die Schätzung der Regressionskoeffizienten erfolgt durch die Methode der gewichteten kleinsten Quadrate. Zur Berücksichtigung der zufallsbedingter Heterogenität wurden alle Regressionen nach dem REM gewichtet. Zwecks der Sensitivitäts-Analyse erfolgte die Gewichtung auch nach dem FEM.

4.5. Ergebnisse

4.5.1. Allgemeine Merkmale der RCTs

Alle Primärstudien wurden zwischen November 1994 und Juni 2002 veröffentlicht. Die durchschnittliche (Median) Dauer des Follow-Ups beträgt 4,8 Jahre (Interquartil-Bereich 3,2 – 5,0). Der Median der Teilnehmerzahl pro RCT liegt bei 5804 (Interquartil-Bereich 4159 - 9014). Sechs Studien schließen Teilnehmer mit und ohne vorherige KHK-Ereignissen ein,

weitere drei RCTs untersuchen die Wirksamkeit von Statinen bei Patienten mit vorherigen Ereignissen (Sekundärprävention) und lediglich zwei Studien beziehen Probanden ohne relevante Krankheitsgeschichte (Primärprävention) ein. Mit Ausnahme von drei Studien [ALLHAT-LLT, 2002; Athyros, 2002; GISSI-P, 2000] werden Statine in allen weiteren RCTs mit Placebo verglichen. Patienten in der Kontrollgruppe in ALLHAT-LLT und GREACE erhielten Standardbehandlung durch den Hausarzt und in GISSIP-P „keine Statine“ [ALLHAT-LLT, 2002; Athyros, 2002; GISSI-P, 2000]. Teilnehmer in der Interventionsgruppe bekamen Pravastatin (6 RCTs), Atorvastatin, Simavastatin, Fluvastatin (jeweils 2 RCTs) oder Lovastatin (1 RCT). Weitere Details werden in der Tabelle 36 deutlich.

Tab. 36 Allgemeine Merkmale der RCTs zu Statinen

Studienbezeichnung	Literaturverweis	Teilnehmerzahl	Dauer des Follow-Up (in Jahren)	Prävention*	Statin
CARDS	[Colhoun, 2004]	2838	3.9	3	Atorvastatin
4S	[4S, 1994]	4444	5.2	2	Simavastatin
ALERT	[Holdaas, 2003]	2102	5.1	3	Fluvastatin
PROSPER	[Shepherd, 2002]	5804	3.2	3	Pravastatin
LIPID	[LIPID, 1998]	9014	5.6	2	Pravastatin
AFCAPS/TexCAPS	[Downs, 1998]	6605	5.3	1	Lovastatin
CARE	[Sacks, 1996]	4159	4.8	2	Pravastatin
WOSCOPS	[Shepherd, 1995]	6595	4.8	1	Pravastatin
ALLHAT-LLT	[ALLHAT-LLT, 2002]	10355	4.8	3	Pravastatin
HPS	[HPS, 2002]	20536	5.0	3	Simavastatin
ASCOT-LLA	[Sever, 2003]	10305	3.2	3	Atorvastatin
LIPS	[Serruys, 2002]	1677	3.1	2	Fluvastatin
GISSI-P	[GISSI-P, 2000]	4271	1.9	2	Pravastatin

* Prävention : 1 = Primärprävention, 2 = Sekundärprävention, 3 = Primär- und Sekundärprävention

4.5.2. Methodische und klinische Heterogenität der RCTs

Die Primärstudien (n= 13) erzielen einen Jadad-Score mit dem Medianwert von 4 Punkten (Interquartil-Bereich 3 – 5). Fast die Hälfte der Studien (n= 6) erhielten alle Punkte des Jadad-Scores (s. Tab. 37). Das Allocation Concealment wurde bei 4 Studien als nicht angemessen beurteilt. Da die Auswertung aller Primärstudien nach dem Intention-To-Treat-Prinzip erfolgte, ist diese Qualitäts-Komponente als potenzielle Heterogenitätsquelle auszuschließen. Daher wurde diese Dimension bei den Subgruppen-Analysen und Meta-Regressionen nicht berücksichtigt, d.h. das Regressions-Modell 3 entfällt. Der Median der

Basis-LDL-K liegt bei 3,8 mmol/L (Interquartil-Bereich 3,4 – 3,9), der des Frauenanteils bei 18,61% (Interquartil-Bereich 15,09 – 32,03) (s. Tab. 38).

Tab. 37 Jadad-Scores der RCTs zu Statinen

Studienbezeichnung	Randomisierung	Doppel-Verblindung	Studienaustritte	Jadad-Score
CARDS	2	2	1	5
4S	2	2	1	5
ALERT	2	2	1	5
PROSPER	2	2	1	5
LIPID	2	2	1	5
AFCAPS/TexCAPS	2	2	1	5
CARE	2	2	0	4
WOSCOPS	2	1	0	3
ALLHAT-LLT	2	0	1	3
HPS	1	1	1	3
ASCOT-LLA	1	2	0	3
LIPS	1	1	1	3
GISSI-P	2	0	0	2

Tab. 38 Methodische und klinische Heterogenität der RCTs zu Statinen

Studienbezeichnung	Jadad-Score	Allocation Concealment	ITT-Analyse	Basis-LDL-K (mmol/L)	Frauenanteil (%)
CARDS	5	angemessen	Ja	3	32,03
4S	5	angemessen	Ja	4.9	18,61
ALERT	5	angemessen	Ja	4.1	34,02
PROSPER	5	angemessen	Ja	3.8	51,69
LIPID	5	unangemessen	Ja	3.9	16,82
AFCAPS/TexCAPS	5	unangemessen	Ja	3.9	15,09
CARE	4	angemessen	Ja	3.6	13,85
WOSCOPS	3	unangemessen	Ja	5	0,00
ALLHAT-LLT	3	angemessen	Ja	3.3	49,00
HPS	3	angemessen	Ja	3.4	24,75
ASCOT-LLA	3	angemessen	Ja	3.4	18,85
LIPS	3	angemessen	Ja	3.4	15,16
GISSI-P	2	unangemessen	Ja	3.9	13,74

4.5.3. Vergleich der synthetischen Funktion von MA-IPD und MA-APD

Die MA-IPD berichtet kein Testergebnis zur Heterogenität zwischen den Primärstudien bezüglich der Gesamtmortalität. Der Cochran-Test (Q-Test) für Heterogenität der Relativen Risiken (RRs) zur Gesamtmortalität in MA-APD liefert ein signifikantes Ergebnis (p-Wert = 0,03). Das Heterogenitäts-Maß (I^2) in der MA-APD, das höher ausfällt, als durch den Zufall

zu erwarten wäre, liegt bei 47,5%. Die MA-IPD und die MA-APD erzielen hoch signifikante Ergebnisse bezüglich der Reduzierung der Gesamtmortalität durch Statine. Berechnet nach dem FEM zeigt die MA-IPD, dass die Statinbehandlung die Gesamtmortalität um 13% reduziert (RR = 0,87; 95%-KI = 0,84 – 0,91). Extrem ähnliche Ergebnisse bezüglich der zusammengefassten Effektgröße und ihres 95%-Konfidenz-Intervalls liefert die MA-APD, wobei das REM ein breiteres Konfidenz-Intervall (RR= 0,87; 95%-KI = 0,82 – 0,93) als das FEM (RR = 0,88; 95%-KI = 0,84 – 0,92) aufweist (s. Tabelle 37).

4.5.4. Ergebnisse der Subgruppen-Analyse

4.5.4.1. Jadad-Score

Berechnet nach dem REM betrug das zusammengefasste RR zur Gesamtmortalität bei RCTs-HQ 0,87 (95%-KI =0,78 – 0,98) und bei RCTs-NQ 0,89 (95%-KI = 0,83 – 0,96). Das Heterogenitäts-Maß (I^2) entsprach 61,1% bzw. 19,4%. RCTs-NQ zeigten einen großeren Effekt sowohl beim FEM als auch beim REM. Das REM wies ein breiteres Konfidenz-Intervall auf als das FEM (s. Tab. 39). Allerdings lieferte der Interaktions-Test sowohl im FEM (z= 1,38; p= 0,17) als auch im REM (z= 0,32; p= 0,75) kein signifikantes Ergebnis.

Tab. 39 Einfluss des Jadad-Scores auf das zusammengefasste RR zur Gesamtmortalität

	Fixed-Effects-Modell		Random-Effects-Modell	
Studienkollektiv (n)	Relatives Risiko	95%-KI	Relatives Risiko	95%-KI
Alle Studien (13)	0,88	0,84 – 0,92	0,87	0,82 – 0,93
RCTs mit hoher Qualität (7)	0,85	0,80 – 0,91	0,87	0,78 – 0,98
RCTs mit niedriger Qualität (6)	0,90	0,85 – 0,94	0,89	0,83 – 0,96

RCTs-HQ: p-Wert für Q-Test = 0,02; I^2 = 61,1%
RCTs-NQ: p-Wert für Q-Test = 0,29; I^2 = 19,4%

4.5.4.2. Allocation Concealment

Berechnet nach dem REM betrug das zusammengefasste RR zur Gesamtmortalität bei RCTs mit hoher Qualität (RCTs-HQ) 0,89 (95%-KI =0,83 – 0,96) und bei RCTs mit niedriger Qualität (RCTs-NQ) 0,81 (95%-KI = 0,74 – 0,88). Das Heterogenitäts-Maß (I^2) entsprach 48,7% bzw. 0%. RCTs-NQ zeigten einen größeren Effekt sowohl beim FEM als auch beim REM. Das REM wies ein breiteres Konfidenz-Intervall auf als das FEM (s. Tab. 40). Während

der Interaktions-Test im FEM einen signifikanten Unterschied zwischen RCTs-NQ und RCTs-HQ zeigte (z= 1.99; p= 0,047), fiel der Unterschied im REM nicht signifikant aus (z= 1,64; p=0,10).

Tab. 40 Einfluss des Allocation Concealment auf das zusammengefasste RR zur Gesamtmortalität

	Fixed-Effects-Modell		Random-Effects-Modell	
Studienkollektiv (n)	Relatives Risiko	95%-KI	Relatives Risiko	95%-KI
Alle Studien (13)	0,88	0,84 – 0,92	0,87	0,82 – 0,93
RCTs mit hoher Qualität (9)	0,90	0,86 – 0,94	0,89	0,83 – 0,96
RCTs mit niedriger Qualität (4)	0,81	0,74 – 0,89	0,81	0,74 – 0,88

RCTs-HQ: p-Wert für Q-Test = 0,05; I^2 = 48,7%
RCTs-NQ: p-Wert für Q-Test = 0,41; I^2 = 0%

4.5.4.3. Frauenanteil

Berechnet nach dem REM betrug das zusammengefasste RR zur Gesamtmortalität bei RCTs mit hohem Frauenanteil (RCTs-HF) 0,90 (95%-KI = 0,82 – 0,98) und bei RCTs mit niedrigem Frauenanteil (RCTs-NF) 0,82 (95%-KI = 0,76 – 0,89). Das Heterogenitäts-Maß (I^2) entsprach 58,7% bzw. 0%. Die Reduzierung der Gesamtmortalität durch Statine war in RCTs-HF geringer als in RCTs-NF. Sie war sowohl beim FEM als auch beim REM zu beobachten. Das REM wies lediglich bei RCTs-HF ein breiteres Konfidenz-Intervall auf als das FEM (s. Tab. 41). Während der Interaktions-Test im FEM einen signifikanten Unterschied zwischen RCTs-NF und RCTs-HF zeigte (z= 1.97; p= 0,049), fiel der Unterschied im REM nicht signifikant aus (z= 1,54; p=0,12).

Tab. 41 Einfluss des Frauenanteils auf das zusammengefasste RR zur Gesamtmortalität

	Fixed-Effects-Modell		Random-Effects-Modell	
Studienkollektiv (n)	Relatives Risiko	95%-KI	Relatives Risiko	95%-KI
Alle Studien (13)	0,88	0,84 – 0,92	0,87	0,82 – 0,93
Studien mit hohem Frauenanteils (7)	0,90	0,86 – 0,94	0,90	0,82 – 0,98
Studien mit niedrigem Frauenanteils (6)	0,82	0,76 – 0,89	0,82	0,76 – 0,89

RCTs-HF: p-Wert für Q-Test = 0,02; I^2 = 58,7%
RCTs-NF: p-Wert für Q-Test = 0,47; I^2 = 0%

4.5.4.4. Basis-LDL-K

Berechnet nach dem REM betrug das zusammengefasste RR zur Gesamtmortalität bei RCTs mit hoher Basis-LDL-K (RCTs-HL) 0,86 (95%-KI = 0,77 – 0,97) und bei RCTs mit niedriger Basis-LDL-K (RCTs-NL) 0,90 (95%-KI = 0,84 – 0,96). Das Heterogenitäts-Maß (I^2) entspricht 58,7% bzw. 19,2%. Die Reduzierung der Gesamtmortalität durch Statine war in RCTs-HL höher als in RCTs-NL. Sowohl beim FEM als auch beim REM war dies zu beobachten. Das REM wies ein breiteres Konfidenz-Intervall auf als das FEM (s. Tab. 42). Allerdings lieferte der Interaktions-Test sowohl im FEM (z= 1,42; p= 0,16) als auch im REM (z= 0,65; p= 0,52) kein signifikantes Ergebnis.

Tab. 42 Einfluss der Basis-LDL-K auf das zusammengefasste RR zur Gesamtmortalität

	Fixed-Effects-Modell		Random-Effects-Modell	
Studienkollektiv (n)	Relatives Risiko	95%-KI	Relatives Risiko	95%-KI
Alle Studien (13)	0,88	0,84 – 0,92	0,87	0,82 – 0,93
Studien mit hoher Basis-LDL-K (7)	0,85	0,79 – 0,90	0,86	0,77 – 0,97
Studien mit niedriger Basis-LDL-K (6)	0,90	0,86 – 0,95	0,90	0,84 – 0,96

RCTs-HL: p-Wert für Q-Test = 0,02; I^2 = 58,7%
RCTs-NL: p-Wert für Q-Test = 0,29; I^2 = 19,2%

4.5.5. Ergebnisse der Meta-Regression

4.5.5.1. Ergebnisse von Modell 1

Gewichtet nach dem FEM, wies der Frauenanteil der Primärstudien einen signifikanten Einfluss auf das RR zur Gesamtmortalität auf. Wenn der Jadad-Score und die Basis-LDL-K konstant blieben, führte eine Steigerung des Frauenanteils um 10% zur Erhöhung der Gesamtmortalität um 5%. Die Gewichtung nach dem REM führte zu einem breiteren 95%-Konfidenz-Intervall und zu einem größeren p-Wert (0,07) für diese Einflussgröße (s. Tab. 43 und 44).

Tab. 43 Ergebnisse der Meta-Regression (Modell 1) nach dem FEM

Kovariable	Punktschätzer	95%-KI	p-Wert
Jadad-Score	0,99	0,93 – 1,04	0,651
Frauenanteil	1,62	1,13 – 2,30	0,027
Basis-LDL-K	0,97	0,85 – 1,11	0,704

Tab. 44 Ergebnisse der Meta-Regression (Modell 1) nach dem REM

Kovariable	Punktschätzer	95%-KI	p-Wert
Jadad-Score	0,99	0,93 – 1,06	0,855
Frauenanteil	1,58	1,01 – 2,46	0,073
Basis-LDL-K	0,98	0,85 – 1,13	0,780

4.5.5.2. Ergebnisse von Modell 2

Gewichtet nach dem FEM, wies der Frauenanteil der Primärstudien einen signifikanten Einfluss auf das RR zur Gesamtmortalität auf. Wenn das Allocation Concealment und die Basis-LDL-K konstant blieben, führte eine Steigerung des Frauenanteils um 10% zur Erhöhung der Gesamtmortalität um 5%. Die Gewichtung nach dem REM führte zu einem breiteren 95%-Konfidenz-Intervall und zu einem größeren p-Wert (0,08) für diese Einflussgröße (s. Tab. 45 und 46).

Tab. 45 Ergebnisse der Meta-Regression (Modell 2) nach dem FEM

Kovariable	Punktschätzer	95%-KI	p-Wert
Allocation Concealment	1,00	0,94 – 1,08	0,894
Frauenanteil	1,57	1,06 – 2,29	0,045
Basis-LDL-K	0,96	0,85 – 1,07	0,484

Tab. 46 Ergebnisse der Meta-Regression (Modell 2) nach dem REM

Kovariable	Punktschätzer	95%-KI	p-Wert
Allocation Concealment	0,99	0,91 – 1,08	0,850
Frauenanteil	1,60	0,99 – 2,58	0,086
Basis-LDL-K	0,97	0,85 – 1,10	0,663

4.6. Zusammenfassung und Diskussion

Die MA-IPD und die MA-APD lieferten extrem ähnliche Ergebnisse zum Einfluss der Statine auf die Gesamtmortalität. Dies geschah, obwohl eine in die MA-IPD einbezogene Primärstudie von der MA-APD ausgeschlossen wurde. Es bestehen wenige Vergleiche der synthetischen Funktion von MA-IPD und MA-APD in der Literatur. Eine mathematische Ableitung wies darauf hin, dass unter Annahme der Homogenität von RCTs, die zusammengefassten Effektgrößen von MA-IPDs und MA-APDs sich nicht unterscheiden

[Olkin, 1998]. Mathew und Nordstrom leiten mathematisch auch keinen Unterschied her, wenn Heterogenität zwischen primären Studien besteht [Mathew, 1999].

Obwohl der Cochran-Test für Heterogenität zwischen den Primärstudien bezüglich des RR zur Gesamtmortalität signifikant ist, wurde dieses Ergebnis in der MA-IPD weder erwähnt noch bei der Synthese berücksichtigt, d.h. die Synthese erfolgt nicht, wie erwartet, nach dem FEM.

Mögliche Heterogenitätsquellen wurden vor der Auswertung bestimmt, ihre Auswahl wurde durch externe Evidenz begründet. Für die Untersuchung methodischer und klinischer Heterogenität wurden jeweils zwei Subgruppen-Analysen durchgeführt. Es wurde nicht auf die Signifikanz der Ergebnisse innerhalb jeder Subgruppe abgehoben, sondern die Unterschiede zwischen den Gruppen wurden mit dem Interaktions-Test geprüft. Allerdings wurde das Signifikanzniveau für die multiplen Interaktions-Tests nicht adjustiert, was zu falsch positiven Ergebnissen führen kann [Schulz, 2005].

Die Bewertung methodischer Qualität primärer Studien in MAs gilt als unausweichlich, da „*many bad studies dont't make a good one*“ [Parmigiani, 2002, S. 125]. Anders als in der vorliegenden MA-APD wurden in der MA-IPD keine methodischen Aspekte der RCTs berücksichtigt. Die methodische Qualität der Primärstudien in dieser MA kann als überdurchschnittlich bewertet werden. 6 Studien erhielten alle Punkte des Jadad-Scores und lediglich eine erzielte 2 Punkte. Das Allocation Concealment wurde bei 9 Studien als angemessen bewertet. Alle Studien wurden nach dem ITT-Prinzip ausgewertet. Eine in einem 3-monatigen Abstand wiederholte Beurteilung der methodischen Qualität dieser RCTs führt zu demselben Ergebnis. Der Einfluss der Verblindung von Primärstudien wurde nicht getrennt untersucht, da die Gesamtmortalität einen objektiven Endpunkt darstellt, der durch Verblindung nicht beeinflusst werden kann [Schulz, 2002c].

Unter Berücksichtigung der zufallsbedingten Heterogenität mittels REM fanden sowohl die Subgruppen-Analysen als auch die Meta-Regressionen keinen signifikanten Einfluss der methodischen Qualität primärer Studien auf die Gesamtmortalität. Dies steht im Einklang mit mehreren meta-epidemiologischen Studien, die keinen Zusammenhang zwischen der Qualität von RCTs und dem Interventionseffekt fanden [Verhagen, 2002; Balk, 2002; Emerson, 1990]. Allerdings ist zu beachten, dass aufgrund der niedrigen Anzahl und der überdurchschnittlichen Qualität der einbezogenen RCTs, ein Einfluss der methodischen Qualität unentdeckt geblieben sein kann [Brookes, 2001]. Zudem zeigen andere empirische Studien eine Überschätzung des Interventionseffekts durch RCTs mit niedriger Qualität

[Moher, 1998; Schulz, 1995] und weitere Studien beobachten eine Unterschätzung des Effekts durch RCTs mit niedriger Qualität [Siersma, 2006 Verhagen, 1998].

Die Variation zwischen den Effektgrößen der RCTs konnte nicht durch die unterschiedliche Basis-LDL-K erklärt werden. Beobachtende Studien haben keine Untergrenze der LDL gefunden, wo niedrigere LDL-Werte nicht mit verringertem KHK-Risiko verbunden sind [Chen, 1991; Stamler, 1993]. Die MA-APD von Edwards und Moore zeigte keinen Einfluss des Basis-Gesamt-Cholesterin-Wertes auf das Ausmaß der Senkung von Gesamt-Cholesterin durch Statine [Edwards, 2003].

