G-2 Gutachten:

Aussagekraft von Subgruppenanalysen

25.03.2004

Fachbereich Evidenz-basierte Medizin

Medizinischer Dienst der Spitzenverbände der Krankenkassen e.V.

Dr. S. Janatzek (geb. Ziegler), Fachbereich Evidenz-basierte Medizin, MDS

unter Mitarbeit von:

C. Arndt, Fachbereich Evidenz-basierte Medizin, MDS

Prof. Dr. J. Windeler, Fachbereich Evidenz-basierte Medizin, MDS

1 Verzeichnisse

1.1 Inhaltsverzeichnis

1.2 Tabellenverzeichnis

1.3 Abkürzungen und Bezeichnungen

Alpha-Adjustierung	siehe Abschnitt 5.3.2, erster Aufzählungspunkt
anti-konservativ	Synonym: liberal Ein statistischer Test ist *anti-konservativ*, wenn sein Fehler 1. Art *größer* als das vorgegebene Signifikanzniveau (α) ist.
a-priori-Subgruppenanalysen	Synonyme: vorab geplante Subgruppenanalysen, prädefinierte Subgruppenanalysen Subgruppenanalyse, die *vor* Studienbeginn festgelegt wurde (siehe Abschnitt 3)
Fehler 1. Art	Der Fehler 1. Art eines statistischen Tests ist die Wahrscheinlichkeit, dass der statistische Test ein signifikantes Ergebnis liefert (d.h. den Unterschied zwischen den Therapiegruppen feststellt), auch wenn in Wahrheit kein Unterschied vorhanden ist.
Fehler 2. Art	Der Fehler 2. Art eines statistischen Tests ist die Wahrscheinlichkeit, dass der statistische Test einen in Wahrheit vorhandenen Unterschied zwischen den Therapiegruppen nicht aufdeckt (d.h. ein nicht-signifikantes Ergebnis liefert).
Interaktionstest	Synonym: Test auf Interaktion Ein Interaktionstest untersucht, ob sich der wahre *Therapieeffekt* (Unterschied zwischen den Therapiegruppen) zwischen den Subgruppen unterscheidet. Ein signifikanter Interaktionstest weist auf unterschiedliche Therapieeffekte in den Subgruppen hin. Siehe auch 5.3.2, zweiter Aufzählungspunkt.
konservativ	Ein statistischer Test ist *konservativ*, wenn sein Fehler 1. Art *kleiner* als das vorgegebene Signifikanzniveau (α) ist.
Niveau	Synonym für „Signifikanzniveau“, siehe Seite 13
post-hoc-Subgruppenanalysen	Subgruppenanalyse, die *nicht* vor Studienbeginn festgelegt wurde (siehe Abschnitt 3)
Power	Die Power eines statistischen Tests ist gleich: 1 – Fehler 2. Art. Die Power ist also die Wahrscheinlichkeit, dass der statistische Test einen in Wahrheit vorhandenen Unterschied zwischen den Therapiegruppen aufdeckt (durch ein signifikantes Ergebnis).
RCT	randomisierte Studie (RCT: **r**andomized **c**ontrolled **t**rial)
stratifizierte Randomisation	separate Randomisierung in einzelnen Patientengruppen (Strata), z.B. getrennte Randomisierung der weiblichen und der männlichen Studienteilnehmer
Subgruppenanalyse	Synonym: Untergruppenanalyse
Subgruppen-spezifische Tests	In jeder Subgruppe wird separat ein Test auf Therapieeffekt durchgeführt.

2 Fragestellung / Auftrag

Am 08.07.2003 wurde der MDS von den Spitzenverbänden der Krankenkassen beauftragt, ein Grundsatzgutachten zur Aussagekraft von post-hoc Subgruppenanalysen im Vergleich zu vorab geplanten Subgruppenanalysen zu erstellen.

In diesem Gutachten soll gemäß Auftrag „eine Bewertung der Aussagekraft von retrospektiven Subgruppenanalysen im Vergleich zu im Rahmen der Primärstudie prädefinierten Subgruppenanalysen“ vorgenommen werden.

3 Einleitung

Subgruppenanalysen (Untergruppenanalysen) sind sehr gebräuchlich. In fast jeder Publikation einer klinischen Studie finden sich eine Reihe von Subgruppenanalysen. Gleichzeitig wird immer wieder vor Überinterpretationen von Subgruppenanalysen gewarnt und generell zur Vorsicht bei der Interpretation geraten.