Frauen sind untervertreten in den RCTs zu Statinen [Bartlett, 2005]. Keine der in der MA-APD eingeschlossenen Studien berichteten von einer nach Geschlecht stratifizierten Randomisierung. Ob der Minimisierungs-Algorithmus, der bei der HPS und der ASCOT-LLA für eine dynamische Randomisierung verwendet wurde, die Variable Geschlecht beinhaltet, ist den Publikationen nicht zu entnehmen. Obwohl die Simulationsstudien von Lambert und Mitarbeitern eine mit großem Abstand niedrigere statistische Power der Meta-Regression im Vergleich zu MA-IPD fand [Lambert, 2002], zeigt die vorliegende Meta-Regression eine signifikante differenzielle Wirksamkeit der Statine hinsichtlich des Frauenanteils. Unter Berücksichtigung der methodischen Heterogenität und der Basis-LDL-K fand die Meta-Regression einen signifikant negativen Zusammenhang zwischen dem Frauenanteil einer RCT und dem Ausmaß der Senkung von Gesamtmortalität durch Statine. Aufgrund der sogenannten „Ecological-Fallacy“ ist dies Ergebnis nicht auf einzelne Frauen zu übertragen [Berlin, 2002]. Die MA-IPD zu Statinen berichtet nicht über den Effekt von Statinen auf die Gesamtmortalität bei Frauen, obwohl die Meta-Analysten von weiteren Autoren danach gefragt wurden [Abraha, 2006]. Weiterhin zeigt die MA-IPD eine nicht signifikant niedrige Wirksamkeit der Statine bei Frauen bezüglich des kombinierten Endpunkts: KHK-Mortalität und nicht-tödlicher Myokardinfarkt. Die kleine Anzahl der Primärstudien ermöglicht keine weiteren Adjustierungen nach wichtigen Confoundern wie Alter und Konzentration von High-Density-Lipoprotein (HDL). Hinweise aus der Literatur deuten an, dass bei älteren Frauen eine niedrige HDL und hohe Triglyceride stärkere Risikomarker für KHK-Mortalität als hohe LDL darstellen [Manolio, 1992].

Die vorliegende Fallstudie deutet auf eine geringfügig verminderte Wirksamkeit der Statine bei der Reduzierung der Gesamtmortalität bei Frauen hin. Es ist empfehlenswert, dieses Ergebnis in zukünftigen RCTs zu Statinen zu prüfen.

Zusammenfassung

Die Berücksichtigung von Heterogenität in MAs schließt die Bewertung der methodischen Variabilität, der klinischen Diversität und der zufälligen Varianz zwischen den Primärstudien ein. Das Ausmaß an Heterogenität in einer MA misst sich an methodischen und klinischen Variationen zwischen den Primärstudien, die über den Zufall hinaus gehen. Bei einem hohen Ausmaß an Heterogenität soll keine MA durchgeführt werden, sondern es gilt zunächst mögliche Heterogenitätsursachen zu untersuchen. Bei homogenen Primärstudien soll das FEM als Synthese-Modell benutzt und mit REM zwecks der Robustheitsprüfung verglichen werden. Bei der Mehrheit der MAs soll allerdings von Variabilität, die über den Zufall hinaus geht, ausgegangen und im REM quantitativ zusammengefasst werden. Aufgrund niedriger statistischer Power der bisherigen Heterogenitäts-Tests und -Maße bei den meisten MAs sollen nicht signifikante Testergebnisse oder niedrige Heterogenitäts-Maße nicht als Entscheidungsgrundlage für die Verwendung vom FEM als Synthese-Modell fungieren [Ioannidis, 2007]. Auf der anderen Seite weist ein signifikanter Heterogenitäts-Test, z.B. der Cochran-Test, oder ein hohes und präzises Heterogenitäts-Maß, z.B. das I^2, auf hohe Heterogenität hin, die nicht nur durch den Zufall bedingt ist. In dem Fall soll die MA im REM durchgeführt werden und mögliche methodische und klinische Ursachen der Variabilität sollen untersucht werden.

In der vorliegenden Arbeit wurden zwei SRs und eine Fallstudie konzipiert und durchgeführt. In der ersten SR wurden 39 empirische Studien zum Einfluss von Bias in RCTs auf die Ergebnisse von MAs identifiziert, kritisch bewertet und in einem REM kombiniert. Die Suchen erfolgten in dem „Cochrane Methodology Register“, in Medline und in den Bibliographien der eingeschlossenen Studien. 134 empirische Vergleiche zwischen RCTs mit hoher und mit niedriger methodischer Qualität konnten extrahiert und im REM zusammengefasst werden. Betrachtet man die Vergleiche einzeln, findet man kein konsistentes Ergebnis bezüglich des Zusammenhangs zwischen der methodischen Qualität von RCTs und deren Effektgröße außer bei der Qualitäts-Komponente: die Berücksichtigung von Studienaustritten, bei der durchgehend kein Zusammenhang gefunden wurde. Es ist allerdings zu vermerken, dass diese Qualitäts-Komponente in den Studien unterschiedlich definiert und operationalisiert ist. Im Durchschnitt überbewerteten RCTs mit niedrigen Qualitäts-Scores, mit unangemessener Randomisierungsmethode, mit unangemessenem Allocation Concealment und ohne jegliche Art der Verblindung die Behandlungswirksamkeit. Da die meisten Vergleiche nicht über die aus den RCTs extrahierten Endpunkte berichteten, konnte keine Differenzierung des Zusammenhangs zwischen Verblindung und Effektgröße nach der Objektivität des Endpunktes vorgenommen werden. Während das Fehlen der Doppel-Verblindung zur

Überschätzung der Interventionswirksamkeit führte, zeigten die Effektgrößen von RCTs mit und ohne Verblindung der Endpunktmessung überraschenderweise keine Unterschiede. Die meisten empirischen Studien zielten nicht darauf ab, die klinischen Heterogenitätsursachen zwischen den RCTs zu untersuchen. Lediglich ein Drittel der Studien untersuchte interventionsbezogene Variationen zwischen den RCTs und nur ein Viertel der Studien suchte nach patientenbezogener Heterogenität. Simultane Kontrolle von methodischen und klinischen Heterogenitätsquellen durch multivariate Regression wurde nur in drei Studien verfolgt. Nach dem Kenntnisstand des Verfassers ist diese SR die erste zum Einfluss von Bias in RCTs auf die Ergebnisse von MAs. Die Ergebnisse dieser SR zeigen im Durchschnitt eine mäßige Überschätzung des Behandlungseffekts durch RCTs mit niedriger methodischer Qualität. Allerdings berücksichtigt die große Mehrheit der in der SR eingeschlossenen Studien nicht die klinische Heterogenität zwischen den RCTs. Diese Tatsache kann zur Verzerrung des gefundenen Zusammenhangs zwischen der Studienqualität und der Effektgröße führen.

Im Rahmen der zweiten SR wurden 70 Vergleiche zwischen den Effektgrößen von MA-IPDs und MA-APDs aus 25 empirischen Studien gefunden und bewertet. Die Suchen erfolgten in dem „Cochrane Methodology Register“, in Medline, im „Cochrane Database of IPD Reviews“, in den Publikationen von 52 kollaborativen Gruppen für MA-IPDs sowie in den Literatur-verzeichnissen der eingeschlossenen Studien. Bei zwei Dritteln der Vergleiche wurde eine Tendenz zur Überschätzung der Effektgröße und zur Reduzierung der Präzision bei MA-APDs im Vergleich zu MA-IPDs beobachtet. Diese Tendenz war oft auf den Einfluss des Patient-Exclusion-Bias und der Verwendung unterschiedlicher Effektmaße, OR bei MA-APDs und Hazard Ratio bei MA-IPDs, zurückzuführen. Allerdings waren die relativen Unterschiede zwischen den Punktschätzern von MA-IPDs und MA-APDs in allen Vergleichen, mit Ausnahme von einem [McCormack, 2004], kleiner als 50%. Bei diesem Vergleich war die Diskrepanz zwischen den Datengrundlagen für MA-IPD (RCT= 20) und für MA-APD (RCTs= 3) groß. Weiterhin ergab der Paired-t-Test keinen signifikanten Unterschied zwischen den beiden Arten von MAs. Dies kann mit der geringen Zahl der Vergleiche oder den geringen Unterschieden zwischen den Punktschätzern von MA-IPDs und MA-APDs zusammenhängen. Die Hälfte der Studien stellte keine Ergebnisse des Heterogenitäts-Tests dar. Lediglich bei einem Viertel der Studien wurde die methodische Qualität in beiden Arten von MAs berücksichtigt. Die klinische Heterogenität wurde nur bei einem Drittel der Studien anhand beider Arten von MAs untersucht. Dabei sind diesbezüglich keine konsistenten Unterschiede zwischen MA-IPDs und MA-APDs zu verzeichnen. Nach dem Kenntnisstand des Verfassers ist diese SR die erste zum Vergleich von MA-IPD und MA-APD. Die Ergebnisse dieser SR deuten darauf hin, dass MA-IPDs und MA-APDs sich bezüglich der

Evidenz-Synthese keine bedeutsamen Unterschiede aufweisen. Anhand der vorhandenen Evidenz kann jedoch keine Aussage über die Unterschiede zwischen MA-IPDs und MA-APDs bezüglich der Heterogenitäts-Analyse gemacht werden.

Im Rahmen einer Fallstudie wurden die Ergebnisse einer publizierten MA-IPD zu Statinen mit einer vom Verfasser durchgeführten MA-APD verglichen. Die MA-IPD und die MA-APD lieferten bemerkenswert ähnliche Ergebnisse zum Einfluss der Statine auf die Gesamtmortalität. Obwohl der Cochran-Test für Heterogenität zwischen den Primärstudien bezüglich des RR zur Gesamtmortalität signifikant ist, wurde dieses Ergebnis in der MA-IPD weder erwähnt noch bei der Synthese berücksichtigt, d.h. die Synthese erfolgt trotzdem im FEM. Mögliche Heterogenitätsquellen wurden im Rahmen der MA-APD vor der Auswertung bestimmt, und ihre Auswahl wurde durch externe Evidenz begründet. Für die Untersuchung methodischer und klinischer Heterogenität zwischen den RCTs wurden Subgruppen-Analyse und Meta-Regression verwendet. Anders als in der vom Verfasser durchgeführten MA-APD wurden in der publizierten MA-IPD keine methodischen Aspekte der RCTs berücksichtigt. Unter der Berücksichtigung der zufallsbedingten Heterogenität mittels eines REM fanden sowohl die Subgruppen-Analysen als auch die Meta-Regressionen keinen signifikanten Einfluss der methodischen Qualität primärer Studien auf die Gesamtmortalität. Die Variation zwischen den Effektgrößen der RCTs konnte nicht durch die unterschiedliche Basis-LDL-K erklärt werden. Unter Berücksichtigung der methodischen Heterogenität und der Basis-LDL-K fand die Meta-Regression einen signifikant negativen Zusammenhang zwischen dem Frauenanteil einer RCT und dem Ausmaß der Senkung von Gesamtmortalität durch Statine. Die vorliegende Fallstudie deutet auf eine geringfügig verminderte Wirksamkeit der Statine bei der Reduzierung der Gesamtmortalität bei Frauen hin. Es ist empfehlenswert, dieses Ergebnis in zukünftigen RCTs zu Statinen zu prüfen.

Betrachtet man die Ergebnisse der beiden SRs zusammen, zeigen sie, dass selten eine umfassende Berücksichtigung diverser Heterogenitätsursachen bei den MAs von RCTs besteht und dass Confounding bei der Untersuchung von Heterogenitätsursachen in MAs äußerst selten in Erwägung gezogen wird. Eine sorgfältige Berücksichtigung zufallsbedingter und methodischer Variation, wie sie in der Fallstudie zu Statinen gezeigt wurde, ist die Voraussetzung für die Untersuchung von klinischer Heterogenität.

Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Dissertation eigenständig verfasst, ohne unerlaubte Hilfe angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, habe ich durch Quellenangaben im Text deutlich gemacht. Die Dissertation wurde in gleicher oder ähnlicher Form noch keiner anderen Institution vorgelegt. Ich habe mich noch keinem anderen Promotionsverfahren unterzogen oder ein solches beantragt.

Bremen, den 25. März 2008

Abdel Moniem Mukhtar

Literaturverzeichnis

4S. (1994) Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: The scandinavian simvastatin survival study (4S). Lancet 344(8934): 1383-1389.

Abraha I, Bonacini I, Montedori A. (2006) Efficacy and safety of cholesterol-lowering treatment. Lancet 367(9509): 469; author reply 470-1.

Abraham NS, Moayyedi P, Daniels B, Veldhuyzen Van Zanten SJ. (2004) Systematic review: The methodological quality of trials affects estimates of treatment efficacy in functional (non-ulcer) dyspepsia. Aliment Pharmacol Ther 19(6): 631-641.

Adetugbo K, Williams H. (2000) How well are randomized controlled trials reported in the dermatology literature? Arch Dermatol 136(3): 381-385.

Agresti A. (2003) Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. Stat Methods Med Res 12(1): 3-21.

Aher S, Ohlsson A. (2006) Late erythropoietin for preventing red blood cell transfusion in preterm and/or low birth weight infants. Cochrane Database Syst Rev 3: CD004868.

Ahmad SR. (1992) USA: Ban on 415 ineffective drug ingredients. Lancet 340(8819): 598-599.

Albers GW, Amarenco P, Easton JD, Sacco RL, Teal P. (2004) Antithrombotic and thrombolytic therapy for ischemic stroke: The seventh ACCP conference on antithrombotic and thrombolytic therapy. Chest 126(3 Suppl): 483S-512S.

Albert JM, Ioannidis JP, Reichelderfer P, Conway B, Coombs RW, et al. (1998) Statistical issues for HIV surrogate endpoints: Point/counterpoint. an NIAID workshop. Stat Med 17(21): 2435-2462.

Albin RL. (2005) Sham surgery controls are mitigated trolleys. J Med Ethics 31(3): 149-152.

Alderson P, Gliddon L, Chalmers I. (2003) Academic recognition of critical appraisal and systematic reviews in british postgraduate medical education. Med Educ 37(4): 386.

Alderson P, Roberts I. (2000) Should journals publish systematic reviews that find no evidence to guide practice? examples from injury research. BMJ 320(7231): 376-377.

ALLHAT-LLT. (2002) Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: The antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT-LLT). JAMA 288(23): 2998-3007.

Allmark P. (2004) Should research samples reflect the diversity of the population? J Med Ethics 30(2): 185-189.

Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. (2003b) Association of funding and conclusions in randomized drug trials: A reflection of treatment effect or adverse events? JAMA 290(7): 921-928.

Als-Nielsen B, Koretz RL, Kjaergard LL, Gluud C. (2003a) Branched-chain amino acids for hepatic encephalopathy. Cochrane Database Syst Rev (2)(2): CD001939.

Altman D, Chalmers I, editors. (1995) Systematic reviews. London: BMJ Publishing Group

Altman DG, Machin D, Bryant TM, Gardner MJ. (2000) Statistics with confidence: Confidence intervals and statistical guidelines. London: BMJ Publishing Group

Altman DG, Bland JM. (2003) Interaction revisited: The difference between two estimates. BMJ 326(7382): 219.

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, et al. (2001) The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Ann Intern Med 134(8): 663-694.

Altman DG, Matthews JN. (1996) Statistics notes. interaction 1: Heterogeneity of effects. BMJ 313(7055): 486.

Anderson GD. (2005) Sex and racial differences in pharmacological response: Where is the evidence? pharmacogenetics, pharmacokinetics, and pharmacodynamics. J Womens Health (Larchmt) 14(1): 19-29.

Anderson KM, Castelli WP, Levy D. (1987) Cholesterol and mortality. 30 years of follow-up from the framingham study. JAMA 257(16): 2176-2180.

Angelos P. (2007) Sham surgery in clinical trials. JAMA 297(14): 1545-6; author reply 1546.

Antman EM, Bennett JS, Daugherty A, Furberg C, Roberts H, et al. (2007) Use of nonsteroidal antiinflammatory drugs: An update for clinicians: A scientific statement from the american heart association. Circulation 115(12): 1634-1642.

Antman EM, Fox KM. (2000) Guidelines for the diagnosis and management of unstable angina and non-Q-wave myocardial infarction: Proposed revisions. international cardiology forum. Am Heart J 139(3): 461-475.

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. treatments for myocardial infarction. JAMA 268(2): 240-248.

Arends LR, Hoes AW, Lubsen J, Grobbee DE, Stijnen T. (2000) Baseline risk as predictor of treatment benefit: Three clinical meta-re-analyses. Stat Med 19(24): 3497-3518.

Armitage P. (1998) Attitudes in clinical trials. Stat Med 17(23): 2675-2683.

Arora NK, McHorney CA. (2000) Patient preferences for medical decision making: Who really wants to participate? Med Care 38(3): 335-341.

Assmann G, Schulte H. (1987) The prospective cardiovascular munster study: Prevalence and prognostic significance of hyperlipidemia in men with systemic hypertension. Am J Cardiol 59(14): 9G-17G.

Assmann SF, Pocock SJ, Enos LE, Kasten LE. (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 355(9209): 1064-1069.

Athyros VG, Papageorgiou AA, Mercouris BR, Athyrou VV, Symeonidis AN, et al. (2002) Treatment with atorvastatin to the national cholesterol educational program goal versus 'usual' care in secondary coronary heart disease prevention. the GREek atorvastatin and coronary-heart-disease evaluation (GREACE) study. Curr Med Res Opin 18(4): 220-228.

Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, et al. (2005) Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system. BMC Health Serv Res 5(1): 25.

Atkins D, Fink K, Slutsky J, Agency for Healthcare Research and Quality, North American Evidence-based Practice Centers. (2005) Better information for better health care: The evidence-based practice center program and the agency for healthcare research and quality. Ann Intern Med 142(12 Pt 2): 1035-1041.

Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, et al. (2004) Grading quality of evidence and strength of recommendations. BMJ 328(7454): 1490.

Bader J, Ismail A, ADA Council on Scientific Affairs, Division of Science, Journal of the American Dental Association. (2004) Survey of systematic reviews in dentistry. J Am Dent Assoc 135(4): 464-473.

Badgett R, Chiquette E, Anagnostelis B, Mulrow C. (October 1999) Locating reports of serious adverse drug reactions. 7th Cochrane Colloquium Rome

Baigent C, Keech A, Kearney PM, Blackwell L, Buck G, et al. (2005) Efficacy and safety of cholesterol-lowering treatment: Prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet 366(9493): 1267-1278.

Bailey CS, Fisher CG, Dvorak MF. (2004) Type II error in the spine surgical literature. Spine 29(10): 1146-1149.

Bailey KR. (1987) Inter-study differences: How should they influence the interpretation and analysis of results? Stat Med 6(3): 351-360.

Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, et al. (2002) Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 287(22): 2973-2982.

Barrett B, Brown R, Mundt M, Dye L, Alt J, et al. (2005) Using benefit harm tradeoffs to estimate sufficiently important difference: The case of the common cold. Med Decis Making 25(1): 47-55.

Barsky AJ, Saintfort R, Rogers MP, Borus JF. (2002) Nonspecific medication side effects and the nocebo phenomenon. JAMA 287(5): 622-627.

Bartlett C, Doyal L, Ebrahim S, Davey P, Bachmann M, et al. (2005) The causes and effects of socio-demographic exclusions from clinical trials. Health Technol Assess 9(38): iii-iv, ix-x, 1-152.

Bartlett C, Davey P, Dieppe P, Doyal L, Ebrahim S, et al. (2003) Women, older persons, and ethnic minorities: Factors associated with their inclusion in randomised trials of statins 1990 to 2001. Heart 89(3): 327-328.

Barton MB, Miller T, Wolff T, Petitti D, LeFevre M, et al. (2007) How to read the new recommendation statement: Methods update from the U.S. preventive services task force. Ann Intern Med 147(2): 123-127.

Bassler D, Ferreira-Gonzalez I, Briel M, Cook DJ, Devereaux PJ, et al. (2007) Systematic reviewers neglect-Bias that results from trials stopped early for benefit. J Clin Epidemiol 60(9): 869-873.

Bastian H, Bender R, Ernst AS, Kaiser T, Kirchner H, et al. Methoden, version 2.0 vom 19.12.2006, institut für qualität und wirtschaftlichkeit im gesundheitswesen (IQWiG).

Baum M, Budzar AU, Cuzick J, Forbes J, Houghton JH, et al. (2002) Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: First results of the ATAC randomised trial. Lancet 359(9324): 2131-2139.

Bebarta V, Luyten D, Heard K. (2003) Emergency medicine animal research: Does use of randomization and blinding affect the results? Acad Emerg Med 10(6): 684-687.

Begg C, Cho M, Eastwood S, Horton R, Moher D, et al. (1996) Improving the quality of reporting of randomized controlled trials. the CONSORT statement. JAMA 276(8): 637-639.

Begg CB, Mazumdar M. (1994) Operating characteristics of a rank correlation test for publication-Bias. Biometrics 50(4): 1088-1101.

Bekelman JE, Li Y, Gross CP. (2003) Scope and impact of financial conflicts of interest in biomedical research: A systematic review. JAMA 289(4): 454-465.