Es gibt viele Beispiele dafür, dass Subgruppenanalysen irreführende Ergebnisse liefern können – sowohl falsch-positive als auch falsch-negative Ergebnisse. Exemplarisch sei hier ein prominentes Beispiel erwähnt: In der randomisierten ISIS-2 Studie mit insgesamt 17187 Patienten wurde beim Vergleich von Aspirin mit Placebo in der Therapie des akuten Herzinfarktes eine signifikante Reduktion der 1-Monats-Mortalität unter Aspirin beobachtet. Eine eigentlich absurde Subgruppenanalyse, bei der die Patienten nach ihrem astrologischen *Sternzeichen* eingeteilt wurden, zeigte, dass Aspirin bei Patienten mit dem Sternzeichen Steinbock besonders wirksam ist, während es bei Patienten mit dem Sternzeichen Waage oder Zwilling keinen Vorteil gegenüber Placebo darstellt (es wurde sogar eine etwas höhere Mortalitätsrate in der Aspirin-Gruppe beobachtet) (ISIS-2 (1988), Peto et al. (1995)).

Angesichts des Aufwandes, den die Durchführung einer klinischen Studie darstellt, ist es verständlich und durchaus in gewissem Maße wünschenswert, dass durch Subgruppenanalysen versucht wird, möglichst viele Informationen aus den erhobenen Daten zu gewinnen.

Subgruppenanalysen können aus verschiedenen Gründen von Interesse sein, z.B. weil man vermutet, dass spezielle Subgruppen von Patienten besonders ausgeprägt von der zu prüfenden Therapie profitieren, oder weil man ausschließen möchte, dass trotz eines positiven Gesamteffektes spezielle Patienten überhaupt nicht von der Therapie profitieren. Auch für eine Steuerung der Ressourcenverteilung ist das Heranziehen von Subgruppenanalysen denkbar: Hat z.B. eine Studie einen signifikanten und relevanten Gesamteffekt einer neuen therapeutischen Maßnahme gezeigt, reichen jedoch die Ressourcen nicht aus, um zukünftig alle in Frage kommenden Patienten mit dieser Maßnahme zu versorgen, so liegt die Frage nahe, welche Patientengruppen am meisten von der Therapie profitieren.

Man kann 2 Typen von Subgruppenanalysen unterscheiden:

- **a-priori-Subgruppenanalysen** und
- **post-hoc-Subgruppenanalysen**.

Unter „a-priori-Subgruppenanalysen“ werden im vorliegenden Gutachten Subgruppenanalysen verstanden, die

- *vor Studienbeginn festgelegt* und üblicherweise im Studienprotokoll dargelegt wurden sowie
- Subgruppenanalysen, die zwar *nach* Studienbeginn, aber *ohne Einsicht in die bereits erhobenen Daten* definiert und in einem Amendment zum Studienprotokoll dargelegt (bzw. auf andere Weise erkennbar festgehalten) wurden.

Alle anderen Subgruppenanalysen werden hier als „post-hoc-Subgruppenanalysen" bezeichnet. Nach dieser Definition umfassen post-hoc-Subgruppenanalysen sowohl

- Subgruppenanalysen, die nach Sicht auf die Daten (***datengetrieben***) durchgeführt wurden, als auch
- Subgruppenanalysen, die nach Studienbeginn, aber ohne Einsicht in die bereits erhobenen Daten definiert, jedoch *nicht erkennbar festgehalten* wurden.

Letztere Art von Subgruppenanalyse (letzter Aufzählungspunkt) kommt jedoch kaum vor. Deshalb sind post-hoc-Subgruppenanalysen im Sinne der obigen Definition in aller Regel *datengetriebene* Subgruppenanalysen.

In der Literatur findet sich keine einheitliche Definition der beiden Begriffe. Während z.B. Parker et al. (2000) die hier angegebene Definition verwenden, werden im Consort-Statement (Altman et al. (2001)) und bei Detsky et al. (1995) unter dem Begriff „post-hoc-Subgruppenanalysen" ausschließlich Subgruppenanalysen verstanden, die nach Sicht auf die Daten durchgeführt wurden.

Im vorliegenden Gutachten wird die Aussagekraft von post-hoc-Subgruppenanalysen mit der Aussagekraft von a-priori-Subgruppenanalysen verglichen. Dazu wird die Aussagekraft beider Typen von Subgruppenanalysen untersucht, um anschließend einen Vergleich anstellen zu können.

Hierbei werden Subgruppenanalysen von Therapiestudien, besonders von randomisierten Studien, fokussiert.

4 Beschreibung des Vorgehens

Es wurde eine systematische Bewertung der Evidenzlage vorgenommen. Dazu wurden folgende Arbeitsschritte durchgeführt:

- Durchführung einer systematischen Recherche
- Entscheidung über Ein- und Ausschluss der Arbeiten anhand der Auswahlkriterien
- Datenextraktion und kritische Bewertung der eingeschlossenen Arbeiten
- Erstellung des Gutachtens.

4.1 Recherche

4.1.1 Vorgelegte Unterlagen

Bei Beginn der Gutachtenerstellung lagen in der Literaturdatenbank des Fachbereichs Evidenz-basierte Medizin einige Artikel zu Subgruppenanalysen vor.

4.1.2 Systematische Recherche

Eine systematische Literaturrecherche wurde in den folgenden Datenbanken vorgenommen: Cancerlit, Euroethics, Ethmed, GEROLIT, MEDIKAT, MEDLINE, Oldmedline, TOXLINE, AnimAlt-ZEBET, Medline Alert, Kluwer-Verlagsdatenbank für Volltexte, Springer-Verlagsdatenbank für Volltexte, Springer PrePrint, Thieme-Verlagsdatenbank für Volltexte und in 2 Datenbanken der Cochrane Library: „Cochrane Database of Methodology Reviews“ und „Cochrane Methodology Register“.

Die Recherche in den Datenbanken wurde im September und November 2003 durchgeführt. Genauere Angaben zu den Datenbanken, zu ihren Zugängen, der Anzahl der Treffer und den Suchstrategien finden sich in Abschnitt 9.2. Auf eine mehrfache Anpassung der Suchstrategie wurde verzichtet. Es wurden Einschränkungen bezüglich des Publikationsjahres vorgenommen (siehe 9.2), um eine im vorgesehenen Zeitrahmen bearbeitbare Anzahl relevanter Arbeiten zu erhalten. Ferner wurde eine Einschränkung bezüglich der Publikationssprache vorgenommen, siehe 9.2.

4.1.3 Umfeldrecherche

Es wurde eine orientierende Recherche im Internet (allgemeine Suchmaschine Google) durchgeführt. Außerdem wurden die Internetseiten der Zulassungsbehörden (FDA, EMEA, BfArM) durchsucht.

4.2 Auswahl der Arbeiten anhand der Auswahlkriterien

4.2.1 Benennung der Auswahlkriterien

Eine Arbeit wurde berücksichtigt, falls sie Aussagen zur Aussagekraft von Subgruppenanalysen enthält.

Dabei wurden auch Arbeiten eingeschlossen, die sich ausschließlich mit a-priori- oder ausschließlich mit post-hoc-Subgruppenanalysen beschäftigen. Vergleichende Aussagen (zum Vergleich von a-priori- mit post-hoc-Subgruppenanalysen) waren nicht zwingend erforderlich.

Ausgewählt wurden sowohl Arbeiten, die Belege bzw. Begründungen für ihre Aussagen zur Aussagekraft von Subgruppenanalysen enthalten als auch Arbeiten, die keine Belege bzw. Begründungen enthalten.

Arbeiten, die lediglich Beispiele von Subgruppenanalysen, aber keine Aussagen zur Aussagekraft von Subgruppenanalysen enthalten, wurden nicht eingeschlossen. Es wurden ausschließlich Aussagen zur Aussagekraft von Subgruppenanalysen in *Einzelstudien* berücksichtigt, Ausführungen zu Subgruppenanalysen in systematischen Reviews bzw. Meta-Analysen wurden nicht berücksichtigt.

Notation: Arbeiten, die die Auswahlkriterien erfüllen, werden als ***„eingeschlossene Arbeiten“*** bezeichnet.
Arbeiten, die die Auswahlkriterien nicht erfüllen, werden als ***„ausgeschlossene Arbeiten“*** bezeichnet.