Bennett DA, Latham NK, Stretton C, Anderson CS. (2004) Capture-recapture is a potentially useful method for assessing publication-Bias. J Clin Epidemiol 57(4): 349-357.

Berger M, Muhlhauser I. (1999) Diabetes care and patient-oriented outcomes. JAMA 281(18): 1676-1678.

Berk PD, Sacks HS. (1999) Assessing the quality of randomized controlled trials: Quality of design is not the only relevant variable. Hepatology 30(5): 1332-1334.

Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. (1995) A random-effects regression model for meta-analysis. Stat Med 14(4): 395-411.

Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI, et al. (2002) Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological-Bias rears its ugly head. Stat Med 21(3): 371-387.

Berlin JA, Colditz GA. (1999) The role of meta-analysis in the regulatory process for foods, drugs, and devices. JAMA 281(9): 830-834.

Berlin JA. (1997) Does blinding of readers affect the results of meta-analyses? university of pennsylvania meta-analysis blinding study group. Lancet 350(9072): 185-186.

Berlin JA, Laird NM, Sacks HS, Chalmers TC. (1989) A comparison of statistical methods for combining event rates from clinical trials. Stat Med 8(2): 141-151.

Bidoli P, Zilembo N, Cortinovis D, Mariani L, Isa L, et al. (2007) Randomized phase II three-arm trial with three platinum-based doublets in metastatic non-small-cell lung cancer. an italian trials in medical oncology study. Ann Oncol 18(3): 461-467.

Biggerstaff BJ, Tweedie RL. (1997) Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Stat Med 16(7): 753-768.

Biondi-Zoccai GG, Lotrionte M, Abbate A, Testa L, Remigi E, et al. (2006) Compliance with QUOROM and quality of reporting of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: Case study. BMJ 332(7535): 202-209.

Bjornsson TD, Wagner JA, Donahue SR, Harper D, Karim A, et al. (2003) A review and assessment of potential sources of ethnic differences in drug responsiveness. J Clin Pharmacol 43(9): 943-967.

Bland JM, Altman DG. (1994) One and two sided tests of significance. BMJ 309(6949): 248.

Boden WE, O'Rourke RA, Teo KK, Hartigan PM, Maron DJ, et al. (2007) Optimal medical therapy with or without PCI for stable coronary disease. N Engl J Med 356(15): 1503-1516.

Bohning D, Sarol J, Rattanasiri S, Viwatwongkasem C, Biggeri A. (2004) A comparison of non-iterative and iterative estimators of heterogeneity variance for the standardized mortality ratio. Biostatistics 5(1): 61-74.

Bohning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, et al. (2002) Some general points in estimating heterogeneity variance with the DerSimonian-laird estimator. Biostatistics 3(4): 445-457.

Boissel JP, Blanchard J, Panak E, Peyrieux JC, Sacks H. (1989) Considerations for the meta-analysis of randomized clinical trials : Summary of a panel discussion. Controlled Clinical Trials 10(3): 254-281.

Bollen CW, Uiterwaal CS, van Vught AJ, van der Tweel I. (2006) Sequential meta-analysis of past clinical trials to determine the use of a new trial. Epidemiology 17(6): 644-649.

Borm GF, Melis RJ, Teerenstra S, Peer PG. (2005) Pseudo cluster randomization: A treatment allocation method to minimize contamination and selection-Bias. Stat Med 24(23): 3535-3547.

Boussageon R, Gueyffier F, Moreau A, Boussageon V. (2006) The difficulty of measurement of placebo effect. Therapie 61(3): 185-190.

Boutron I, Guittet L, Estellat C, Moher D, Hrobjartsson A, et al. (2007) Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. PLoS Med 4(2): e61.

Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, et al. (2006) Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: A systematic review. PLoS Med 3(10): e425.

Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. (2007) Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. Stat Med 26(1): 53-77.

Braitman LE, Rosenbaum PR. (2002) Rare outcomes, common treatments: Analytic strategies using propensity scores. Ann Intern Med 137(8): 693-695.

Bravata DM, McDonald KM, Shojania KG, Sundaram V, Owens DK. (2005) Challenges in systematic reviews: Synthesis of topics related to the delivery, organization, and financing of health care. Ann Intern Med 142(12 Pt 2): 1056-1065.

Brehaut JC, Poses R, Shojania KG, Lott A, Man-Son-Hing M, et al. (2007) Do physician outcome judgments and judgment-Biases contribute to inappropriate use of treatments? study protocol. Implement Sci 2: 18.

Brenner BM, Cooper ME, de Zeeuw D, Keane WF, Mitch WE, et al. (2001) Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. N Engl J Med 345(12): 861-869.

Brewin CR, Bradley C. (1989) Patient preferences and randomised clinical trials. BMJ 299(6694): 313-315.

Brinkhaus B, Pach D, Ludtke R, Willich SN. (2008) Who controls the placebo? introducing a placebo quality checklist for pharmacological trials. Contemp Clin Trials 29(2): 149-156.

Brockwell SE, Gordon IR. (2001) A comparison of statistical methods for meta-analysis. Stat Med 20(6): 825-840.

Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, et al. (2001) Subgroup analyses in randomised controlled trials: Quantifying the risks of false-positives and false-negatives. Health Technol Assess 5(33): 1-56.

Brown CG, Kelen GD, Ashton JJ, Werman HA. (1987) The beta error and sample size determination in clinical trials in emergency medicine. Ann Emerg Med 16(2): 183-187.

Brozek JL, Guyatt GH, Schunemann HJ. (2006) How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health Qual Life Outcomes 4: 69.

Burton A, Altman DG, Royston P, Holder RL. (2006) The design of simulation studies in medical statistics. Stat Med 25(24): 4279-4292.

Buyse M, Molenberghs G. (1998) Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics 54(3): 1014-1029.

Buyse ME. (1989) Analysis of clinical trial outcomes: Some comments on subgroup analyses. Control Clin Trials 10(4 Suppl): 187S-194S.

Caldwell PH, Murphy SB, Butow PN, Craig JC. (2004) Clinical trials in children. Lancet 364(9436): 803-811.

Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, et al. (2000) Framework for design and evaluation of complex interventions to improve health. BMJ 321(7262): 694-696.

Campbell MK, Elbourne DR, Altman DG, CONSORT group. (2004) CONSORT statement: Extension to cluster randomised trials. BMJ 328(7441): 702-708.

Carroll RJ, Ruppert D, Stefanski LA. (1995) Measurement error in non-linear models. New York: Wiley

Carroll D, Tramer M, McQuay H, Nye B, Moore A. (1996) Randomization is important in studies with pain outcomes: Systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. Br J Anaesth 77(6): 798-803.

Carroll RJ, Roeder K, Wasserman L. (1999) Flexible parametric measurement error models. Biometrics 55(1): 44-54.

Carter BL. (2002) Blood pressure as a surrogate end point for hypertension. Ann Pharmacother 36(1): 87-92.

CAST Investigators. (1989) Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. the cardiac arrhythmia suppression trial (CAST) investigators. N Engl J Med 321(6): 406-412.

Caubet JF, Tosteson TD, Dong EW, Naylon EM, Whiting GW, et al. (1997) Maximum androgen blockade in advanced prostate cancer: A meta-analysis of published randomized controlled trials using nonsteroidal antiandrogens. Urology 49(1): 71-78.

Ceballos C, Valdizan JR, Artal A, Almarcegui C, Allepuz C, et al. (2000) Why evidence-based medicine? 20 years of meta-analysis. An Med Interna 17(10): 521-526.

Chalmers I, Hedges LV, Cooper H. (2002) A brief history of research synthesis. Eval Health Prof 25(1): 12-37.

Chalmers I. (2006) Role of systematic reviews in detecting plagiarism: Case of asim kurjak. BMJ 333(7568): 594-596.

Chalmers I. (2005) Academia's failure to support systematic reviews. Lancet 365(9458): 469.

Chan AW, Altman DG. (2005) Epidemiology and reporting of randomised trials published in PubMed journals. Lancet 365(9465): 1159-1162.

Chan AW, Krleza-Jeric K, Schmid I, Altman DG. (2004b) Outcome reporting-Bias in randomized trials funded by the canadian institutes of health research. CMAJ 171(7): 735-740.

Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. (2004a) Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. JAMA 291(20): 2457-2465.

Chan KB, Man-Son-Hing M, Molnar FJ, Laupacis A. (2001) How well is the clinical importance of study results reported? an assessment of randomized controlled trials. CMAJ 165(9): 1197-1202.

Chang BH, Waternaux C, Lipsitz S. (2001) Meta-analysis of binary data: Which within study variance estimate to use? Stat Med 20(13): 1947-1956.

Chapman A, Middleton P, Maddern G. (July 2002) Early updates of systematic reviews - a waste of resources? 4th Symposium on Systematic Reviews: Pushing the Boundaries Oxford

Chen W, Ghosh D, Raghunathan TE, Sargent DJ. (2007) A false-discovery-rate-based loss framework for selection of interactions. Stat Med

Chen Z, Peto R, Collins R, MacMahon S, Lu J, et al. (1991) Serum cholesterol concentration and coronary heart disease in population with low cholesterol concentrations. BMJ 303(6797): 276-282.

Chipman H. (1996) Bayesian variable selection with related predictors. The Canadian Journal of Statistics 24: 17-36.

Choi L, Dominici F, Zeger SL, Ouyang P. (2005) Estimating treatment efficacy over time: A logistic regression model for binary longitudinal outcomes. Stat Med 24(18): 2789-2805.

Christiansen PE, Behnke K, Black CH, Ohrstrom JK, Bork-Rasmussen H, et al. (1996) Paroxetine and amitriptyline in the treatment of depression in general practice. Acta Psychiatr Scand 93(3): 158-163.

Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, et al. (1999) Assessing the quality of randomized trials: Reliability of the jadad scale. Control Clin Trials 20(5): 448-452.

Clark WF, Garg AX, Blake PG, Rock GA, Heidenheim AP, et al. (2003) Effect of awareness of a randomized controlled trial on use of experimental therapy. JAMA 290(10): 1351-1355.

Clarke M, Stewart L. (October 1997) Individual patient data or published meta-analysis: A systematic review. 5th Cochrane Colloquium and Second International Conference Scientific Basis of Health Services Amsterdam

Clarke L, Clarke M, Clarke T. (2007a) How useful are cochrane reviews in identifying research needs? J Health Serv Res Policy 12(2): 101-103.

Clarke M. (2007c) The cochrane collaboration and the cochrane library. Otolaryngol Head Neck Surg 137(4 Suppl): S52-4.

Clarke M, Hopewell S, Chalmers I. (2007b) Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: A status report. J R Soc Med 100(4): 187-190.

Clarke M, Alderson P, Chalmers I. (2002) Discussion sections in reports of controlled trials published in general medical journals. JAMA 287(21): 2799-2801.

Clarke M, Chalmers I. (1998a) Discussion sections in reports of controlled trials published in general medical journals: Islands in search of continents? JAMA 280(3): 280-282.

Clarke M, Godwin J. (1998b) Systematic reviews using individual patient data: A map for the minefields? Ann Oncol 9(8): 827-833.

Colhoun HM, Betteridge DJ, Durrington PN, Hitman GA, Neil HA, et al. (2004) Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the collaborative atorvastatin diabetes study (CARDS): Multicentre randomised placebo-controlled trial. Lancet 364(9435): 685-696.

Collier A, Heilig L, Schilling L, Williams H, Dellavalle RP. (2006) Cochrane skin group systematic reviews are more methodologically rigorous than other systematic reviews in dermatology. Br J Dermatol 155(6): 1230-1235.

Collins R, Peto R, MacMahon S, Hebert P, Fiebach NH, et al. (1990) Blood pressure, stroke, and coronary heart disease. part 2, short-term reductions in blood pressure: Overview of randomised drug trials in their epidemiological context. Lancet 335(8693): 827-838.

Cook DI, Gebski VJ, Keech AC. (2004) Subgroup analysis in clinical trials. Med J Aust 180(6): 289-291.

Cook DJ, Sackett DL, Spitzer WO. (1995) Methodologic guidelines for systematic reviews of randomized control trials in health care from the potsdam consultation on meta-analysis. J Clin Epidemiol 48(1): 167-171.

Cooper H, Hedges LV, editors. (1994) The handbook of research synthesis. New York: Russell Sage Foundation

Cooper NJ, Jones DR, Sutton AJ. (2005) The use of systematic reviews when designing studies. Clin Trials 2(3): 260-264.

Copas J, Jackson D. (2004) A bound for publication-Bias based on the fraction of unpublished studies. Biometrics 60(1): 146-153.

Copas J. (2003) A simple confidence interval for meta-analysis. K. sidik and J. N. jonkman, statistics in medicine 2002; 21:3153-3159. Stat Med 22(16): 2667-2668.

Crowley P, Chalmers I, Keirse MJ. (1990) The effects of corticosteroid administration before preterm delivery: An overview of the evidence from controlled trials. Br J Obstet Gynaecol 97(1): 11-25.

Cuervo LG, Clarke M. (2003) Balancing benefits and harms in health care. BMJ 327(7406): 65-66.

Curb JD, McTiernan A, Heckbert SR, Kooperberg C, Stanford J, et al. (2003) Outcomes ascertainment and adjudication methods in the women's health initiative. Ann Epidemiol 13(9 Suppl): S122-8.

Cuzick J. (1999) Interaction, subgroup analysis and sample size. IARC Sci Publ (148)(148): 109-121.

D'Amico R, Pifferi S, Leonetti C, Torri V, Tinazzi A, et al. (1998) Effectiveness of antibiotic prophylaxis in critically ill adult patients: Systematic review of randomised controlled trials. BMJ 316(7140): 1275-1285.

Davidoff F, Haynes B, Sackett D, Smith R. (1995) Evidence based medicine. BMJ 310(6987): 1085-1086.

Dawes M, Sampson U. (2003) Knowledge management in clinical practice: A systematic review of information seeking behavior in physicians. Int J Med Inform 71(1): 9-15.

de Craen AJ, van Vliet HA, Helmerhorst FM. (2005) An analysis of systematic reviews indicated low incorpororation of results from clinical trial quality assessment. J Clin Epidemiol 58(3): 311-313.

DeAngelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, et al. (2004) Registration of clinical trials: A statement from the international committee of medical journal editors. Ned Tijdschr Geneeskd 148(38): 1870-1871.

DECODE. (1998) Will new diagnostic criteria for diabetes mellitus change phenotype of patients with diabetes? reanalysis of european epidemiological data. DECODE study group on behalf of the european diabetes epidemiology study group. BMJ 317(7155): 371-375.

Dehue T. (2000) From deception trials to control reagents. the introduction of the control group about a century ago. Am Psychol 55(2): 264-268.

Delaney A, Bagshaw SM, Ferland A, Laupland K, Manns B, et al. (2007) The quality of reports of critical care meta-analyses in the cochrane database of systematic reviews: An independent appraisal. Crit Care Med 35(2): 589-594.

Delgado-Rodríguez M. (2001) Glossary on meta-analysis. J Epidemiol Community Health 55(8): 534-536.

Demirtas H. (2007) The design of simulation studies in medical statistics by andrea burton, douglas G. altman, patrick royston and roger L. holder, statistics in medicine 2006; 25:4279-4292. Stat Med 26(20): 3818-3821.

Derksen S, Keselman HJ. (1992) Backward, forward, and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Statist Psych 45: 265-282.

Derry S, Kong Loke Y, Aronson JK. (2001) Incomplete evidence: The inadequacy of databases in tracing published adverse drug reactions in clinical trials. BMC Med Res Methodol 1: 7.

DerSimonian R, Kacker R. (2007) Random-effects model for meta-analysis of clinical trials: An update. Contemp Clin Trials 28(2): 105-114.

DerSimonian R, Laird N. (1986) Meta-analysis in clinical trials. Control Clin Trials 7(3): 177-188.

Devereaux PJ, Bhandari M, Clarke M, Montori VM, Cook DJ, et al. (2005) Need for expertise based randomised controlled trials. BMJ 330(7482): 88.

Devereaux PJ, Choi PT, El-Dika S, Bhandari M, Montori VM, et al. (2004) An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. J Clin Epidemiol 57(12): 1232-1236.

Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, et al. (2001) Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: Observational study. BMJ 323(7323): 1218-1222.

Di Blasi Z, Harkness E, Ernst E, Georgiou A, Kleijnen J. (2001) Influence of context effects on health outcomes: A systematic review. Lancet 357(9258): 757-762.

Dicker A, Armstrong D. (1995) Patients' views of priority setting in health care: An interview survey in one practice. BMJ 311(7013): 1137-1139.

Dickersin K, Rennie D. (2003) Registering clinical trials. JAMA 290(4): 516-523.

Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, et al. (2002) Development of the cochrane collaboration's CENTRAL register of controlled clinical trials. Eval Health Prof 25(1): 38-64.

Dickersin K, Berlin JA. (1992) Meta-analysis: State-of-the-science. Epidemiol Rev 14: 154-176.

Diggle PJ, Verbyla AP. (1998) Nonparametric estimation of covariance structure in longitudinal data. Biometrics 54(2): 401-415.

Dincer F, Linde K. (2003) Sham interventions in randomized clinical trials of acupuncture--a review. Complement Ther Med 11(4): 235-242.

Djulbegovic B, Frohlich A, Bennett CL. (2005) Acting on imperfect evidence: How much regret are we ready to accept? J Clin Oncol 23(28): 6822-6825.

Djulbegovic B, Lacevic M, Cantor A, Fields KK, Bennett CL, et al. (2000) The uncertainty principle and industry-sponsored research. Lancet 356(9230): 635-638.

Dobson AJ. (2001) Introduction to generalized linear models. London: Chapman and Hall/CRC

Dohoo I, Stryhn H, Sanchez J. (2007) Evaluation of underlying risk as a source of heterogeneity in meta-analyses: A simulation study of bayesian and frequentist implementations of three models. Prev Vet Med 81(1-3): 38-55.

Downs JR, Clearfield M, Weis S, Whitney E, Shapiro DR, et al. (1998) Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: Results of AFCAPS/TexCAPS. air Force/Texas coronary atherosclerosis prevention study. JAMA 279(20): 1615-1622.

Draborg E, Gyrd-Hansen D, Poulsen PB, Horder M. (2005) International comparison of the definition and the practical application of health technology assessment. Int J Technol Assess Health Care 21(1): 89-95.

Dubben HH, Beck-Bornholdt HP. (2005) Systematic review of publication-Bias in studies on publication-Bias. BMJ 331(7514): 433-434.

Duchateau L, Pignon JP, Bijnens L, Bertin S, Bourhis J, et al. (2001) Individual patient-versus literature-based meta-analysis of survival data: Time to event and event rate at a particular time can make a difference, an example based on head and neck cancer. Control Clin Trials 22(5): 538-547.

Dundar Y, Dodd S, Dickson R, Walley T, Haycox A, et al. (2006) Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies. Health Technol Assess 10(5): iii-iv, ix-145.

Duval S, Tweedie R. (2000) Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication-Bias in meta-analysis. Biometrics 56(2): 455-463.

Edlund MJ, Overall JE, Rhoades HM. (1985) Beta, or type II error in psychiatric controlled clinical trials. J Psychiatr Res 19(4): 563-567.

Edwards JE, Moore RA. (2003) Statins in hypercholesterolaemia: A dose-specific meta-analysis of lipid changes in randomised, double blind trials. BMC Fam Pract 4: 18.

Egger M. (2005) Systematic reviews in health care : Meta-analysis in context. London: BMJ Books.

Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? empirical study. Health Technol Assess 7(1): 1-76.

Egger M, Davey Smith G, Schneider M, Minder C. (1997)-Bias in meta-analysis detected by a simple, graphical test. BMJ 315(7109): 629-634.

Eisenbud R, Assmann SF, Kalish LA, van Der Horst C, Collier AC, et al. (2001) Differences in difficulty adjudicating clinical events in patients with advanced HIV disease. J Acquir Immune Defic Syndr 28(1): 43-46.

Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. (2004) Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. Clin Trials 1(1): 80-90.

Ellenberg SS, Temple R. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 2: Practical issues and specific cases. Ann Intern Med 133(6): 464-470.

EMEA. (2006) Draft: Guideline on clinical trials in small populations. European Medicines Agency (EMEA) CHMP/EWP/83561

EMEA. (2002) Points to consider on multiplicity issues in clinical trials. European Medicines Agency (EMEA) CPMP/EWP/908/99

EMEA. (2001) Points to consider on validity and interpretation of meta-analyses, and one pivotal study. European Medicines Agency (EMEA) CPMP/EWP/2330/99

Emerson JD. (1994) Combining estimates of the odds ratio: The state of the art. Stat Methods Med Res 3(2): 157-178.

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. (2000) Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. Stat Med 19(13): 1707-1728.

Ernst E, Harkness E. (2001) Spinal manipulation: A systematic review of sham-controlled, double-blind, randomized clinical trials. J Pain Symptom Manage 22(4): 879-889.

Ernst E, Pittler MH. (2001) Assessment of therapeutic safety in systematic reviews: Literature review. BMJ 323(7312): 546.