4.2.2 Vorgehensweise

Die durch die Recherche erhaltenen Abstracts wurden gelesen und es wurde entschieden, welche Artikel die oben aufgeführten Auswahlkriterien sicher nicht erfüllen. Diese Artikel wurden ausgeschlossen, die restlichen Artikel wurden als Volltexte beschafft. Jeder Volltext-Artikel wurde gelesen und es wurde anhand der Auswahlkriterien entschieden, ob er berücksichtigt wird oder nicht.

Die Literaturverzeichnisse der eingeschlossenen Artikel sowie weiterer relevanter Artikel wurden daraufhin geprüft, ob sie Hinweise auf weitere wichtige Quellen enthalten. Wurden solche Quellen identifiziert, wurden die Artikel ebenfalls beschafft und es wurde entschieden, ob sie die Auswahlkriterien erfüllen. Dabei wurden im Unterschied zur systematischen Recherche in den Literaturdatenbanken auch Arbeiten berücksichtigt, die vor 1990 publiziert wurden.

Die Entscheidung darüber, ob die Auswahlkriterien erfüllt sind, erfolgte durch nur eine Person.

Alle durch die Recherche erhaltenen Artikel und Abstracts wurden in ein Literaturverwaltungssystem (Reference Manager®) eingegeben, mit Ein- oder Ausschluss gekennzeichnet und zur Einsicht archiviert.

4.3 Bearbeitung der ausgewählten Arbeiten

4.3.1 Datenextraktion und Bewertung

Aus den eingeschlossenen Arbeiten wurden die Aussagen zur Aussagekraft von Subgruppenanalysen sowie die dort angegebenen Belege / Begründungen für diese Aussagen, soweit verfügbar, extrahiert. Die Datenextraktion erfolgte durch eine Person.

4.3.2 Zusammenfassende Bewertung

Die zusammenfassende Bewertung der Aussagekraft von Subgruppenanalysen erfolgte durch eine narrative Zusammenschau der Aussagen der eingeschlossenen Arbeiten.

5 Ergebnisse

5.1 Ergebnis der Recherche

In den beim DIMDI verfügbaren kostenfreien Literaturdatenbanken wurden insgesamt 130 Treffer erzielt. 28 dieser Treffer wurden im Volltext beschafft und bearbeitet. Von diesen erfüllten 18 Arbeiten die Auswahlkriterien. Die Recherche in den beiden Datenbanken „Cochrane Database of Methodology Reviews“ und „Cochrane Methodology Register“ der Cochrane Library lieferte 26 Treffer. 8 der Treffer wurden als Volltexte bearbeitet, eine der 8 Arbeiten wurde eingeschlossen. Durch das Prüfen der Literaturverzeichnisse der eingeschlossenen und relevanten Publikationen wurden weitere 18 Artikel identifiziert, die die Auswahlkriterien erfüllen. Mit der Google-Suchmaschine wurden 2 Treffer gefunden; beide erfüllten nicht die Auswahlkriterien. Die Recherche bei den Zulassungsbehörden lieferte eine einzuschließende Arbeit. Zusätzlich wurde exemplarisch ein Standard-Lehrbuch (Altman, 1991) herangezogen und eingeschlossen.

Insgesamt wurden 38 Arbeiten gefunden, die die Auswahlkriterien erfüllen. Genaue Angaben zum Ergebnis der Recherche finden sich in Abschnitt 9.2.

5.2 Ausgeschlossene Arbeiten

Die Arbeiten, die nach Durchsicht des Abstracts oder des Volltextes ausgeschlossen wurden, sind in der Tabelle „Ausgeschlossene Publikationen“ (Abschnitt 9.1) aufgeführt. Die Tabelle enthält den Ausschlussgrund für jeden Artikel.

5.3 Eingeschlossene Arbeiten

5.3.1 Einzeldarstellung

Es wurden 38 Publikationen identifiziert, die die in 4.2.1 beschriebenen Auswahlkriterien erfüllen. Diese lassen sich in die folgenden 5 Kategorien einteilen:

- 5 Arbeiten, die **empirische Untersuchungen** enthalten,
- 2 Arbeiten, die **statistische Simulationsstudien** enthalten,
- 1 Arbeit, die eine **durch Simulation generierte Studie** enthält,
- 1 Arbeit, die **mathematische Begründungen** für ihre Aussagen enthält,
- 29 Arbeiten, die **keine Belege bzw. Begründungen** für ihre Aussagen enthalten und ferner in keine der vorgenannten 4 Kategorien gehören (einige dieser Arbeiten enthalten Beispiele, die der Illustration dienen).