European Commission. Promotion of paediatric research in the european union. 2002. (http://www.efpia.org/4_pos/sci_regu/Paeds020123.pdf) Zugang am 8. Juni 2006

European Environment Agency. (2001) Late lessons from early warnings: The precautionary principle 1896–2000. Environmental Issue Report No. 22

FDA. (2006) Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: Draft guidance. U.S. department of health and human services FDA center for drug evaluation and research; U.S. department of health and human services FDA center for biologics evaluation and research; U.S. department of health and human services FDA center for devices and radiological health. Health Qual Life Outcomes 4: 79.

Feinstein AR. (2001) The blame-X syndrome: Problems and lessons in nosology, spectrum, and etiology. J Clin Epidemiol 54(5): 433-439.

Feinstein AR. (1998) The problem of cogent subgroups: A clinicostatistical tragedy. J Clin Epidemiol 51(4): 297-299.

Fenton JJ, Elmore JG. (2004) Balancing mammography's benefits and harms. BMJ 328(7453): E301-2.

Fergusson D, Glass KC, Hutton B, Shapiro S. (2005) Randomized controlled trials of aprotinin in cardiac surgery: Could clinical equipoise have stopped the bleeding? Clin Trials 2(3): 218-29; discussion 229-32.

Fergusson D, Laupacis A, Salmi LR, McAlister FA, Huet C. (2000) What should be included in meta-analyses? an exploration of methodological issues using the ISPOT meta-analyses. Int J Technol Assess Health Care 16(4): 1109-1119.

Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, et al. (2007) Problems with use of composite end points in cardiovascular trials: Systematic review of randomised controlled trials. BMJ 334(7597): 786.

Fibrinolytic Therapy Trialists'. (1994) Indications for fibrinolytic therapy in suspected acute myocardial infarction: Collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. fibrinolytic therapy trialists' (FTT) collaborative group. Lancet 343(8893): 311-322.

Flanagin A, Fontanarosa PB, DeAngelis CD. (2002) Authorship for research groups. JAMA 288(24): 3166-3168.

Fleiss JL. (1993) The statistical basis of meta-analysis. Stat Methods Med Res 2(2): 121-145.

Fleiss JL. (1986) Analysis of data from multiclinic trials. Control Clin Trials 7(4): 267-275.

Fortin PR, Lew RA, Liang MH, Wright EA, Beckett LA, et al. (1995) Validation of a meta-analysis: The effects of fish oil in rheumatoid arthritis. J Clin Epidemiol 48(11): 1379-1390.

Frank S, Kieburtz K, Holloway R, Kim SY. (2005) What is the risk of sham surgery in parkinson disease clinical trials? A review of published reports. Neurology 65(7): 1101-1103.

Franzosi MG, Santoro E, Santoro L. (1997b) Prospective meta-analysis using individual patient data vs meta-analysis of published reports: The case of ACE-inhibitors in myocardial infarction Controlled Clinical Trials 18(3, Supplement 1): S183.

Freemantle N, Calvert M. (2007) Composite and surrogate outcomes in randomised controlled trials. BMJ 334(7597): 756-757.

Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. (2003) Composite outcomes in randomized trials: Greater precision but with greater uncertainty? JAMA 289(19): 2554-2559.

French SD, McDonald S, McKenzie JE, Green SE. (2005) Investing in updating: How do conclusions change when cochrane systematic reviews are updated? BMC Med Res Methodol 5: 33.

Fritz E, Ludwig H. (2000) Interferon-alpha treatment in multiple myeloma: Meta-analysis of 30 randomised trials among 3948 patients. Ann Oncol 11(11): 1427-1436.

Furberg CD. (1983) Effect of antiarrhythmic drugs on mortality after myocardial infarction. Am J Cardiol 52(6): 32C-36C.

Furlan AD, Clarke J, Esmail R, Sinclair S, Irvin E, et al. (2001) A critical review of reviews on the treatment of chronic low back pain. Spine 26(7): E155-62.

Gage BF, van Walraven C, Pearce L, Hart RG, Koudstaal PJ, et al. (2004) Selecting patients with atrial fibrillation for anticoagulation: Stroke risk stratification in patients taking aspirin. Circulation 110(16): 2287-2292.

Geddes J, Freemantle N, Harrison P, Bebbington P. (2000) Atypical antipsychotics in the treatment of schizophrenia: Systematic overview and meta-regression analysis. BMJ 321(7273): 1371-1376.

Gelfand LA, Strunk DR, Tu XM, Noble RE, Derubeis RJ. (2006)-Bias resulting from the use of 'assay sensitivity' as an inclusion criterion for meta-analysis. Stat Med 25(6): 943-955.

Gheorghiade M, Schultz L, Tilley B, Kao W, Goldstein S. (1991) Subgroup analysis of clinical trials. Am J Cardiol 67(4): 330-332.

Gifford F. (2001) Uncertainty about clinical equipoise. clinical equipoise and the uncertainty principles both require further scrutiny. BMJ 322(7289): 795.

GISSI-P. (2000) Results of the low-dose (20 mg) pravastatin GISSI prevenzione trial in 4271 patients with recent myocardial infarction: Do stopped trials contribute to overall knowledge? GISSI prevenzione investigators (gruppo italiano per lo studio della sopravvivenza nell'infarto miocardico). Ital Heart J 1(12): 810-820.

Glass GV. (1976) Primary, secondary, and meta-analysis of research. Educational Researcher 5: 3-8.

Glasziou P, Meats E, Heneghan C, Shepperd S. (September 2007) Inadequate descriptions of treatments in published reports: A common but correctable barrier to research uptake. Third International Clinical Trials Symposium: Improving Health Care in the New Millennium Sydney

Glasziou P, Chalmers I, Rawlins M, McCulloch P. (2007) When are randomised trials unnecessary? picking signal from noise. BMJ 334(7589): 349-351.

Glasziou P, Djulbegovic B, Burls A. (2006) Are systematic reviews more cost-effective than randomised trials? Lancet 367(9528): 2057-2058.

Glasziou P, Guyatt GH, Dans AL, Dans LF, Straus S, et al. (1998) Applying the results of trials and systematic reviews to individual patients. ACP J Club 129(3): A15-6.

Gluud C, Christensen E. (2001) Ursodeoxycholic acid for primary biliary cirrhosis. Cochrane Database Syst Rev (1)(1): CD000551.

Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, et al. (2003) Pragmatic controlled clinical trials in primary care: The struggle between external and internal validity. BMC Med Res Methodol 3: 28.

Goldbeck-Wood S. (1998) Denmark takes a lead on research ethics. BMJ 316(7139): 1189.

Golder S, McIntosh HM, Duffy S, Glanville J, Centre for Reviews and Dissemination and UK Cochrane Centre Search Filters Design Group. (2006c) Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J 23(1): 3-12.

Golder S, McIntosh HM, Loke Y. (2006b) Identifying systematic reviews of the adverse effects of health care interventions. BMC Med Res Methodol 6: 22.

Golder S, Loke Y, McIntosh HM. (2006a) Room for improvement? A survey of the methods used in systematic reviews of adverse effects. BMC Med Res Methodol 6: 3.

Goldstein H. (2003) Multilevel statistical models. London: Arnold

Goodman SN. (2007) Stopping at nothing? some dilemmas of data monitoring in clinical trials. Ann Intern Med 146(12): 882-887.

Goodman SN, Berlin JA. (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 121(3): 200-206.

Gorelick P, Sechenova O, Hennekens CH. (2006) Evolving perspectives on clopidogrel in the treatment of ischemic stroke. J Cardiovasc Pharmacol Ther 11(4): 245-248.

Gottlieb SS, McCarter RJ, Vogel RA. (1998) Effect of beta-blockade on mortality among high-risk and low-risk patients after myocardial infarction. N Engl J Med 339(8): 489-497.

Gotzsche PC, Hrobjartsson A, Maric K, Tendal B. (2007) Data extraction errors in meta-analyses that use standardized mean differences. JAMA 298(4): 430-437.

Gotzsche PC. (1994) Is there logic in the placebo? Lancet 344(8927): 925-926.

Gotzsche PC. (1989a) Methodology and overt and hidden-Bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Control Clin Trials 10(1): 31-56.

Grady D, Chaput L, Kristof M. (2003) Diagnosis and treatment of coronary heart disease in women: Systematic reviews of evidence on selected topics. Evid Rep Technol Assess (Summ) (81)(81): 1-4.

Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, et al. (2005) Issues in data monitoring and interim analysis of trials. Health Technol Assess 9(7): 1-238, iii-iv.

Greenland S, O'Rourke K. (2001) On the-Bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2(4): 463-471.

Greenland S. (1994) Invited commentary: A critical look at some popular meta-analytic methods. Am J Epidemiol 140(3): 290-296.

Greenland S, Salvan A. (1990)-Bias in the one-step method for pooling study results. Stat Med 9(3): 247-252.

Gregoire G, Derderian F, Le Lorier J. (1995) Selecting the language of the publications included in a meta-analysis: Is there a tower of babel-Bias? J Clin Epidemiol 48(1): 159-163.

Griffiths F, Green E, Tsouroufli M. (2005) The nature of medical evidence and its inherent uncertainty for the clinical consultation: Qualitative study. BMJ 330(7490): 511.

Grimes DA, Hubacher D, Nanda K, Schulz KF, Moher D, et al. (2005) The good clinical practice guideline: A bronze standard for clinical research. Lancet 366(9480): 172-174.

Grimshaw J. (2004) So what has the cochrane collaboration ever done for us? A report card on the first 10 years. CMAJ 171(7): 747-749.

Gueyffier F, Boissel JP, Pocock S, Boutitie F, Coope J, et al. (1999) Identification of risk factors in hypertensive patients: Contribution of randomized controlled trials through an individual patient database. Circulation 100(18): e88-94.

Guillemin F, Associate Editor for the Cochrane Back Review Group. (2006) The cochrane collaboration is in its fourteenth year. Joint Bone Spine 73(3): 236-238.

Guirguis-Blake J, Calonge N, Miller T, Siu A, Teutsch S, et al. (2007) Current processes of the U.S. preventive services task force: Refining evidence-based recommendation development. Ann Intern Med 147(2): 117-122.

Gunnell D, Ashby D. (2004) Antidepressants and suicide: What is the balance of benefit and harm. BMJ 329(7456): 34-38.

Guyatt G, Montori V, Devereaux PJ, Schunemann H, Bhandari M. (2004) Patients at the center: In our practice, and in our use of language. ACP J Club 140(1): A11-2.

Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, et al. (1986) Determining optimal therapy--randomized trials in individual patients. N Engl J Med 314(14): 889-892.

Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, et al. (2002) Methods to explain the clinical significance of health status measures. Mayo Clin Proc 77(4): 371-383.

Hadhazy V, Ezzo J, Berman B. (October 1999) How valuable is effort to contact authors to obtain missing data in systematic reviews. 7th Cochrane Colloquium Rome

Hahn S, Puffer S, Torgerson DJ, Watson J. (2005) Methodological-Bias in cluster randomised trials. BMC Med Res Methodol 5(1): 10.

Hall SM, Brannick MT. (2002) Comparison of two random-effects methods of meta-analysis. J Appl Psychol 87(2): 377-389.

Harden A, Garcia J, Oliver S, Rees R, Shepherd J, et al. (2004) Applying systematic review methods to studies of people's views: An example from public health research. J Epidemiol Community Health 58(9): 794-800.

Hardy RJ, Thompson SG. (1998) Detecting and describing heterogeneity in meta-analysis. Stat Med 17(8): 841-856.

Hardy RJ, Thompson SG. (1996) A likelihood approach to meta-analysis with random effects. Stat Med 15(6): 619-629.

Harrell FE. (2001) Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis. New York: Springer

Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, et al. (2001) Current methods of the US preventive services task force: A review of the process. Am J Prev Med 20(3 Suppl): 21-35.

Harrison JE. (2003) Clinical trials in orthodontics II: Assessment of the quality of reporting of clinical trials published in three orthodontic journals between 1989 and 1998. J Orthod 30(4): 309-15; discussion 297-8.

Hart D. (2005) Risk-benefit evaluation of medicinal products. an element of health technology assessment. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 48(2): 204-214.

Hart RG, Pearce LA, Aguilar MI. (2007) Meta-analysis: Antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. Ann Intern Med 146(12): 857-867.

Hartung J, Knapp G. (2001b) A refined method for the meta-analysis of controlled clinical trials with binary outcome. Stat Med 20(24): 3875-3889.

Hartung J, Knapp G. (2001a) On tests of the overall treatment effect in meta-analysis with normally distributed responses. Stat Med 20(12): 1771-1782.

Havelaar AH, De Hollander AE, Teunis PF, Evers EG, Van Kranen HJ, et al. (2000) Balancing the risks and benefits of drinking water disinfection: Disability adjusted life-years on the scale. Environ Health Perspect 108(4): 315-321.

Hawe P, Shiell A, Riley T. (2004) Complex interventions: How "out of control" can a randomised controlled trial be? BMJ 328(7455): 1561-1563.

Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, et al. (2005) Optimal search strategies for retrieving scientifically strong studies of treatment from medline: Analytical survey. BMJ 330(7501): 1179.

Heath I. (2007) The road to hell.. BMJ 335(7631): 1185.

Hedges LV, Vevea JL. (1998) Fixed- and random-effects models in meta-analysis. Psychological Methods 3(4): 486-504.

Hedges LV, Olkin I. (1985) Statistical methods for meta-analysis. Orlando: Academic Press

Hedges LV, Pigott TD. (2004) The power of statistical tests for moderators in meta-analysis. Psychol Methods 9(4): 426-445.

Hedges LV, Pigott TD. (2001) The power of statistical tests in meta-analysis. Psychol Methods 6(3): 203-217.

Heiat A, Gross CP, Krumholz HM. (2002) Representation of the elderly, women, and minorities in heart failure clinical trials. Arch Intern Med 162(15): 1682-1688.

Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. (2006) Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? Am Heart J 151(2): 257-264.

Hetherington J, Dickersin K, Chalmers I, Meinert CL. (1989) Retrospective and prospective identification of unpublished controlled trials: Lessons from a survey of obstetricians and pediatricians. Pediatrics 84(2): 374-380.

Higgins J. (October 1999) How should we interpret updated meta-analyses? 7th Cochrane Colloquium Rome

Higgins JPT, Green S, editors. (2006) Cochrane handbook for systematic reviews of interventions 4.2.6 [updated september 2006]. In: The Cochrane Library, Issue 4, 2006 Chichester, UK: John Wiley & Sons, Ltd.

Higgins J, Thompson S, Deeks J, Altman D. (2002b) Statistical heterogeneity in systematic reviews of clinical trials: A critical appraisal of guidelines and practice. J Health Serv Res Policy 7(1): 51-61.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. (2003) Measuring inconsistency in meta-analyses. BMJ 327(7414): 557-560.

Higgins JP, Thompson SG. (2002a) Quantifying heterogeneity in a meta-analysis. Stat Med 21(11): 1539-1558.

Hill CL, LaValley MP, Felson DT. (2002) Discrepancy between published report and actual conduct of randomized clinical trials. J Clin Epidemiol 55(8): 783-786.

Holbrook A, Labiris R, Goldsmith CH, Ota K, Harb S, et al. (2007) Influence of decision aids on patient preferences for anticoagulant therapy: A randomized trial. CMAJ 176(11): 1583-1587.

Holdaas H, Fellstrom B, Jardine AG, Holme I, Nyberg G, et al. (2003) Effect of fluvastatin on cardiac outcomes in renal transplant recipients: A multicentre, randomised, placebo-controlled trial. Lancet 361(9374): 2024-2031.

Hollis S, Campbell F. (1999) What is meant by intention to treat analysis? survey of published randomised controlled trials. BMJ 319(7211): 670-674.

Hopayian K. (2001) The need for caution in interpreting high quality systematic reviews. BMJ 323(7314): 681-684.

Hopewell S, Clarke M, Stewart L, Tierney J. (2007b) Time to publication for results of clinical trials. Cochrane Database Syst Rev (2)(2): MR000011.

Hopewell S, McDonald S, Clarke M, Egger M. (2007a) Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev (2)(2): MR000010.

Hopewell S, Clarke M, Askie L. (2006) Reporting of trials presented in conference abstracts needs to be improved. J Clin Epidemiol 59(7): 681-684.

Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. (2002) A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. Stat Med 21(11): 1625-1634.

Hostetter TH. (2001) Prevention of end-stage renal disease due to type 2 diabetes. N Engl J Med 345(12): 910-912.

HPS. (2002) MRC/BHF heart protection study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: A randomised placebo-controlled trial. heart protection study collaborative group (HPS). Lancet 360(9326): 7-22.

Hrobjartsson A, Gotzsche PC. (2004) Is the placebo powerless? update of a systematic review with 52 new randomized trials comparing placebo with no treatment. J Intern Med 256(2): 91-100.

Huang JQ, Zheng GF, Irvine EJ, Karlberg J. (2004) Assessing heterogeneity in meta-analyses of helicobacter pylori infection-related clinical studies: A critical appraisal. Chin J Dig Dis 5(3): 126-133.

Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. (2006) Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods 11(2): 193-206.

Hunt DL, McKibbon KA. (1997) Locating and appraising systematic reviews. Ann Intern Med 126(7): 532-538.

Hunt MM. (1997) How science takes stock : The story of meta-analysis. New York: Russell Sage Found.

Hunter JE, Schmit FL. (1990) Methods of meta-analysis: Correcting error and-Bias in research findings. Newbury Park, CA: SAGE Publications

Huth EJ. (1986) Irresponsible authorship and wasteful publication. Ann Intern Med 104(2): 257-259.

Iberg JR. (1991) Applying statistical control theory to bring together clinical supervision and psychotherapy research. J Consult Clin Psychol 59(4): 575-586.

ICH. (2000) E10: Choice of control group and related issues in clinical trials. International Conference on Harmonisation (ICH) CPMP/ICH/364/96

ICH. (1998b) E9: Statistical principles for clinical trials. International Conference on Harmonisation (ICH) CPMP/ICH/363/96

ICH. (1998a) E5(R1): Ethnic factors in the acceptability of foreign clinical data. International Conference on Harmonisation (ICH) CPMP/ICH/289/95

ICH. (1993) E7: Studies in support of special populations : Geriatrics. International Conference on Harmonisation (ICH) CPMP/ICH/379/95

IMPACT. (1995) Efficacy of adjuvant fluorouracil and folinic acid in colon cancer. international multicentre pooled analysis of colon cancer trials (IMPACT) investigators. Lancet 345(8955): 939-944.

Ioannidis JP, Patsopoulos NA, Evangelou E. (2007c) Uncertainty in heterogeneity estimates in meta-analyses. BMJ 335(7626): 914-916.

Ioannidis JP. (2007b) Limitations are not properly acknowledged in the scientific literature. J Clin Epidemiol 60(4): 324-329.

Ioannidis JP, Trikalinos TA. (2007a) The appropriateness of asymmetry tests for publication-Bias in meta-analyses: A large survey. CMAJ 176(8): 1091-1096.

Ioannidis JP, Mulrow CD, Goodman SN. (2006) Adverse events: The more you search, the more you find. Ann Intern Med 144(4): 298-300.

Ioannidis JP. (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, et al. (2004) Better reporting of harms in randomized trials: An extension of the CONSORT statement. Ann Intern Med 141(10): 781-788.

Ioannidis JP, Contopoulos-Ioannidis DG, Lau J. (1999) Recursive cumulative meta-analysis: A diagnostic for the evolution of total randomized evidence from group and individual patient data. J Clin Epidemiol 52(4): 281-291.

Ioannidis JP. (1998) Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 279(4): 281-286.

Ioannidis JP, Cappelleri JC, Sacks HS, Lau J. (1997) The relationship between study design, results, and reporting of randomized clinical trials of HIV infection. Control Clin Trials 18(5): 431-444.

IOM. (2001) Exploring the biological contributions to human health: Does sex matter? Institute of Medicine (IOM) http://www.iom.edu/Object.File/Master/4/130/DoesSexMatter8pager.pdf(Zugang am 3. August 2005)

Jackson D. (2006) The power of the standard test for the presence of heterogeneity in meta-analysis. Stat Med 25(15): 2688-2699.

Jadad AR, Moher M, Browman GP, Booker L, Sigouin C, et al. (2000) Systematic reviews and meta-analyses on treatment of asthma: Critical evaluation. BMJ 320(7234): 537-540.

Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, et al. (1998) Methodology and reports of systematic reviews and meta-analyses: A comparison of cochrane reviews with articles published in paper-based journals. JAMA 280(3): 278-280.

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, et al. (1996b) Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 17(1): 1-12.

Jadad AR, McQuay HJ. (1996a) Meta-analyses to evaluate analgesic interventions: A systematic qualitative review of their methodology. J Clin Epidemiol 49(2): 235-243.

Jeng GT, Scott JR, Burmeister LF. (1995a) A comparison of meta-analytic results using literature vs individual patient data. paternal cell immunization for recurrent miscarriage. JAMA 274(10): 830-836.

Jennison C, Turnbull BW. (1993) Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. Biometrics 49(3): 741-752.

Johansen HK, Gotzsche PC. (1999) Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. JAMA 282(18): 1752-1759.

Jorgensen AW, Hilden J, Gotzsche PC. (2006) Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: Systematic review. BMJ 333(7572): 782.

Juni P, Tallon D, Egger M. (July 2000) 'Garbage in - garbage out'? assessment of the quality of controlled trials in meta-analyses published in leading journals. Third Symposium on Systematic Reviews: Beyond the Basics Oxford

Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, et al. (2004) Risk of cardiovascular events and rofecoxib: Cumulative meta-analysis. Lancet 364(9450): 2021-2029.

Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. (2002) Direction and impact of language-Bias in meta-analyses of controlled trials: Empirical study. Int J Epidemiol 31(1): 115-123.

Juni P, Witschi A, Bloch R, Egger M. (1999) The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282(11): 1054-1060.

Kacker RN. (2004) Combining information from interlaboratory evaluations using a random effects model. Metrologia 41(3): 132-136.

Kajermo KN, Nordstrom G, Krusebrant A, Lutzen K. (2001) Nurses' experiences of research utilization within the framework of an educational programme. J Clin Nurs 10(5): 671-681.

Kamprath S, Timmer A. (März 2007) Bedarf an systematischen übersichtsarbeiten am beispiel des diabetes mellitus typ 2. 8. Jahrestagung des Deutschen Netzwerks Evidenzbasierte Medizin Berlin

Kaptchuk TJ, Stason WB, Davis RB, Legedza AR, Schnyer RN, et al. (2006) Sham device v inert pill: Randomised controlled trial of two placebo treatments. BMJ 332(7538): 391-397.

Kearney PM, Baigent C, Godwin J, Halls H, Emberson JR, et al. (2006) Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? meta-analysis of randomised trials. BMJ 332(7553): 1302-1308.

Kent G. (1996) Volunteering children for bone marrow donation. studies show large discrepancies between views of surrogate decision makers and patients. BMJ 313(7048): 49-50.

Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. (1999) Stratified randomization for clinical trials. J Clin Epidemiol 52(1): 19-26.

Kerr NL. (1998) HARKing: Hypothesizing after the results are known. Pers Soc Psychol Rev 2(3): 196-217.

Khan KS, ter Riet G, Glanville J, Sowden A, Kleijnen J. (2001) Undertaking systematic reviews of research on effectiveness, CRD's guidance for those carrying out or commissioning reviews. Centre for Reviews and Dissemination CRD Report Number 4 (2nd Edition)(York, UK: York Publishing Ltd)

Khan KS, Daya S, Collins JA, Walter SD. (1996b) Empirical evidence of-Bias in infertility research: Overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. Fertil Steril 65(5): 939-945.

Khan KS, Daya S, Jadad A. (1996a) The importance of quality of primary studies in producing unbiased systematic reviews. Arch Intern Med 156(6): 661-666.

Kieser M, Friede T. (2007) Planning and analysis of three-arm non-inferiority trials with binary endpoints. Stat Med 26(2): 253-273.

King M, Nazareth I, Lampe F, Bower P, Chandler M, et al. (2005) Conceptual framework and systematic review of the effects of participants' and professionals' preferences in randomised controlled trials. Health Technol Assess 9(35): 1-186, iii-iv.

Kirwan BA, Lubsen J, de Brouwer S, Danchin N, Battler A, et al. (2007) Diagnostic criteria and adjudication process both determine published event-rates: The ACTION trial experience. Contemp Clin Trials 28(6): 720-729.

Kjaergard LL, Villumsen J, Gluud C. (2001) Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 135(11): 982-989.

Kjaergard LL, Nikolova D, Gluud C. (1999) Randomized clinical trials in HEPATOLOGY: Predictors of quality. Hepatology 30(5): 1134-1138.

Klar N, Donner A. (2001) Current and future challenges in the design and analysis of cluster randomization trials. Stat Med 20(24): 3729-3740.

Knapp G, Biggerstaff BJ, Hartung J. (2006) Assessing the amount of heterogeneity in random-effects meta-analysis. Biom J 48(2): 271-285.

Knapp G, Hartung J. (2003) Improved tests for a random effects meta-regression with a single covariate. Stat Med 22(17): 2693-2710.

Konstam MA, Weir MR, Reicin A, Shapiro D, Sperling RS, et al. (2001) Cardiovascular thrombotic events in controlled, clinical trials of rofecoxib. Circulation 104(19): 2280-2288.

Koopman L, van der Heijden GJ, Glasziou PP, Grobbee DE, Rovers MM. (2007) A systematic review of analytical methods used to study subgroups in (individual patient data) meta-analyses. J Clin Epidemiol 60(10): 1002-1009.

Kotzin S. (August 2005)
Journal selection for medline. World Library and Information Congress: 71st IFLA General Conference and Council Oslo

Kraemer HC, Frank E, Kupfer DJ. (2006) Moderators of treatment outcomes: Clinical, research, and policy importance. JAMA 296(10): 1286-1289.

Kraemer HC, Robinson TN. (2005) Are certain multicenter randomized clinical trial structures misleading clinical and policy decisions? Contemp Clin Trials 26(5): 518-529.

Kravitz RL, Duan N, Braslow J. (2004) Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 82(4): 661-687.

Kruse RL, Alper BS, Reust C, Stevermer JJ, Shannon S, et al. (2002) Intention-to-treat analysis: Who is in? who is out? J Fam Pract 51(11): 969-971.

Kulinskaya E, Dollinger MB, Knight E, Gao H. (2004) A welch-type test for homogeneity of contrasts under heteroscedasticity with application to meta-analysis. Stat Med 23(23): 3655-3670.

Kunz R, Vist G, Oxman AD. (2002) Randomisation to protect against selection-Bias in healthcare trials. Cochrane Database Syst Rev (2)(2): MR000012.

Kunz R, Oxman AD. (1998) The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 317(7167): 1185-1190.

Lagakos SW. (2006) The challenge of subgroup analyses--reporting without distorting. N Engl J Med 354(16): 1667-1669.

Laine C, Horton R, DeAngelis CD, Drazen JM, Frizelle FA, et al. (2007) Clinical trial registration: Looking back and moving ahead. Lancet 369(9577): 1909-1911.

Lambert PC, Sutton AJ, Abrams KR, Jones DR. (2002) A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J Clin Epidemiol 55(1): 86-94.

LaRosa JC, He J, Vupputuri S. (1999) Effect of statins on risk of coronary disease: A meta-analysis of randomized controlled trials. JAMA 282(24): 2340-2346.

Lassere M. (2007) The biomarker-surrogacy evaluation schema: A review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of

evidence schema for evaluating the status of biomarkers as surrogate endpoints. Stat Methods Med Res

Last JM, Spasoff RA, International Epidemiological Association. (2001) A dictionary of epidemiology. Oxford u.a.: Oxford Univ. Press.

Latronico N, Botteri M, Minelli C, Zanotti C, Bertolini G, et al. (2002) Quality of reporting of randomised controlled trials in the intensive care literature. A systematic analysis of papers published in intensive care medicine over 26 years. Intensive Care Med 28(9): 1316-1323.

Lau J, Ioannidis JP, Schmid CH. (1998) Summing up evidence: One answer is not always enough. Lancet 351(9096): 123-127.

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, et al. (1992) Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 327(4): 248-254.

Laupacis A, Straus S. (2007) Systematic reviews: Time to address clinical and policy relevance as well as methodological rigor. Ann Intern Med 147(4): 273-274.

Lavis JN, Ross SE, Hurley JE, Hohenadel JM, Stoddart GL, et al. (2002) Examining the role of health services research in public policymaking. Milbank Q 80(1): 125-154.

Law MR, Wald NJ, Rudnicka AR. (2003) Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: Systematic review and meta-analysis. BMJ 326(7404): 1423.

le Chevalier T. (1996) Chemotherapy for advanced NSCLC. will meta-analysis provide the answer? Chest 109(5 Suppl): 107S-109S.

le Chevalier T, Pujol JL, Douillard JY, Alberola V, Monnier A, et al. (1994) A three-arm trial of vinorelbine (navelbine) plus cisplatin, vindesine plus cisplatin, and single-agent vinorelbine in the treatment of non-small cell lung cancer: An expanded analysis. Semin Oncol 21(5 Suppl 10): 28-33; discussion 33-4.

Leber P. (2000) The use of placebo control groups in the assessment of psychiatric drugs: An historical context. Biol Psychiatry 47(8): 699-706.

Leber PD. (1989) Hazards of inference: The active control investigation. Epilepsia 30 Suppl 1: S57-63; discussion S64-8.

Lee KJ, Thompson SG. (2008) Flexible parametric models for random-effects distributions. Stat Med 27(3): 418-434.

Lee WL, Bausell RB, Berman BM. (2001) The growth of health-related meta-analyses published from 1980 to 2000. Eval Health Prof 24(3): 327-335.

Lewis EJ, Hunsicker LG, Clarke WR, Berl T, Pohl MA, et al. (2001) Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. N Engl J Med 345(12): 851-860.

Lexchin J, Bero LA, Djulbegovic B, Clark O. (2003) Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. BMJ 326(7400): 1167-1170.

Lilford RJ. (2001) Equipoise is not synonymous with uncertainity. BMJ 323(7312): 574.

Lilford RJ, Braunholtz D. (1996) The statistical basis of public policy: A paradigm shift is overdue. BMJ 313(7057): 603-607.

Lilford RJ, Jackson J. (1995) Equipoise and the ethics of randomization. J R Soc Med 88(10): 552-559.

Lindbaek M, Hjortdahl P. (1999) How do two meta-analyses of similar data reach opposite conclusions? BMJ 318(7187): 873-874.

Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, et al. (1999) Impact of study quality on outcome in placebo-controlled trials of homeopathy. J Clin Epidemiol 52(7): 631-636.

LIPID. (1998) Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. the long-term intervention with pravastatin in ischaemic disease (LIPID) study group. N Engl J Med 339(19): 1349-1357.

Little P, Everitt H, Williamson I, Warner G, Moore M, et al. (2001) Observational study of effect of patient centredness and positive approach on outcomes of general practice consultations. BMJ 323(7318): 908-911.

Lochner HV, Bhandari M, Tornetta P,3rd. (2001) Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. J Bone Joint Surg Am 83-A(11): 1650-1655.

Loke YK, Derry S, Pritchard-Copley A. (2002) Appetite suppressants and valvular heart disease - a systematic review. BMC Clin Pharmacol 2: 6.

London AJ, Kadane JB. (2002) Placebos that harm: Sham surgery controls in clinical trials. Stat Methods Med Res 11(5): 413-427.

Lyman GH, Kuderer NM. (2005) The strengths and limitations of meta-analyses based on aggregate data. BMC Med Res Methodol 5(1): 14.

Macaskill P, Walter SD, Irwig L. (2001) A comparison of methods to detect publication-Bias in meta-analysis. Stat Med 20(4): 641-654.

Macedo A, Farre M, Banos JE. (2003) Placebo effect and placebos: What are we talking about? some conceptual and historical considerations. Eur J Clin Pharmacol 59(4): 337-342.

Mackay HC, Barkham M, Rees A, Stiles WB. (2003) Appraisal of published reviews of research on psychotherapy and counseling with adults 1990-1998. J Consult Clin Psychol 71(4): 652-656.

MacLean CH, Morton SC, Ofman JJ, Roth EA, Shekelle PG, et al. (2003) How useful are unpublished data from the food and drug administration in meta-analysis? J Clin Epidemiol 56(1): 44-51.

Madhok V, Fahey T. (2005) N-of-1 trials: An opportunity to tailor treatment in individual patients. Br J Gen Pract 55(512): 172.

Mahaffey KW, Roe MT, Dyke CK, Newby LK, Kleiman NS, et al. (2002) Misreporting of myocardial infarction end points: Results of adjudication by a central clinical events committee in the PARAGON-B trial. second platelet IIb/IIIa antagonist for the reduction of acute coronary syndrome events in a global organization network trial. Am Heart J 143(2): 242-248.

Mahaffey KW, Harrington RA, Akkerhuis M, Kleiman NS, Berdan LG, et al. (2001) Systematic adjudication of myocardial infarction end-points in an international clinical trial. Curr Control Trials Cardiovasc Med 2(4): 180-186.

Mahon J, Laupacis A, Donner A, Wood T. (1996) Randomised study of n of 1 trials versus standard practice. BMJ 312(7038): 1069-1074.

Makambi KH. (2004) The effect of the heterogeneity variance estimator on some tests of treatment efficacy. J Biopharm Stat 14(2): 439-449.

Mallett S, Clarke M. (2003) How many cochrane reviews are needed to cover existing evidence on the effects of health care interventions? ACP J Club 139(1): A11.

Mamdani MM, Tu JV. (2001) Did the major clinical trials of statins affect prescribing behaviour? CMAJ 164(12): 1695-1696.

Manheimer E, Ezzo J, Hadhazy V, Berman B. (2006) Published reports of acupuncture trials showed important limitations. J Clin Epidemiol 59(2): 107-113.

Manolio TA, Pearson TA, Wenger NK, Barrett-Connor E, Payne GH, et al. (1992) Cholesterol and heart disease in older persons and women. review of an NHLBI workshop. Ann Epidemiol 2(1-2): 161-176.

Man-Son-Hing M, Gage BF, Montgomery AA, Howitt A, Thomson R, et al. (2005) Preference-based antithrombotic therapy in atrial fibrillation: Implications for clinical decision making. Med Decis Making 25(5): 548-559.

Man-Son-Hing M, Laupacis A, O'Rourke K, Molnar FJ, Mahon J, et al. (2002) Determination of the clinical importance of study results. J Gen Intern Med 17(6): 469-476.

Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, et al. (2007) The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: An unblinded randomised controlled trial. Lancet 369(9566): 1016-1026.

Mathew T, Nordstrom K. (1999) On the equivalence of meta-analysis using literature and using individual patient data. Biometrics 55(4): 1221-1223.

Matt GE, Navarro AM. (1997) What meta-analyses have and have not taught us about psychotherapy effects: A review and future directions. Clin Psychol Rev 17(1): 1-32.

Matthews JN, Altman DG. (1996) Interaction 3: How to examine heterogeneity. BMJ 313(7061): 862.

Mazumdar M, Glassman JR. (2000) Categorizing a prognostic variable: Review of methods, code for easy implementation and applications to decision-making about cancer treatments. Stat Med 19(1): 113-132.

McAlindon TE, LaValley MP, Gulin JP, Felson DT. (2000) Glucosamine and chondroitin for treatment of osteoarthritis: A systematic quality assessment and meta-analysis. JAMA 283(11): 1469-1475.

McAlister FA, Clark HD, van Walraven C, Straus SE, Lawson FM, et al. (1999) The medical review article revisited: Has the science improved? Ann Intern Med 131(12): 947-951.

McAlister FA, Teo KK, Taher M, Montague TJ, Humen D, et al. (1999) Insights into the contemporary epidemiology and outpatient management of congestive heart failure. Am Heart J 138(1 Pt 1): 87-94.

McAlister FA, Clark HD, Wells PS, Laupacis A. (1998) Perioperative allogeneic blood transfusion does not cause adverse sequelae in patients with cancer: A meta-analysis of unconfounded studies. Br J Surg 85(2): 171-178.

McCormack K, Grant A, Scott N, EU Hernia Trialists Collaboration. (2004) Value of updating a systematic review in surgery using individual patient data. Br J Surg 91(4): 495-499.

McIntosh MW. (1996) The population risk as an explanatory variable in research synthesis of clinical trials. Stat Med 15(16): 1713-1728.

McLean AJ, Le Couteur DG. (2004) Aging biology and geriatric clinical pharmacology. Pharmacol Rev 56(2): 163-184.

McMahon AD. (2002) Study control, violators, inclusion criteria and defining explanatory and pragmatic trials. Stat Med 21(10): 1365-1376.

Meissner K, Distel H, Mitzdorf U. (2007) Evidence for placebo effects on physical but not on biochemical outcome parameters: A review of clinical trials. BMC Med 5: 3.

Menzies D, Pai M, Comstock G. (2007) Meta-analysis: New tests for the diagnosis of latent tuberculosis infection: Areas of uncertainty and recommendations for research. Ann Intern Med 146(5): 340-354.

Michiels S, Piedbois P, Burdett S, Syz N, Stewart L, et al. (2005) Meta-analysis when only the median survival times are known: A comparison with individual patient data results. Int J Technol Assess Health Care 21(1): 119-125.

Mignini LE, Khan KS. (2006) Methodological quality of systematic reviews of animal studies: A survey of reviews of basic research. BMC Med Res Methodol 6: 10.

Miller FG, Brody H. (2002) What makes placebo-controlled trials unethical? Am J Bioeth 2(2): 3-9.

Mills E, Loke YK, Wu P, Montori VM, Perri D, et al. (2004) Determining the reporting quality of RCTs in clinical pharmacology. Br J Clin Pharmacol 58(1): 61-65.

Minelli C, Abrams KR, Sutton AJ, Cooper NJ. (2004) Benefits and harms associated with hormone replacement therapy: Clinical decision analysis. BMJ 328(7436): 371.

Mittlbock M, Heinzl H. (2006) A simulation study comparing properties of heterogeneity measures in meta-analyses. Stat Med 25(24): 4321-4333.

Moerman DE, Jonas WB. (2002) Deconstructing the placebo effect and finding the meaning response. Ann Intern Med 136(6): 471-476.

Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. (2007) Epidemiology and reporting characteristics of systematic reviews. PLoS Med 4(3): e78.

Moher D, Tsertsvadze A. (2006) Systematic reviews: When is an update an update? Lancet 367(9514): 881-883.

Moher D, Pham B, Lawson ML, Klassen TP. (2003) The inclusion of reports of randomised trials published in languages other than english in systematic reviews. Health Technol Assess 7(41): 1-90.

Moher D, Schulz KF, Altman DG. (2001) The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357(9263): 1191-1194.

Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, et al. (1999a) Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. quality of reporting of meta-analyses. Lancet 354(9193): 1896-1900.

Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, et al. (1999b) Assessing the quality of reports of randomised trials: Implications for the conduct of meta-analyses. Health Technol Assess 3(12): i-iv, 1-98.

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, et al. (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352(9128): 609-613.

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, et al. (1995) Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials 16(1): 62-73.

Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, et al. (2005) Assessment of methodological quality of primary studies by systematic reviews: Results of the metaquality cross sectional study. BMJ 330(7499): 1053.

Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, et al. (2002) Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Control Clin Trials 23(6): 607-625.

Montenegro R, Needleman I, Moles D, Tonetti M. (2002) Quality of RCTs in periodontology--a systematic review. J Dent Res 81(12): 866-870.

Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, et al. (2005b) Randomized trials stopped early for benefit: A systematic review. JAMA 294(17): 2203-2209.

Montori VM, Wilczynski NL, Morgan D, Haynes RB, Hedges Team. (2005a) Optimal search strategies for retrieving systematic reviews from medline: Analytical survey. BMJ 330(7482): 68.

Montori VM, Wilczynski NL, Morgan D, Haynes RB, Hedges Team. (2003) Systematic reviews: A cross-sectional study of location and citation counts. BMC Med 1: 2.

Moore RA, Derry S, Phillips CJ, McQuay HJ. (2006) Nonsteroidal anti-inflammatory drugs (NSAIDs), cyxlooxygenase-2 selective inhibitors (coxibs) and gastrointestinal harm: Review of clinical trials and clinical practice. BMC Musculoskelet Disord 7: 79.

Morgenstern H. (1982) Uses of ecologic analysis in epidemiologic research. Am J Public Health 72(12): 1336-1344.

Morris CN. (1983) Parametric empirical bayes inference: Theory and applications. J Am Statist Assoc 78: 47-65.

Moye LA, Tita AT. (2002) Defending the rationale for the two-tailed test in clinical research. Circulation 105(25): 3062-3065.

MRC. (2000) A framework for the development and evaluation of randomised controlled trials for complex interventions to improve health. medical research council (MRC). London: MRC

MRFIT. (1982) Multiple risk factor intervention trial. risk factor changes and mortality results. multiple risk factor intervention trial research group (MRFIT). JAMA 248(12): 1465-1477.

Mueller PS, Montori VM, Bassler D, Koenig BA, Guyatt GH. (2007) Ethical issues in stopping randomized trials early because of apparent benefit. Ann Intern Med 146(12): 878-881.

Muhlhauser I, Meyer G. (2006) Surrogate end point fallacies -- the urge for randomized trials with clinical endpoints. Psychother Psychosom Med Psychol 56(5): 193-201.

Mukhtar AM, Timm J. (März 2006) Berücksichtigung methodischer qualität bei der untersuchung von klinischer heterogenität in meta-analyse Gemeinsame Jahrestagung Der Deutschen Region Der Internationalen Biometrischen Gesellschaft Und Des Deutschen Netzwerks Evidenz-Basierte Medizin Bochum

Mulrow CD. (1994) Rationale for systematic reviews. BMJ 309(6954): 597-599.

Mulrow CD. (1987) The medical review article: State of the science. Ann Intern Med 106(3): 485-488.

Murchie P, Hannaford PC, Wyke S, Nicolson MC, Campbell NC. (2007) Designing an integrated follow-up programme for people treated for cutaneous malignant melanoma: A practical application of the MRC framework for the design and evaluation of complex interventions to improve health. Fam Pract 24(3): 283-292.