Die 38 Publikationen sind – getrennt nach diesen 5 Kategorien – in Tabelle 2 bis Tabelle 6 im Anhang 9.1 dargestellt. Die Tabellen enthalten die Angaben zur Aussagekraft von Subgruppenanalysen, die aus den Publikationen extrahiert wurden.

Die 5 Publikationen mit **empirischen Untersuchungen** berichten insgesamt über 4 empirische Untersuchungen. Das Ziel dieser empirischen Untersuchungen ist es *nicht*, die *Aussagekraft* von Subgruppenanalysen zu untersuchen, sondern zu untersuchen, wie gut die Qualität der in Studienpublikationen dargestellten Subgruppenanalysen ist. Die empirischen Untersuchungen sind folglich *keine* Belege / Beweise für Aussagen zur Aussagekraft von Subgruppenanalysen. Deshalb wurden die konkreten Ergebnisse der empirischen Untersuchungen für das vorliegende Gutachten nicht extrahiert, extrahiert wurden lediglich die Angaben zur Aussagekraft von Subgruppenanalysen.

In den beiden Arbeiten mit **statistischen Simulationsstudien** wird über die selbe Simulationsstudie berichtet. Ziel dieser Simulationsstudie war es, die beiden gebräuchlichsten statistischen Methoden für Subgruppenanalysen, das separate Auswerten der Subgruppen sowie den Test auf Interaktion, hinsichtlich ihres Fehlers 1. Art und ihrer Power zu untersuchen. Die Ergebnisse dieser Untersuchung werden im vorliegenden Gutachten nur qualitativ dargestellt, da sie nur eingeschränkt mit der *Aussagekraft* von Subgruppenanalysen zu tun haben: Für die Frage nach der Aussagekraft ist lediglich relevant, welche Auswertungstechniken zu falschen Schlussfolgerungen führen können; diese Informationen wurden neben weiteren Äußerungen zur Aussagekraft von Subgruppenanalysen aus den beiden Publikationen extrahiert.

Die Arbeit, die eine **durch Simulation generierte Studie** enthält, berichtet über eine simulierte randomisierte Studie (Lee, 1980). Ausgehend von echten Patientendaten wurde die randomisierte Studie so simuliert, dass ihr *wahrer* Therapieeffekt – im Unterschied zu realen Studien – bekannt war. Ziel dieser Arbeit war es, zu illustrieren, dass Subgruppenanalysen falsch-positive Ergebnisse liefern können.

Die Arbeit, die **mathematische Begründungen** enthält (Cui, 2002), zeigt, dass es bei Subgruppenanalysen in randomisierten Studien passieren kann, dass die Therapiegruppen innerhalb der Subgruppen aufgrund von Imbalancen nicht vergleichbar sind – besonders dann,

wenn die Randomisierung nicht stratifiziert nach der Subgruppen-bildenden Variable durchgeführt wurde oder wenn die Subgruppen klein sind.

Die meisten der eingeschlossenen Arbeiten enthalten erwartungsgemäß **keine Belege bzw. Begründungen** für ihre Aussagen. Dass solche Belege kaum zu erwarten sind, wird im nächsten Abschnitt näher ausgeführt.

In den meisten der eingeschlossenen Arbeiten werden Subgruppenanalysen in *randomisierten* Studien besprochen (in 20 der 38 Arbeiten).

5.3.2 Zusammenfassende Darstellung

In den 38 eingeschlossenen Publikationen finden sich die nachfolgend aufgelisteten 8 Kernaussagen; eine Erläuterung, welche Aussagen sich in welchen Publikationen finden, wird im Anschluss an diese Auflistung gegeben:

1) Die häufig in Subgruppenanalysen angewendete Auswertungstechnik, die Subgruppen ausschließlich separat auszuwerten, ist inadäquat. Die adäquate Auswertungstechnik besteht darin, zunächst einen Test auf Interaktion durchzuführen und wenn dieser signifikant ist, die Subgruppen separat auszuwerten.