Murphy AW, Esterman A, Pilotto LS. (2006) Cluster randomized controlled trials in primary care: An introduction. Eur J Gen Pract 12(2): 70-73.

Nakagawa K, Ishizaki T. (2000) Therapeutic relevance of pharmacogenetic factors in cardiovascular medicine. Pharmacol Ther 86(1): 1-28.

National Research Council. (1992) Combining information: Statistical issues and opportunities for research. Washington, DC: National Academic Press.

Naylor CD, Llewellyn-Thomas HA. (1994) Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? J Clin Epidemiol 47(7): 787-795.

Nelder JA. (1998) The selection of terms in response-surface models - how strong is the weak heredity principle? The American Statistician 52: 315-318.

Ni Mhurchu C, Dunshea-Mooij CA, Bennett D, Rodgers A. (2005) Chitosan for overweight or obesity. Cochrane Database Syst Rev (3)(3): CD003892.

Nowak A, Findlay M, Culjak G, Stockler M. (2004) Tamoxifen for hepatocellular carcinoma. Cochrane Database Syst Rev (3)(3): CD001024.

O'Rourke K. (2006) An historical perspective on meta-analysis: Dealing quantitatively with varying study results. The James Lind Library (www.jameslindlibrary.org)(Zugang am 02.02.07)

Oakley GP,Jr, Johnston RB,Jr. (2004) Balancing benefits and harms in public health prevention programmes mandated by governments. BMJ 329(7456): 41-3; discussion 43-4.

Ohlssen DI, Sharples LD, Spiegelhalter DJ. (2007) Flexible random-effects models using bayesian semi-parametric models: Applications to institutional comparisons. Stat Med 26(9): 2088-2112.

Olade RA. (2004) Evidence-based practice and research utilization activities among rural nurses. J Nurs Scholarsh 36(3): 220-225.

Olkin I, Sampson A. (1998) Comparison of meta-analysis versus analysis of variance of individual patient data. Biometrics 54(1): 317-322.

Olschewski M, Scheurlen H. (1985) Comprehensive cohort study: An alternative to randomized consent design in a breast preservation trial. Methods Inf Med 24(3): 131-134.

Olshansky B. (2007) Placebo and nocebo in cardiovascular health: Implications for healthcare, research, and the doctor-patient relationship. J Am Coll Cardiol 49(4): 415-421.

Ortiz Z, Shea B, Suarez-Almazor ME, Moher D, Wells GA, et al. (1998) The efficacy of folic acid and folinic acid in reducing methotrexate gastrointestinal toxicity in rheumatoid arthritis. A metaanalysis of randomized controlled trials. J Rheumatol 25(1): 36-43.

Otto MW, Nierenberg AA. (2002) Assay sensitivity, failed clinical trials, and the conduct of science. Psychother Psychosom 71(5): 241-243.

Overall JE, Doyle SR. (1994) Estimating sample sizes for repeated measurement designs. Control Clin Trials 15(2): 100-123.

Owens DK, Shachter RD, Nease RF,Jr. (1997) Representation and analysis of medical decision problems with influence diagrams. Med Decis Making 17(3): 241-262.

Oxman AD, Lavis JN, Fretheim A. (2007) Use of evidence in WHO recommendations. Lancet 369(9576): 1883-1889.

Oxman AD, Guyatt GH. (1992) A consumer's guide to subgroup analyses. Ann Intern Med 116(1): 78-84.

Oxman AD, Guyatt GH. (1991) Validation of an index of the quality of review articles. J Clin Epidemiol 44(11): 1271-1278.

Oxman AD, Guyatt GH. (1988) Guidelines for reading literature reviews. CMAJ 138(8): 697-703.

Panpanich R, Lerttrakarnnon P, Laopaiboon M. (2004) Azithromycin for acute lower respiratory tract infections. Cochrane Database Syst Rev (4)(4): CD001954.

Parker AB, Naylor CD. (2000) Subgroups, treatment effects, and baseline risks: Some lessons from major cardiovascular trials. Am Heart J 139(6): 952-961.

Parmigiani G. (2002) Modeling in medical decision making. Chichester: Wiley.

Parving HH, Lehnert H, Brochner-Mortensen J, Gomis R, Andersen S, et al. (2001) The effect of irbesartan on the development of diabetic nephropathy in patients with type 2 diabetes. N Engl J Med 345(12): 870-878.

Patsopoulos NA, Analatos AA, Ioannidis JP. (2005) Relative citation impact of various study designs in the health sciences. JAMA 293(19): 2362-2366.

Paul SR, Donner A. (1992) Small sample performance of tests of homogeneity of odds ratios in K 2 x 2 tables. Stat Med 11(2): 159-165.

Paule RC, Mandel J. (1982) Consensus values and weighting factors. J Res Natl Bur Stand 87(5): 377-385.

Perel P, Roberts I, Sena E, Wheble P, Briscoe C, et al. (2007) Comparison of treatment effects between animal experiments and clinical trials: Systematic review. BMJ 334(7586): 197.

Petersen JL, Haque G, Hellkamp AS, Flaker GC, Mark Estes NA,3rd, et al. (2006) Comparing classifications of death in the mode selection trial: Agreement and disagreement among site investigators and a clinical events committee. Contemp Clin Trials 27(3): 260-268.

Peters-Klimm F, Muller-Tasch T, Schellberg D, Gensichen J, Muth C, et al. (2007) Rationale, design and conduct of a randomised controlled trial evaluating a primary care-based complex intervention to improve the quality of life of heart failure patients: HICMan (heidelberg integrated case management). BMC Cardiovasc Disord 7: 25.

Petitti DB. (1994) Meta-analysis, decision analysis and costeffectiveness analysis. New York: Oxford University Press

Petitti DB. (2001) Approaches to heterogeneity in meta-analysis. Stat Med 20(23): 3625-3633.

Peto R. (1987) Why do we need systematic overviews of randomized trials? Stat Med 6(3): 233-244.

Petticrew M. (2003) Why certain systematic reviews reach uncertain conclusions. BMJ 326(7392): 756-758.

Petticrew M. (2001) Systematic reviews from astronomy to zoology: Myths and misconceptions. BMJ 322(7278): 98-101.

Petticrew M, Song F, Wilson P, Wright K. (1999) Quality-assessed reviews of health care interventions and the database of abstracts of reviews of effectiveness (DARE). NHS CRD review, dissemination, and information teams. Int J Technol Assess Health Care 15(4): 671-678.

Petticrew M, Kennedy SC. (1997) Detecting the effects of thromboprophylaxis: The case of the rogue reviews. BMJ 315(7109): 665-668.

Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, et al. (2004) Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 8(36): iii-iv, ix-xi, 1-158.

Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, et al. (2006) Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. JAMA 295(10): 1152-1160.

Pignon JP, Courtial F, Syz N. (October 2001) Difficulties in searching meta-analyses based on individual patient data on potentially curative treatment in oncology. 9th Cochrane Colloquium Lyon

Pignon JP, Arriagada R. (1993) Meta-analysis. Lancet 341(8850): 964-965.

Pignone M, Phillips C, Mulrow C. (2000) Use of lipid lowering drugs for primary prevention of coronary heart disease: Meta-analysis of randomised trials. BMJ 321(7267): 983-986.

Pildal J, Chan AW, Hrobjartsson A, Forfang E, Altman D, et al. (October 2004b) How often does unclear allocation concealment in randomised trials reflect inadequate reporting of adequate methods? 12th Cochrane Colloquium Ottawa

Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, et al. (2007) Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol 36(4): 847-857.

Pitt B, Mancini GB, Ellis SG, Rosman HS, Park JS, et al. (1995) Pravastatin limitation of atherosclerosis in the coronary arteries (PLAC I): Reduction in atherosclerosis progression and clinical events. PLAC I investigation. J Am Coll Cardiol 26(5): 1133-1139.

Plint AC, Moher D, Morrison A, Schulz K, Altman DG, et al. (2006) Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust 185(5): 263-267.

Pocock S, Wang D, Wilhelmsen L, Hennekens CH. (2005) The data monitoring experience in the candesartan in heart failure assessment of reduction in mortality and morbidity (CHARM) program. Am Heart J 149(5): 939-943.

Pocock SJ, Assmann SE, Enos LE, Kasten LE. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. Stat Med 21(19): 2917-2930.

Poeze M, Greve JW, Ramsay G. (2005) Meta-analysis of hemodynamic optimization: Relationship to methodological quality. Crit Care 9(6): R771-9.

Poole C, Greenland S. (1999) Random-effects meta-analyses are not always conservative. Am J Epidemiol 150(5): 469-475.

Poses RM, McClish DK, Smith WR, Chaput de Saintonge DM, Huber EC, et al. (1997) Physicians' judgments of the risks of cardiac procedures. differences between cardiologists and other internists. Med Care 35(6): 603-617.

POST-CABG. (1997) The effect of aggressive lowering of low-density lipoprotein cholesterol levels and low-dose anticoagulation on obstructive changes in saphenous-vein coronary-artery bypass grafts. the post coronary artery bypass graft trial investigators. N Engl J Med 336(3): 153-162.

Potter J, Langhorne P, Roberts M. (1998) Routine protein energy supplementation in adults: Systematic review. BMJ 317(7157): 495-501.

Potts M, Prata N, Walsh J, Grossman A. (2006) Parachute approach to evidence based medicine. BMJ 333(7570): 701-703.

Prins JM, Buller HR. (1996) Meta-analysis: The final answer, or even more confusion? Lancet 348(9021): 199.

Pusic A, Liu JC, Chen CM, Cano S, Davidge K, et al. (2007) A systematic review of patient-reported outcome measures in head and neck cancer surgery. Otolaryngol Head Neck Surg 136(4): 525-535.

Rambaldi A, Jacobs BP, Iaquinto G, Gluud C. (2005) Milk thistle for alcoholic and/or hepatitis B or C virus liver diseases. Cochrane Database Syst Rev (2)(2): CD003620.

Raudenbush SW, Bryk AS. (1985) Empirical bayes metaanalysis. J Educ Statist 10(2): 75-98.

Riley RD, Simmonds MC, Look MP. (2007) Evidence synthesis combining individual patient data and aggregate data: A systematic review identified current practice and possible methods. J Clin Epidemiol 60(5): 431-439.

Rochon PA, Gurwitz JH, Simms RW, Fortin PR, Felson DT, et al. (1994) A study of manufacturer-supported trials of nonsteroidal anti-inflammatory drugs in the treatment of arthritis. Arch Intern Med 154(2): 157-163.

Roderick P, Ferris G, Wilson K, Halls H, Jackson D, et al. (2005) Towards evidence-based guidelines for the prevention of venous thromboembolism: Systematic reviews of mechanical methods, oral anticoagulation, dextran and regional anaesthesia as thromboprophylaxis. Health Technol Assess 9(49): iii-iv, ix-x, 1-78.

Rosser WW. (1999) Application of evidence from randomised controlled trials to general practice. Lancet 353(9153): 661-664.

Rothstein HR, Sutton AJ, Borenstein M, editors. (2005) Publication-Bias in meta-analysis. prevention, assessment and adjustments. Sussex: John Wiley and Sons

Rothwell PM. (2005) Treating individuals 2. subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. Lancet 365(9454): 176-186.

Royle P, Waugh N. (2004) Should systematic reviews include searches for published errata? Health Info Libr J 21(1): 14-20.

Rucker G. (1989) A two-stage trial design for testing treatment, self-selection and treatment preference effects. Stat Med 8(4): 477-485.

Ruiz-Canela M, Martinez-Gonzalez MA, de Irala-Estevez J. (2000) Intention to treat analysis is related to methodological quality. BMJ 320(7240): 1007-1008.

Rukhin AL, Biggerstaff BJ, Vangel MG. (2000) Restricted maximum likelihood estimation of common mean and the Mandel–Paule algorithm. J Stat Plan Inference 83(2): 319-330.

Sackett DL. (2001) Uncertainty about clinical equipoise. there is another exchange on equipoise and uncertainty. BMJ 322(7289): 795-796.

Sacks FM, Pfeffer MA, Moye LA, Rouleau JL, Rutherford JD, et al. (1996) The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. cholesterol and recurrent events trial investigators. N Engl J Med 335(14): 1001-1009.

Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. (1987) Meta-analyses of randomized controlled trials. N Engl J Med 316(8): 450-455.

Safer DJ. (2002) Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. J Nerv Ment Dis 190(9): 583-592.

Saillourglenisson F, Chene G, Salmi LR, Hafner R, Salamon R. (2000) Effect of dapsone on survival in HIV infected patients: A meta- analysis of finished trials. Rev Epidemiol Sante Publique 48(1): 17-30.

Sander L, Kitcher H. (2006) Systematic and other reviews: Terms and definitions used by UK organisations and selected databases. systematic review and delphi survey. National Institute for Health and Clinical Excellence

Sanders S, Del Mar C. (2005) Clever searching for evidence. BMJ 330(7501): 1162-1163.

Scheinfeld N. (2003) Getting started with evidence-based research. Dermatol Surg 29(6): 572.

Scherer RW, Langenberg P, von Elm E. (2007) Full publication of results initially presented in abstracts. Cochrane Database Syst Rev (2)(2): MR000005.

Schmid CH. (1999) Exploring heterogeneity in randomized trials via meta-analysis. Drug Inf J 33(211): 224.

Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. (2004) Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol 57(7): 683-697.

Schmid CH, Landa M, Jafar TH, Giatras I, Karim T, et al. (2003) Constructing a database of individual clinical trials for longitudinal analysis. Control Clin Trials 24(3): 324-340.

Schmid CH, Lau J, McIntosh MW, Cappelleri JC. (1998) An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med 17(17): 1923-1942.

Schmidt FL, Oh IS, Hayes TL. (2007) Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. Br J Math Stat Psychol

Schneider B. (1989) Analysis of clinical trial outcomes: Alternative approaches to subgroup analysis. Control Clin Trials 10(4 Suppl): 176S-186S.

Schroeder K, Fahey T. (2004) Over-the-counter medications for acute cough in children and adults in ambulatory settings. Cochrane Database Syst Rev (4)(4): CD001831.

Schulman JL, Nelson SJ. (2005) Systematic reviews and medical vocabulary: Rapid responses to: Montori,V.M.; wilczynski, N.L.; morgan, D.; haynes, R.B.; hedges team. BMJ 330(7482): 68.

Schulz KF, Grimes DA. (2005) Multiplicity in randomised trials II: Subgroup and interim analyses. Lancet 365(9471): 1657-1661.

Schulz KF, Chalmers I, Altman DG. (2002c) The landscape and lexicon of blinding in randomized trials. Ann Intern Med 136(3): 254-259.

Schulz KF, Grimes DA. (2002b) Allocation concealment in randomised trials: Defending against deciphering. Lancet 359(9306): 614-618.

Schulz KF, Grimes DA. (2002a) Generation of allocation sequences in randomised trials: Chance, not choice. Lancet 359(9305): 515-519.

Schulz KF, Grimes DA, Altman DG, Hayes RJ. (1996) Blinding and exclusions after allocation in randomised controlled trials: Survey of published parallel group trials in obstetrics and gynaecology. BMJ 312(7033): 742-744.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. (1995) Empirical evidence of-Bias. dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273(5): 408-412.

Schulz KF, Chalmers I, Grimes DA, Altman DG. (1994b) Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 272(2): 125-128.

Schunemann HJ, Oxman AD, Fretheim A. (2006) Improving the use of research evidence in guideline development: 6. determining which outcomes are important. Health Res Policy Syst 4: 18.

Schunemann HJ, Guyatt GH. (2005) Commentary--goodbye M(C)ID! hello MID, where do you come from? Health Serv Res 40(2): 593-597.

Schwarzer G, Antes G, Schumacher M. (2002) Inflation of type I error rate in two statistical tests for the detection of publication-Bias in meta-analyses with binary outcomes. Stat Med 21(17): 2465-2477.

Senn S. (1994) Regression to the mean and crossover trials revisited. Stat Med 13(11): 1181-1186.

Serruys PW, de Feyter P, Macaya C, Kokott N, Puel J, et al. (2002) Fluvastatin for prevention of cardiac events following successful first percutaneous coronary intervention: A randomized controlled trial. JAMA 287(24): 3215-3222.

Sever PS, Dahlof B, Poulter NR, Wedel H, Beevers G, et al. (2003) Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the anglo-scandinavian cardiac outcomes trial--lipid lowering arm (ASCOT-LLA): A multicentre randomised controlled trial. Lancet 361(9364): 1149-1158.

Shaffer ML, Watterberg KL. (2006) Joint distribution approaches to simultaneously quantifying benefit and risk. BMC Med Res Methodol 6: 48.

Shang A, Huwiler-Muntener K, Nartey L, Juni P, Dorig S, et al. (2005) Are the clinical effects of homoeopathy placebo effects? comparative study of placebo-controlled trials of homoeopathy and allopathy. Lancet 366(9487): 726-732.

Shea B, Dubé C, Moher D. (Systematic Reviews in Health Care: Meta-analysis in context. London: BMJ books; 2001. pp. 122–139.) Assessing the quality of reports of systematic reviews: The QUOROM statement compared to other tools. In: Egger, M ; Smith, G D ; Altman, D G ; Editor

Shea B, Bouter LM, Grimshaw JM, Francis D, Ortiz Z, et al. (2006) Scope for improvement in the quality of reporting of systematic reviews. from the cochrane musculoskeletal group. J Rheumatol 33(1): 9-15.

Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, et al. (2007) Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 7: 10.

Shekelle PG, Ortiz E, Rhodes S, Morton SC, Eccles MP, et al. (2001) Validity of the agency for healthcare research and quality clinical practice guidelines: How quickly do guidelines become outdated? JAMA 286(12): 1461-1467.

Shepherd J, Blauw GJ, Murphy MB, Bollen EL, Buckley BM, et al. (2002) Pravastatin in elderly individuals at risk of vascular disease (PROSPER): A randomised controlled trial. Lancet 360(9346): 1623-1630.

Shepherd J, Cobbe SM, Ford I, Isles CG, Lorimer AR, et al. (1995) Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. west of scotland coronary prevention study group. N Engl J Med 333(20): 1301-1307.

Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, et al. (2007) How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 147(4): 224-233.

Shojania KG, Bero LA. (2001) Taking advantage of the explosion of systematic reviews: An efficient MEDLINE search strategy. Eff Clin Pract 4(4): 157-162.

Sidik K, Jonkman JN. (2005a) Simple heterogeneity variance estimation for metaanalysis. J Royal Stat Soc Series C: Appl Stat 54: 367-384.

Sidik K, Jonkman JN. (2007) A comparison of heterogeneity variance estimators in combining results of studies. Stat Med 26(9): 1964-1981.

Sidik K, Jonkman JN. (2005b) A note on variance estimation in random effects meta-regression. J Biopharm Stat 15(5): 823-838.

Sidik K, Jonkman JN. (2002) A simple confidence interval for meta-analysis. Stat Med 21(21): 3153-3159.

Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, et al. (2006) Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. Stat Med Published online 20 November 2006 in Wiley InterScience (DOI: 10.1002/sim.2752)

Siest G, Jeannesson E, Visvikis-Siest S. (2007) Enzymes and pharmacogenetics of cardiovascular drugs. Clin Chim Acta 381(1): 26-31.

SIGN. (2008) SIGN 50: A guideline developers' handbook. scottish intercollegiate guideline network (SIGN). http://www.sign.ac.uk/pdf/sign50.pdf (Zugang am 11. Februar 2008)

Silagy CA, Middleton P, Hopewell S. (2002) Publishing protocols of systematic reviews: Comparing what was done to what was planned. JAMA 287(21): 2831-2834.

Silva Filho CR, Saconato H, Conterno LO, Marques I, Atallah AN. (2005) Assessment of clinical trial quality and its impact on meta-analyses. Rev Saude Publica 39(6): 865-873.

Simes RJ. (1986) Publication-Bias: The case for an international registry of clinical trials. J Clin Oncol 4(10): 1529-1541.

Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, et al. (2005) Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. Clin Trials 2(3): 209-217.

Siragusa S, Cosmi B, Piovella F, Hirsh J, Ginsberg JS. (1996) Low-molecular-weight heparins and unfractionated heparin in the treatment of patients with acute venous thromboembolism: Results of a meta-analysis. Am J Med 100(3): 269-277.

Smith CL, Stein GE. (2002) Viral load as a surrogate end point in HIV disease. Ann Pharmacother 36(2): 280-287.

Smith CT, Williamson PR, Marson AG. (2005) An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes. J Eval Clin Pract 11(5): 468-478.

Smith GC, Pell JP. (2003) Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. BMJ 327(7429): 1459-1461.

Soares HP, Daniels S, Kumar A, Clarke M, Scott C, et al. (2004) Bad reporting does not mean bad methods for randomised trials: Observational study of randomised controlled trials performed by the radiation therapy oncology group. BMJ 328(7430): 22-24.

Song F, Sheldon TA, Sutton AJ, Abrams KR, Jones DR. (2001) Methods for exploring heterogeneity in meta-analysis. Eval Health Prof 24(2): 126-151.

Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. (2000) Publication and related-Biases. Health Technol Assess 4(10): 1-115.

Song F, Freemantle N, Sheldon TA, House A, Watson P, et al. (1993) Selective serotonin reuptake inhibitors: Meta-analysis of efficacy and acceptability. BMJ 306(6879): 683-687.

Song FJ, Fry-Smith A, Davenport C, Bayliss S, Adi Y, et al. (2004) Identification and assessment of ongoing trials in health technology assessment reviews. Health Technol Assess 8(44): iii, 1-87.

Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. (2000) Bayesian methods in health technology assessment: A review. Health Technol Assess 4(38): 1-130.

Spooner C, Rowe BH, Saunders LD, Milner RA. (October 1998) Nedocromil sodium as treatment of exercise-induced bronchoconstriction: A comparison of results from a metaanalysis with individual patient data. 6th Cochrane Colloquium Baltimore

Stamler J, Vaccaro O, Neaton JD, Wentworth D. (1993) Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the multiple risk factor intervention trial. Diabetes Care 16(2): 434-444.

Steinberg KK, Smith SJ, Stroup DF, Olkin I, Lee NC, et al. (1997) Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. Am J Epidemiol 145(10): 917-925.

Sterne JA, Egger M, Smith GD. (2001) Systematic reviews in health care: Investigating and dealing with publication and other-Biases in meta-analysis. BMJ 323(7304): 101-105.

Sterne JA, Gavaghan D, Egger M. (2000) Publication and related-Bias in meta-analysis: Power of statistical tests and prevalence in the literature. J Clin Epidemiol 53(11): 1119-1129.

Stewart LA, Tierney JF. (2002) To IPD or not to IPD? advantages and disadvantages of systematic reviews using individual patient data. Eval Health Prof 25(1): 76-97.

Stewart LA, Parmar MK. (1993) Meta-analysis of the literature or of individual patient data: Is there a difference? Lancet 341(8842): 418-422.

Straus SE. (2002) Individualizing treatment decisions. the likelihood of being helped or harmed. Eval Health Prof 25(2): 210-224.

Streiner DL, Joffe R. (1998) The adequacy of reporting randomized, controlled trials in the evaluation of antidepressants. Can J Psychiatry 43(10): 1026-1030.

Strippoli GF, Craig JC, Schena FP. (2004) The number, quality, and coverage of randomized controlled trials in nephrology. J Am Soc Nephrol 15(2): 411-419.

Sutton AJ, Higgins JP. (2008) Recent developments in meta-analysis. Stat Med 27(5): 625-650.

Sutton AJ, Kendrick D, Coupland CA. (2008) Meta-analysis of individual- and aggregate-level data. Stat Med 27(5): 651-669.

Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, et al. (2007) Evidence-based sample size calculations based upon updated meta-analysis. Stat Med 26(12): 2479-2500.

Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. (2000) Empirical assessment of effect of publication-Bias on meta-analyses. BMJ 320(7249): 1574-1577.

Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. (1998) Systematic reviews of trials and other studies. Health Technol Assess 2(19): 1-276.

Sweeney KG, MacAuley D, Gray DP. (1998) Personal significance: The third dimension. Lancet 351(9096): 134-136.

Swingler GH, Volmink J, Ioannidis JP. (2003) Number of published systematic reviews and global burden of disease: Database analysis. BMJ 327(7423): 1083-1084.

Switzer FSI, Paese PW, Drasgow F. (1992) Bootstrap estimates of standard errors in validity generalization. Journal of Applied Psychology 77(2): 123-129.

Szczech LA, Berlin JA, Feldman HI. (1998) The effect of antilymphocyte induction therapy on renal allograft survival. A meta-analysis of individual patient-level data. anti-lymphocyte antibody induction therapy study group. Ann Intern Med 128(10): 817-826.

Takkouche B, Cadarso-Suarez C, Spiegelman D. (1999) Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. Am J Epidemiol 150(2): 206-215.

Tanne JH. (2006) NEJM stands by its criticism of vioxx study. BMJ 332(7540): 505.

Teerenstra S, Melis RJ, Peer PG, Borm GF. (2006) Pseudo cluster randomization dealt with selection-Bias and contamination in clinical trials. J Clin Epidemiol 59(4): 381-386.

Temple R. (2002) Policy developments in regulatory approval. Stat Med 21(19): 2939-2948.

Temple R, Ellenberg SS. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: Ethical and scientific issues. Ann Intern Med 133(6): 455-463.

Teo KK, Yusuf S, Furberg CD. (1993) Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction. an overview of results from randomized controlled trials. JAMA 270(13): 1589-1595.

Teramukai S, Matsuyama Y, Mizuno S, Sakamoto J. (2004) Individual patient-level and study-level meta-analysis for investigating modifiers of treatment effect. Jpn J Clin Oncol 34(12): 717-721.

Terrin N, Schmid CH, Lau J. (2005) In an empirical evaluation of the funnel plot, researchers could not visually identify publication-Bias. J Clin Epidemiol 58(9): 894-901.

Terrin N, Schmid CH, Lau J, Olkin I. (2003) Adjusting for publication-Bias in the presence of heterogeneity. Stat Med 22(13): 2113-2126.

Thompson SG, Sharp SJ. (1999) Explaining heterogeneity in meta-analysis: A comparison of methods. Stat Med 18(20): 2693-2708.

Thompson SG, Smith TC, Sharp SJ. (1997) Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 16(23): 2741-2758.

Thompson SG, Pocock SJ. (1991) Can meta-analyses be trusted? Lancet 338(8775): 1127-1130.

Thornley B, Adams C. (1998) Content and quality of 2000 controlled trials in schizophrenia over 50 years. BMJ 317(7167): 1181-1184.

Thornton A, Lee P. (2000) Publication-Bias in meta-analysis: Its causes and consequences. J Clin Epidemiol 53(2): 207-216.

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. ; (): -. J R Stat Soc Ser B 58(1): 267-288.

Tierney JF, Stewart LA. (2005) Investigating patient exclusion-Bias in meta-analysis. Int J Epidemiol 34(1): 79-87.

Tierney JF, Clarke M, Stewart LA. (2000) Is there-Bias in the publication of individual patient data meta-analyses? Int J Technol Assess Health Care 16(2): 657-667.

Torgerson DJ. (2001) Contamination in trials: Is cluster randomisation the answer? BMJ 322(7282): 355-357.

Tramer MR, Reynolds DJ, Moore RA, McQuay HJ. (1997) Impact of covert duplicate publication on meta-analysis: A case study. BMJ 315(7109): 635-640.

Tu JV, Hannan EL, Anderson GM, Iron K, Wu K, et al. (1998) The fall and rise of carotid endarterectomy in the united states and canada. N Engl J Med 339(20): 1441-1447.

Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, et al. (2005) Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: The minimal clinically important improvement. Ann Rheum Dis 64(1): 29-33.

Tudur Smith C, Williamson PR. (2007) A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. Clin Trials 4(6): 621-630.

Tudur C, Williamson PR, Khan S, Best LY. (2001) The value of the aggregate data approach in meta-analysis with time-to-event outcomes. J R Stat Soc Ser A 164(2): 357-370.

Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. (2000) A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 19(24): 3417-3432.

Van den Poel, Dirk., Lariviere B. (2004) Attrition analysis for financial services using proportional hazard models. European Journal of Operational Research 157(1): 196-217.

Van Houwelingen HC, Zwinderman KH, Stijnen T. (1993) A bivariate approach to meta-analysis. Stat Med 12(24): 2273-2284.

van Nieuwenhoven CA, Buskens E, van Tiel FH, Bonten MJ. (2001) Relationship between methodological trial quality and the effects of selective digestive decontamination on pneumonia and mortality in critically ill patients. JAMA 286(3): 335-340.

van Walraven C, Mahon JL, Moher D, Bohm C, Laupacis A. (1999) Surveying physicians to determine the minimal important difference: Implications for sample-size calculation. J Clin Epidemiol 52(8): 717-723.

Verhagen AP, de Vet HC, Vermeer F, Widdershoven JW, de Bie RA, et al. (2002) The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 18(1): 11-23.

Verhagen AP, de Bie RA, Lenssen AF, de Vet HC, Kessels AG, et al. (2000) Impact of quality items on study outcome. treatments in acute lateral ankle sprains. Int J Technol Assess Health Care 16(4): 1136-1146.

Vickers A, Goyal N, Harland R, Rees R. (1998) Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials 19(2): 159-166.

Vickers AJ. (2003) How many repeated measures in repeated measures designs? statistical issues for comparative trials. BMC Med Res Methodol 3: 22.

Viechtbauer W. (2007b) Hypothesis tests for population heterogeneity in meta-analysis. Br J Math Stat Psychol 60(Pt 1): 29-60.

Viechtbauer W. (2007a) Confidence intervals for the amount of heterogeneity in meta-analysis. Stat Med 26(1): 37-52.

Villar J, Mackey ME, Carroli G, Donner A. (2001) Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: Comparison of fixed and random effects models. Stat Med 20(23): 3635-3647.

Villari P, Manzoli L, Boccia A. (2004) Methodological quality of studies and patient age as major sources of variation in efficacy estimates of influenza vaccination in healthy adults: A meta-analysis. Vaccine 22(25-26): 3475-3486.

Walker AE, McLeer SK, DAMOCLES Group. (2004) Small group processes relevant to data monitoring committees in controlled clinical trials: An overview of reviews. Clin Trials 1(3): 282-296.

Walsh JM, Pignone M. (2004) Drug treatment of hyperlipidemia in women. JAMA 291(18): 2243-2252.

Walter SD. (1997) Variation in baseline risk as an explanation of heterogeneity in meta-analysis. Stat Med 16(24): 2883-2900.

Walter SD, Cook DJ, Guyatt GH, King D, Troyan S. (1997) Outcome assessment for clinical trials: How many adjudicators do we need? canadian lung oncology group. Control Clin Trials 18(1): 27-42.

Wang CT, Lin J, Chang CJ, Lin YT, Hou SM. (2004) Therapeutic effects of hyaluronic acid on osteoarthritis of the knee. A meta-analysis of randomized controlled trials. J Bone Joint Surg Am 86-A(3): 538-545.

Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. (2007) Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med 357(21): 2189-2194.

Wang SJ, Hung HM, Tsong Y. (2002) Utility and pitfalls of some statistical methods in active controlled clinical trials. Control Clin Trials 23(1): 15-28.

Weber WW. (2001) The legacy of pharmacogenetics and potential applications. Mutat Res 479(1-2): 1-18.

Weed DL. (2000) Interpreting epidemiological evidence: How meta-analysis and causal inference methods are related. Int J Epidemiol 29(3): 387-390.

Wennberg JE, Barry MJ, Fowler FJ, Mulley A. (1993) Outcomes research, PORTs, and health care reform. Ann N Y Acad Sci 703: 52-62.

West SL, King V, Carey TS, Kathleen N, Lohr MD, et al. (2002) Systems to rate the strength of scientific evidence. Agency for Healthcare Research and Quality Evidence Report, Technology Assessment No. 47.(AHRQ Publication No. 02-E016. Rockville)

Whelan T, Sawka C, Levine M, Gafni A, Reyno L, et al. (2003) Helping patients make informed choices: A randomized trial of a decision aid for adjuvant chemotherapy in lymph node-negative breast cancer. J Natl Cancer Inst 95(8): 581-587.

Whittington CJ, Kendall T, Fonagy P, Cottrell D, Cotgrove A, et al. (2004) Selective serotonin reuptake inhibitors in childhood depression: Systematic review of published versus unpublished data. Lancet 363(9418): 1341-1345.

Wiebe N, Vandermeer B, Platt RW, Klassen TP, Moher D, et al. (2006) A systematic review identifies a lack of standardization in methods for handling missing variance data. J Clin Epidemiol 59(4): 342-353.

Wieland S, Dickersin K. (2005) Selective exposure reporting and medline indexing limited the search sensitivity for observational studies of the adverse effects of oral contraceptives. J Clin Epidemiol 58(6): 560-567.

Wilczynski NL, Haynes RB, Hedges Team. (2007) EMBASE search strategies achieved high sensitivity and specificity for retrieving methodologically sound systematic reviews. J Clin Epidemiol 60(1): 29-33.

Wilkes MM, Navickis RJ. (2001) Patient survival after human albumin administration. A meta-analysis of randomized, controlled trials. Ann Intern Med 135(3): 149-164.

Williams HC, Seed P. (1993) Inadequate size of 'negative' clinical trials in dermatology. Br J Dermatol 128(3): 317-326.

Williamson PR, Gamble C. (2007) Application and investigation of a bound for outcome reporting-Bias. Trials 8: 9.

Williamson PR, Marson AG, Tudur C, Hutton JL, Chadwick D. (2000) Individual patient data meta-analysis of randomized anti-epileptic drug monotherapy trials. J Eval Clin Pract 6(2): 205-214.

Winkelstein WJ. (1998) The first use of meta-analysis? Am J Epidemiol 147(8): 717.

Winkens B, Schouten HJ, van Breukelen GJ, Berger MP. (2005) Optimal time-points in clinical trials with linearly divergent treatment effects. Stat Med 24(24): 3743-3756.

Wittes J, Barrett-Connor E, Braunwald E, Chesney M, Cohen HJ, et al. (2007) Monitoring the randomized trials of the women's health initiative: The experience of the data and safety monitoring board. Clin Trials 4(3): 218-234.

Woods SW, Gueorguieva RV, Baker CB, Makuch RW. (2005) Control group-Bias in randomized atypical antipsychotic medication trials for schizophrenia. Arch Gen Psychiatry 62(9): 961-970.

Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, et al. (2007) Triangulating patient and clinician perspectives on clinically important differences in health-related quality of life among patients with heart disease. Health Serv Res 42(6 Pt 1): 2257-74; discussion 2294-323.

Yank V, Rennie D, Bero LA. (September 2005) Are authors' financial ties with pharmaceutical companies associated with positive results or conclusions in meta-analyses on antihypertensive medications? 5th International Congress on Peer Review and Biomedical Publication Chicago

Yong WP, Innocenti F, Ratain MJ. (2006) The role of pharmacogenetics in cancer therapeutics. Br J Clin Pharmacol 62(1): 35-46.

Yusuf S, Wittes J, Probstfield J, Tyroler HA. (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 266(1): 93-98.

Yusuf S, Peto R, Lewis J, Collins R, Sleight P. (1985) Beta blockade during and after myocardial infarction: An overview of the randomized trials. Prog Cardiovasc Dis 27(5): 335-371.

Zeger SL, Liang KY. (1992) An overview of methods for the analysis of longitudinal data. Stat Med 11(14-15): 1825-1839.

Zelen M. (1979) A new design for randomized clinical trials. N Engl J Med 300(22): 1242-1245.

Anhang 1: Aus der SR in Abschnitt 2 ausgeschlossene Studien

Aker PD, Gross AR, Goldsmith CH, Peloso P. (1996) Conservative management of mechanical neck pain: Systematic overview and meta-analysis. BMJ 313(7068): 1291-1296.

Als-Nielsen B, Gluud LL, Gluud C. (October 2004b) Methodological quality and treatment effects in randomised trials: A review of six empirical studies. 12th Cochrane Colloquium Ottawa

Als-Nielsen B, Chen W, Gluud LL, Siersma V, Hilden J, et al. (October 2004a) Are trial size and reported methodological quality associated with treatment effects? observational study of 523 randomised trials. 12th Cochrane Colloquium Ottawa

Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. (2003b) Association of funding and conclusions in randomized drug trials: A reflection of treatment effect or adverse events? JAMA 290(7): 921-928.

Aronson R, Offman HJ, Joffe RT, Naylor CD. (1996) Triiodothyronine augmentation in the treatment of refractory depression. A meta-analysis. Arch Gen Psychiatry 53(9): 842-848.

Arrich J, Piribauer F, Mad P, Schmid D, Klaushofer K, et al. (2005) Intra-articular hyaluronic acid for the treatment of osteoarthritis of the knee: Systematic review and meta-analysis. CMAJ 172(8): 1039-1043.

Assendelft WJ, Koes BW, van der Heijden GJ, Bouter LM. (1996) The effectiveness of chiropractic for treatment of low back pain: An update and attempt at statistical pooling. J Manipulative Physiol Ther 19(8): 499-507.

Assendelft WJ, Koes BW, van der Heijden GJ, Bouter LM. (1992) The efficacy of chiropractic manipulation for back pain: Blinded review of relevant randomized clinical trials. J Manipulative Physiol Ther 15(8): 487-494.

Bausell RB, Lee WL, Soeken KL, Li YF, Berman BM. (2004) Larger effect sizes were associated with higher quality ratings in complementary and alternative medicine randomized controlled trials. J Clin Epidemiol 57(5): 438-446.

Beckerman H, de Bie RA, Bouter LM, De Cuyper HJ, Oostendorp RA. (1992) The efficacy of laser therapy for musculoskeletal and skin disorders: A criteria-based meta-analysis of randomized clinical trials. Phys Ther 72(7): 483-491.

Berard A, Bravo G. (1998) Combining studies using effect sizes and quality scores: Application to bone loss in postmenopausal women. J Clin Epidemiol 51(10): 801-807.

Bjordal JM. (2003) A quantitative study of-Bias in systematic reviews . Advances in Physiotherapy 5(2): 83-96.

Brittain E, Lin D. (2005) A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. Stat Med 24(1): 1-10.

Brosseau L, Milne S, Robinson V, Marchand S, Shea B, et al. (2002) Efficacy of the transcutaneous electrical nerve stimulation for the treatment of chronic low back pain: A meta-analysis. Spine 27(6): 596-603.

Brouwers MC, Johnston ME, Charette ML, Hanna SE, Jadad AR, et al. (2005) Evaluating the role of quality assessment of primary studies in systematic reviews of cancer practice guidelines. BMC Med Res Methodol 5(1): 8.

Brown SA. (1992) Meta-analysis of diabetes patient education research: Variations in intervention effects across studies. Res Nurs Health 15(6): 409-419.

Chalmers TC, Lau J, Cappelleri JC, Schmid CH. (1994) Comparing results from the largest studies with meta-analyses of smaller studies. Controlled Clinical Trials 15(3, Supplement 1): 63S.

Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, et al. (1987) Meta-analysis of clinical trials as a scientific discipline. I: Control of-Bias and comparison with large co-operative trials. Stat Med 6(3): 315-328.

Chalmers TC, Celano P, Sacks HS, Smith H,Jr. (1983)-Bias in treatment assignment in controlled clinical trials. N Engl J Med 309(22): 1358-1361.

Chalmers TC, Matta RJ, Smith H,Jr, Kunzler AM. (1977) Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med 297(20): 1091-1096.

Chalmers TC, Block JB, Lee S. (1972) Controlled studies in clinical cancer research. N Engl J Med 287(2): 75-78.

Clifford TJ, Barrowman NJ, Moher D. (2002) Funding source, trial outcome and reporting quality: Are they related? results of a pilot study. BMC Health Serv Res 2(1): 18.

Colditz GA, Miller JN, Mosteller F. (1989) How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med 8(4): 441-454.

Colditz GA, Miller JN, Mosteller F. (1988) Measuring gain in the evaluation of medical technology. the probability of a better outcome. Int J Technol Assess Health Care 4(4): 637-642.

Coleman BD, Khan KM, Maffulli N, Cook JL, Wark JD. (2000) Studies of surgical outcome after patellar tendinopathy: Clinical significance of methodological deficiencies and guidelines for future studies. victorian institute of sport tendon study group. Scand J Med Sci Sports 10(1): 2-11.

Cosmi B, Conti E, Coccheri S. (2001) Anticoagulants (heparin, low molecular weight heparin and oral anticoagulants) for intermittent claudication. Cochrane Database Syst Rev (3)(3): CD001999.

de Oliveira IR, Dardennes RM, Amorim ES, Diquet B, de Sena EP, et al. (1995) Is there a relationship between antipsychotic blood levels and their clinical efficacy? an analysis of studies design and methodology. Fundam Clin Pharmacol 9(5): 488-502.

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. (1992) Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 45(3): 255-265.

Dube S, Heyen F, Jenicek M. (1997) Adjuvant chemotherapy in colorectal carcinoma: Results of a meta-analysis. Dis Colon Rectum 40(1): 35-41.

Egger M, Juni P, Bartlett C, Sterne J. (October 2001) Importance of different sources of-Bias in systematic reviews of controlled trials: Systematic review of empirical studies. 9th Cochrane Colloquium Lyon

Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. (1990) An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Control Clin Trials 11(5): 339-352.

Ferriter M, Huband N. (2005) Does the non-randomized controlled study have a place in the systematic review? A pilot study. Crim Behav Ment Health 15(2): 111-120.

Fortin PR, Lew RA, Liang MH, Wright EA, Beckett LA, et al. (1995) Validation of a meta-analysis: The effects of fish oil in rheumatoid arthritis. J Clin Epidemiol 48(11): 1379-1390.

Fried LF, Orchard TJ, Kasiske BL. (2001) Effect of lipid reduction on the progression of renal disease: A meta-analysis. Kidney Int 59(1): 260-269.

Friedenreich CM, Brant RF, Riboli E. (1994) Influence of methodologic factors in a pooled analysis of 13 case-control studies of colorectal cancer and dietary fiber. Epidemiology 5(1): 66-79.