Erläuterung:

Das alleinige separate Auswerten der Subgruppen durch Subgruppen-spezifische Tests ist anti-konservativ, d.h. das vorgegebene Signifikanzniveau wird überschritten. Diese Auswertungstechnik führt demnach zu häufig (häufiger als mit der vorgegebenen Irrtumswahrscheinlichkeit / Signifikanzniveau) zu einem „falsch-positiven" Ergebnis (obwohl in Wahrheit kein Unterschied zwischen den Therapiegruppen vorhanden ist, wird in mindestens einer Subgruppe durch das statistische Verfahren ein Unterschied behauptet).

Auch kann durch das Gegenüberstellen der Subgruppen-spezifischen P-Werte *nicht* auf das Vorhandensein oder Nicht-Vorhandensein eines Subgruppeneffektes geschlossen werden.

Die Durchführung eines sogenannten **Interaktionstests** ist zunächst erforderlich. Ein Interaktionstest untersucht, ob sich der wahre *Therapieeffekt* (Unterschied zwischen den Therapiegruppen) zwischen den Subgruppen unterscheidet. Ein signifikanter Interaktionstest weist auf einen Subgruppeneffekt, d.h. auf unterschiedliche Therapieeffekte in den Subgruppen, hin. Ein solcher Subgruppeneffekt bedeutet konkret, dass die Wirksamkeit oder Sicherheit der zu prüfenden therapeutischen Maßnahme in den verschiedenen Subgruppen *unterschiedlich* ist.

Beleg der Aussage:

siehe Seite 18 („Zu 1: ...")

Vorgehen:

oben beschrieben (fett gedruckter Text), siehe außerdem Punkt 2 (hierarchische Testprozedur bzw. Alpha-Adjustierung)

2) In Subgruppenanalysen tritt das sogenannte „Problem des multiplen Testens“ auf. Deshalb ist es erforderlich, sich auf wenige Subgruppenanalysen zu beschränken und möglichst eine statistische Korrektur für multiples Testen vorzunehmen (bzw. eine (hierarchische) Testprozedur anzuwenden, die das Niveau einhält).

Erläuterung:

Beim Durchführen von Subgruppenanalysen werden in der Regel mehrere statistische Tests durchgeführt. Je mehr statistische Tests aber zu einer Studie durchgeführt werden, umso wahrscheinlicher ist es, dass einige der Tests fälschlicherweise signifikante Ergebnisse liefern (d.h. Unterschiede zwischen den Therapiegruppen feststellen, obwohl in Wahrheit keine Unterschiede vorhanden sind).

Das Konzept des statistischen Tests sieht vor, dass diese Wahrscheinlichkeit eines falsch-positiven (fälschlich signifikanten) Ergebnisses eines statistischen Tests höchstens so groß sein kann wie das vorgegebene Signifikanzniveau α (meist wird α=0,05 gewählt). Die Wahrscheinlichkeit eines falsch-positiven Ergebnisses (Fehler 1. Art) ist also bei einem adäquaten statistischen Test „gedeckelt“. Bei der Durchführung mehrerer statistischer Tests zu einer Studie funktioniert diese Deckelung nicht mehr; die Wahrscheinlichkeit eines falsch-positiven Ergebnisses wird größer als das vorgegebene Signifikanzniveau α.

Werden in einer Studie *k* unabhängige statistische Tests (z.B. Tests in disjunkten Subgruppen) zum Signifikanzniveau α durchgeführt, so ist die Wahrscheinlichkeit, dass mindestens einer der *k* Tests ein signifikantes Ergebnis liefert, auch wenn in Wahrheit bei keinem der *k* Vergleiche ein Unterschied zwischen den Therapiegruppen vorliegt, gleich

$$1-(1-\alpha)^k .$$

Werden z.B. 10 Tests zum Signifikanzniveau α = 0,05 durchgeführt, so ist mit 40%-iger Wahrscheinlichkeit mindestens einer der 10 Tests signifikant, obwohl in Wahrheit kein Unterschied zwischen den Therapiegruppen vorliegt. Bei 20 Tests beträgt die Wahrscheinlichkeit 64%, bei 50 Tests 92% und bei 100 Tests 99%.

Das Problem des multiplen Testens kann auch auf folgende Weise illustriert werden: Werden 100 statistische Tests durchgeführt, alle zum Signifikanzniveau 0,05, so ist zu erwarten, dass 5 dieser Tests ein signifikantes Ergebnis liefern, obwohl in Wahrheit kein Unterschied zwischen den Therapiegruppen vorhanden ist.