Geffers C, Zuschneid I, Eckmanns T, Ruden H, Gastmeier P. (2003) The relationship between methodological trial quality and the effects of impregnated central venous catheters. Intensive Care Med 29(3): 403-409.

Gilbert JP, McPeek B, Mosteller F. (1977) Progress in surgery and anesthesia: Benefits and risks of innovative surgery. In J. P. Bunker, B.A. Barnes & F. Mosteller (eds.) (1977). Costs, Risks and Benefits of Surgery, pp. 124-169, NY: Oxford University Press.

Gillman MW, Runyan DK. (1984)-Bias in treatment assignment in controlled clinical trials. N Engl J Med 310(24): 1610-1611.

Glasziou PP, Woodward AJ, Mahon CM. (1995) Mammographic screening trials for women aged under 50. A quality assessment and meta-analysis. Med J Aust 162(12): 625-629.

Gluud LL. (2006)-Bias in clinical intervention research. Am J Epidemiol 163(6): 493-501.

Gotzsche PC, Olsen O. (2000) Is screening for breast cancer with mammography justifiable? Lancet 355(9198): 129-134.

Gotzsche PC. (1993) Meta-analysis of NSAIDs: Contribution of drugs, doses, trial designs, and meta-analytic techniques. Scand J Rheumatol 22(6): 255-260.

Gotzsche PC. (1989b) Multiple publication of reports of drug trials. Eur J Clin Pharmacol 36(5): 429-432.

Greenberg RP, Bornstein RF, Zborowski MJ, Fisher S, Greenberg MD. (1994) A meta-analysis of fluoxetine outcome in the treatment of depression. J Nerv Ment Dis 182(10): 547-551.

Gueyffier F, Froment A, Gouton M. (1996) New meta-analysis of treatment trials of hypertension: Improving the estimate of therapeutic benefit. J Hum Hypertens 10(1): 1-8.

Heyland DK, Kernerman P, Gafni A, Cook DJ. (1996) Economic evaluations in the critical care literature: Do they help us improve the efficiency of our unit? Crit Care Med 24(9): 1591-1598.

Holme I, Ekelund LG, Hjermann I, Leren P. (1994) Quality-adjusted meta-analysis of the hypertension/coronary dilemma. Am J Hypertens 7(8): 703-712.

Hovell MF. (1982) The experimental evidence for weight-loss treatment of essential hypertension: A critical review. Am J Public Health 72(4): 359-368.

Husereau D, Clifford T, Aker P, Leduc D, Mensinkai S. (2003) Spinal manipulation for infantile colic. Canadian Coordinating Office for Health Technology Assessment Technology report no 42(Ottawa)

Imperiale TF, McCullough AJ. (1990) Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomized trials. Ann Intern Med 113(4): 299-307.

Jeng GT, Scott JR, Burmeister LF. (1995a) A comparison of meta-analytic results using literature vs individual patient data. paternal cell immunization for recurrent miscarriage. JAMA 274(10): 830-836.

Karassa FB, Tatsioni A, Ioannidis JP. (2003) Design, quality, and-Bias in randomized controlled trials of systemic lupus erythematosus. J Rheumatol 30(5): 979-984.

Kelly A, Kavanagh J, Thomas J, Sterne J, Egger M. (October 2001) The last word in trial quality? the impact of selection and performance-Bias within a series of integrated systematic reviews. 9th Cochrane Colloquium Lyon

Keren R, Chan E. (2002) A meta-analysis of randomized, controlled trials comparing short- and long-course antibiotic therapy for urinary tract infections in children. Pediatrics 109(5): E70-0.

Khan A, Kolts RL, Thase ME, Krishnan KR, Brown W. (2004) Research design features and patient characteristics associated with the outcome of antidepressant clinical trials. Am J Psychiatry 161(11): 2045-2049.

Kjaergard LL, Gluud C. (2002) Funding, disease area, and internal validity of hepatobiliary randomized clinical trials. Am J Gastroenterol 97(11): 2708-2713.

Kjaergard LL, Villumsen J, Gluud C. (October 1999) Quality of randomised clinical trials affects estimates of intervention efficacy. 7th Cochrane Colloquium Rome

Kleijnen J, Knipschild P, ter Riet G. (1991a) Clinical trials of homoeopathy. BMJ 302(6772): 316-323.

Kleijnen J, Knipschild P, ter Riet G. (1991b) Trials of homeopathy. BMJ 302(6782): 960.

Kleijnen J, ter Riet G, Knipschild P. (1991c) Acupuncture and asthma: A review of controlled trials. Thorax 46(11): 799-802.

Klein S, Simes J, Blackburn GL. (1986) Total parenteral nutrition and cancer clinical trials. Cancer 58(6): 1378-1386.

Koes BW, Assendelft WJ, van der Heijden GJ, Bouter LM. (1996) Spinal manipulation for low back pain. an updated systematic review of randomized clinical trials. Spine 21(24): 2860-71; discussion 2872-3.

Koes BW, Bouter LM, van der Heijden GJ. (1995) Methodological quality of randomized clinical trials on treatment efficacy in low back pain. Spine 20(2): 228-235.

Koes BW, van Tulder MW, van der Windt WM, Bouter LM. (1994) The efficacy of back schools: A review of randomized clinical trials. J Clin Epidemiol 47(8): 851-862.

Kunz R, Vist G, Oxman AD. (2002) Randomisation to protect against selection-Bias in healthcare trials. Cochrane Database Syst Rev (2)(2): MR000012.

Kunz R, Oxman AD. (1998) The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 317(7167): 1185-1190.

Kyriakidi M, Ioannidis JP. (2002) Design and quality considerations for randomized controlled trials in systemic sclerosis. Arthritis Rheum 47(1): 73-81.

Laupacis A, Fergusson D. (1998) Erythropoietin to minimize perioperative blood transfusion: A systematic review of randomized trials. the international study of peri-operative transfusion (ISPOT) investigators. Transfus Med 8(4): 309-317.

le Chevalier T. (1996) Chemotherapy for advanced NSCLC. will meta-analysis provide the answer? Chest 109(5 Suppl): 107S-109S.

Legg L, Leonardi-Bee J, Langhorne P, Walker M. (October 2003) Is getting individual patient data for meta-analyses worthwhile? 11th Cochrane Colloquium Barcelona

Liberati A, Himel HN, Chalmers TC. (1986) A quality assessment of randomized control trials of primary treatment of breast cancer. J Clin Oncol 4(6): 942-951.

Linde K, Melchart D. (1998) Randomized controlled trials of individualized homeopathy: A state-of-the-art review. J Altern Complement Med 4(4): 371-388.

Linde K, Clausius N, Ramirez G, Melchart D, Eitel F, et al. (1997) Are the clinical effects of homeopathy placebo effects? A meta-analysis of placebo-controlled trials. Lancet 350(9081): 834-843.

Lipsey MW, Wilson DB. (1993) The efficacy of psychological, educational, and behavioral treatment. confirmation from meta-analysis. Am Psychol 48(12): 1181-1209.

Macarthur C, Foran PJ, Bailar JC,3rd. (1995) Qualitative assessment of studies included in a meta-analysis: DES and the risk of pregnancy loss. J Clin Epidemiol 48(6): 739-747.

Macedo A, Farre M, Banos JE. (2006) A meta-analysis of the placebo response in acute migraine and how this response may be influenced by some of the characteristics of clinical trials. Eur J Clin Pharmacol 62(3): 161-172.

Macleod MR, O'Collins T, Howells DW, Donnan GA. (2004) Pooling of animal experimental data reveals influence of study design and publication-Bias. Stroke 35(5): 1203-1208.

Marsoni S, Torri W, Taiana A, Gambino A, Grilli R, et al. (1990) Critical review of the quality and development of randomized clinical trials (RCTs) and their influence on the treatment of advanced epithelial ovarian cancer. Ann Oncol 1(5): 343-350.

Maxfield L, Hyer L. (2002) The relationship between efficacy and methodology in studies investigating EMDR treatment of PTSD. J Clin Psychol 58(1): 23-41.

McCormack K, Grant A, Scott N, EU Hernia Trialists Collaboration. (2004) Value of updating a systematic review in surgery using individual patient data. Br J Surg 91(4): 495-499.

McQuay H, Carroll D, Moore A. (1996) Variation in the placebo effect in randomised controlled trials of analgesics: All is as blind as it seems. Pain 64(2): 331-335.

Miller JN, Colditz GA, Mosteller F. (1989) How study design affects outcomes in comparisons of therapy. II: Surgical. Stat Med 8(4): 455-466.

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, et al. (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352(9128): 609-613.

Moncrieff J. (2003) A comparison of antidepressant trials using active and inert placebos. Int J Methods Psychiatr Res 12(3): 117-127.

Mortimer D, French S. (2006) Can dissenting findings regarding the comparative effectiveness of ICSI and IVF be explained by a learning curve? J Assist Reprod Genet 23(1): 33-36.

Morrison B, Lilford RJ, Ernst E. (2000) Methodological rigour and results of clinical trials of homoeopathic remedies Perfusion 13(3): 132-138.

Naylor CD, Detsky AS, O'Rourke K, Fonberg E. (1987) Does treatment with essential amino acids and hypertonic glucose improve survival in acute renal failure?: A meta-analysis. Ren Fail 10(3-4): 141-152.

Nurmohamed MT, Rosendaal FR, Buller HR, Dekker E, Hommes DW, et al. (1992) Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: A meta-analysis. Lancet 340(8812): 152-156.

O'Donovan PA, Vandekerckhove P, Lilford RJ, Hughes E. (1993) Treatment of male infertility: Is it effective? review and meta-analyses of published randomized controlled trials. Hum Reprod 8(8): 1209-1222.

Pildal J, Hrobjartsson A, Jorgensen K, Hilden J, Altman D, et al. (October 2004a) How often do positive conclusions drawn from meta-analyses remain substantiated if only data from randomised trials with adequate allocation concealment are considered? 12th Cochrane Colloquium Ottawa

Pogue JM, Yusuf S. (1997) Cumulating evidence from randomized trials: Utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials 18(6): 580-93; discussion 661-6.

Sack M, Lempa W, Lamprecht F. (2001) Study quality and effect-sizes - a metaanalysis of EMDR-treatment for posttraumatic stress disorder. Psychother Psychosom Med Psychol 51(9-10): 350-355.

Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. (2004) Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol 57(7): 683-697.

Schulz KF, Grimes DA, Altman DG, Hayes RJ. (1996) Blinding and exclusions after allocation in randomised controlled trials: Survey of published parallel group trials in obstetrics and gynaecology. BMJ 312(7033): 742-744.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. (1994a) Failure to conceal treatment allocation schedules in trials influenced estimates of treatment effects. Controlled Clinical Trials 15(3, Supplement 1): 63-64.

Shaikh W, Vayda E, Feldman W. (1976) A systematic review of the literature on evaluative studies of tonsillectomy and adenoidectomy. Pediatrics 57(3): 401-407.

Shapiro DA, Shapiro D. (1982) Meta-analysis of comparative therapy outcome studies: A replication and refinement. Psychol Bull 92(3): 581-604.

Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, et al. (October 2004) Multivariate modelling for meta-epidemiological assessment of the association between trial quality and apparent treatment effects in randomised clinical trials. 12th Cochrane Colloquium Ottawa

Silva Filho C. (October 2004) Impact of scoring the quality of clinical trials in results of meta-analysis 12th Cochrane Colloquium Ottawa

Sitter H, Nies C, Krack W, Celik I, Pruente H, et al. (July 2002) Influence of trial design on validity of trial results for cholecystectomy trials. 4th Symposium on Systematic Reviews: Pushing the Boundaries Oxford

Stanton MD, Shadish WR. (1997) Outcome, attrition, and family-couples treatment for drug abuse: A meta-analysis and review of the controlled, comparative studies. Psychol Bull 122(2): 170-191.

Stein DJ, Ipser JC, Seedat S. (2006) Pharmacotherapy for post traumatic stress disorder (PTSD). Cochrane Database Syst Rev (1)(1): CD002795.

Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. (2005) Association between compliance with methodological standards of diagnostic research and reported test accuracy: Meta-analysis of focused assessment of US for trauma. Radiology 236(1): 102-111.

Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, et al. (2002) Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 21(11): 1513-1524.

Szczech LA, Berlin JA, Feldman HI. (1998) The effect of antilymphocyte induction therapy on renal allograft survival. A meta-analysis of individual patient-level data. anti-lymphocyte antibody induction therapy study group. Ann Intern Med 128(10): 817-826.

Towheed TE, Hochberg MC. (1997) A systematic review of randomized controlled trials of pharmacological therapy in osteoarthritis of the knee, with an emphasis on trial methodology. Semin Arthritis Rheum 26(5): 755-770.

Vranos G, Tatsioni A, Polyzoidis K, Ioannidis JP. (2004) Randomized trials of neurosurgical interventions: A systematic appraisal. Neurosurgery 55(1): 18-25; discussion 25-6.

Wahlbeck K, Tuunainen A, Gilbody S, Adams CE. (2000) Influence of methodology on outcomes of randomised clozapine trials. Pharmacopsychiatry 33(2): 54-59.

Westwood ME, Whiting PF, Kleijnen J. (2005) How does study quality affect the results of a diagnostic meta-analysis? BMC Med Res Methodol 5(1): 20.

Whiting P, Harbord R, Kleijnen J. (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5: 19.

Wilson DB, Lipsey MW. (2001) The role of method in treatment effectiveness research: Evidence from meta-analysis. Psychol Methods 6(4): 413-429.

Wortman P. (1983) Meta-analysis: A validity perspective. Annual Review of Psychology 34, 223–26: 223-226.

Anhang 2: Aus der SR in Abschnitt 3 ausgeschlossene Studien

Angelillo IF, Villari P. (2003) Meta-analysis of published studies or meta-analysis of individual data? caesarean section in HIV-positive women as a study case. Public Health 117(5): 323-328.

Benjamin RS. (1999) Evidence for using adjuvant chemotherapy as standard treatment of soft tissue sarcoma. Semin Radiat Oncol 9(4): 349-351.

Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. (July 2000) Anti-lymphocyte antibody-induction therapy study group. individual patient versus group-level data metaregressions for the investigation of treatment effect modifiers: Ecological-Bias rears its ugly head. Third Symposium on Systematic Reviews: Beyond the Basics Oxford

Clarke M, Stewart L. (1998c) Re: "comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies". Am J Epidemiol 148(1): 102-103.

Clarke M, Stewart L, Pignon JP, Bijnens L. (1998d) Individual patient data meta-analysis in cancer. Br J Cancer 77(11): 2036-2044.

Clarke M, Stewart L. (October 1997) Individual patient data or published meta-analysis: A systematic review. 5th Cochrane Colloquium and Second International Conference Scientific Basis of Health Services Amsterdam

Clauser C, Nieri M, Franceschi D, Pagliaro U, Pini-Prato G. (2003) Evidence-based mucogingival therapy. part 2: Ordinary and individual patient data meta-analyses of surgical treatment of recession using complete root coverage as the outcome variable. J Periodontol 74(5): 741-756.

Collette L, Suciu S, Bijnens L, Sylvester R. (January 1998) Including literature data in individual patient data meta-analyses for time-to-event endpoints. First Symposium on Systematic Reviews: Beyond the Basics Oxford

Delmas PD, Li Z, Cooper C. (2004) Relationship between changes in bone mineral density and fracture risk reduction with antiresorptive drugs: Some issues with meta-analyses. J Bone Miner Res 19(2): 330-337.

Duchateau L, Collette L, Suciu S, Sylvester R. (January 1999) The impact of including survival information from literature on the overall x2 in an IPD meta-analysis. Second Symposium on Systematic Reviews: Beyond the Basics Oxford

Duchateau L, Collette L, Sylvester R, Pignon JP. (2000) Estimating number of events from the kaplan-meier curve for incorporation in a literature-based meta-analysis: What you don't see you can't get! Biometrics 56(3): 886-892.

Franzosi MG, Santoro E, Santoro L. (October 1997a) Use of individual patient data vs. published reports in a meta-analysis: The case of ace-inhibitors in myocardial infarction. 5th Cochrane Colloquium and Second International Conference Scientific Basis of Health Services Amsterdam

Grant A, EU Hernia Trialists Collaboration. (Juni 1999) Published data alone for meta-analysis may be sufficient: The experience of the EU hernia trialists collaboration. 15th Annual Meeting of the International Society of Technology Assessment in Health Care Edinburgh

Gueyffier F, Boissel JP, Boutitie F. (October 1997) What could be expected from systematic reviews on individual patient data? examples from hypertension treatment. 5th Cochrane Colloquium and Second International Conference Scientific Basis of Health Services Amsterdam

Jeng GT, Scott JR, Burmeister LF. (1995b) Paternal cell immunization for recurrent miscarriage: A comparison of meta-analyses with and without individual patient data. Controlled Clinical Trials 16(3, Supplement 1): 40S.

Koopman L, van-der HG, Glasziou P, Grobbee R, Rovers M. (October 2005) Methodology of subgroup analyses used in individual patient data meta-analyses versus meta-analyses on published data. 13th Cochrane Colloquium Melbourne

Lambert PC, Sutton AJ, Abrams KR, Jones DR. (2002) A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J Clin Epidemiol 55(1): 86-94.

Lambert PC, Sutton AJ, Abrams KR, Jones DR. (October 2001) A comparison of individual patient data meta-analysis with the use of summary patient-level covariates in meta-regression. 9th Cochrane Colloquium Lyon

Legg L, Leonardi-Bee J, Langhorne P, Walker M. (October 2003) Is getting individual patient data for meta-analyses worthwhile? 11th Cochrane Colloquium Barcelona

Li Z, Meredith MP. (2003) Exploring the relationship between surrogates and clinical outcomes: Analysis of individual patient data vs. meta-regression on group-level summary statistics. J Biopharm Stat 13(4): 777-792.

Ludwig H, Fritz E. (2000) Interferon in multiple myeloma--summary of treatment results and clinical implications. Acta Oncol 39(7): 815-821.

Man-Son-Hing M, Wells G, Lau A. (1998) Quinine for nocturnal leg cramps: A meta-analysis including unpublished data. J Gen Intern Med 13(9): 600-606.

Marino P, Pampallona S, Preatoni A, Cantoni A, Invernizzi F. (1994) Chemotherapy vs supportive care in advanced non-small-cell lung cancer. results of a meta-analysis of the literature. Chest 106(3): 861-865.

Mathew T, Nordstrom K. (1999) On the equivalence of meta-analysis using literature and using individual patient data. Biometrics 55(4): 1221-1223.

McCormack K, Scott N. (June 2003) Outcome reporting-Bias and individual patient data meta-analysis: A case study in surgery. improving outcomes through health technology assessment. 19th Annual Meeting of the International Society of Technology Assessment in Health Care Alberta

McCormack K, Scott N. (October 2001b) Outcome reporting-Bias and individual patient data meta-analysis: A case study in surgery. the EU hernia trialists collaboration. 9th Cochrane Colloquium Lyon

McCormack K, Scott N. (October 2001a) Are trials with individual patient data available different from trials without individual patient data available? the EU hernia trialists collaboration. 9th Cochrane Colloquium Lyon

Michiels S, Syz N, Piedbois P, Burdett S, Stewart L, et al. (August 2002) Is it possible to perform a meta-analysis when only the median survival time is known? A comparison with individual patient data results. 10th Cochrane Colloquium Stavanger

Olkin I, Sampson A. (1998) Comparison of meta-analysis versus analysis of variance of individual patient data. Biometrics 54(1): 317-322.

Pignon JP, Luc D, Bertin S, Sylvester R, Institut GR. (October 1999) Individual patient and literature based meta-analyses of chemotherapy in head and neck cancer: Reasons for differing results. 7th Cochrane Colloquium Rome

Parmar MK, Torri V, Stewart L. (1998) Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 17(24): 2815-2834.

Pignon JP, Bourhis J. (1995) Meta-analysis of chemotherapy in head and neck cancer: Individual patient data vs literature data. Br J Cancer 72(4): 1062-1063.

Scott NW, Webb K, Graham P, Grant AM, EU Hernia Trialists Collaboration. (June 2001) The added value of obtaining individual patient data for systematic reviews of randomised trials: The experience of the EU hernia trialists collaboration. 17th Annual Meeting of the International Society of Technology Assessment in Health Care Philadelphia

Steinberg KK, Smith SJ, Stroup DF, Olkin I, Lee NC, et al. (1997) Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. Am J Epidemiol 145(10): 917-925.

Stewart LA, Parmar MKB, Clarke M. (October 1995) Benefits of meta-analysis using individual patient data (IPD). Scientific Basis of Health Services: International Conference London

Williamson PR, Marson A, Hutton J, Chadwick D. (Januray 1998) Individual patient versus aggregate data meta-analysis for time-to-event outcomes: Empirical evidence from epilepsy. First Symposium on Systematic Reviews: Beyond the Basics Oxford

Williamson PR, Hutton J, Marson A, Chadwick D. (1997) Individual patient data meta-analyses for time-to-event outcomes: An example from epilepsy. Controlled Clinical Trials 18(3, Supplement 1): S184.