Bedenkt man, dass bei der Durchführung von Subgruppenanalysen häufig nicht nur mehrere Subgruppen-bildende Variablen (z.B. Geschlecht, Alter), sondern darüber hinaus auch mehrere Zielgrößen herangezogen werden, so wird deutlich, dass vielfach eine *große Anzahl* statistischer Tests durchgeführt wird und damit das Problem des multiplen Testens hoch relevant wird. Selbst wenn nur *eine* Subgruppenanalyse durchgeführt wird, handelt es sich um multiples Testen, da sowohl ein Test auf Gesamteffekt (Hauptauswertung der Studie) als auch der Interaktionstest und falls dieser signifikant ist noch die Subgruppen-spezifischen Tests durchgeführt werden, siehe Punkt 1.

Beleg der Aussage:

Das Problem des multiplen Testens ist bereits bewiesen, hierfür sind *keine neuen Belege mehr erforderlich*.

Vorgehen:

Es gibt statistische Methoden, um für multiples Testen zu korrigieren, so dass insgesamt der Fehler 1. Art (Wahrscheinlichkeit, dass mindestens einer der durchgeführten Tests ein signifikantes Ergebnis liefert, wenn in Wahrheit kein Unterschied zwischen den Therapien vorliegt) das vorgegebene Signifikanzniveau α nicht überschreitet. Eine einfache, aber konservative Methode ist die sogenannte **Alpha-Adjustierung** oder Bonferroni-Adjustierung: Hierbei werden die einzelnen statistischen Tests nicht zum Signifikanzniveau α, sondern zum Signifikanzniveau

$$\frac{\alpha}{\text{Anzahl statistischer Tests}}$$

durchgeführt. Ein statistischer Test ist erst dann statistisch signifikant, falls sein P-Wert kleiner ist als dieses Signifikanzniveau. Beispielsweise sind also für ein Gesamtniveau von α=0,05 und 20 Tests die P-Werte der einzelnen Tests mit 0,0025 (anstatt mit 0,05, wie es bei der Durchführung nur eines Tests der Fall wäre) zu vergleichen.

Der Nachteil solcher Alpha-Adjustierungen ist, dass die *Power* der Tests, d.h. die Wahrscheinlichkeit, dass tatsächlich vorhandene Effekte durch die statistischen Tests aufgedeckt werden können, *gering* wird.

Generell kann das Problem des multiplen Testens dadurch gelöst werden, dass eine Testprozedur (bestehend aus allen durchzuführenden Tests) verwendet wird, die insgesamt das vorgegebenen (multiple) Signifikanzniveau einhält. Eine Möglichkeit einer solchen Testprozedur ist die oben beschriebene Bonferroni-Adjustierung. Es kann aber auch **(hierarchische) Testprozeduren** geben, die das vorgegebene Signifikanzniveau α einhalten, obwohl alle einzelnen Tests ebenfalls zum Signifikanzniveau α durchgeführt werden.

Eine mögliche solche Testprozedur besteht darin, alle zu prüfenden Hypothesen und die Reihenfolge ihrer Testung vorab (im Studienprotokoll) festzulegen und nur solange zu testen, bis der erste Test ein nicht-signifikantes Ergebnis (zum Niveau α) liefert. Dann werden diese und alle nachfolgenden Hypothesen nicht verworfen, alle vorangegangenen Hypothesen hingegen werden verworfen (hierarchisches Testen bei **a-priori geordneten Hypothesen** (Maurer et al. (1995), Koch et al. (1996), Bauer (1991)), Spezialfall der Abschlusstest-Prozedur (Marcus et al. (1976)).

Es wäre also insbesondere vorab festzulegen, in welcher Reihenfolge die verschiedenen Subgruppenanalysen und in welcher Reihenfolge die Subgruppen-spezifischen Tests, die sich einem signifikanten Interaktionstest anschließen, durchgeführt werden sollen. Bei Durchführung nur einer Subgruppenanalyse würde diese Testprozedur so aussehen, dass zuerst der Test auf Gesamteffekt durchgeführt wird. Nur wenn dieser signifikant ist, wird der Interaktionstest durchgeführt. Nur wenn dieser signifikant ist, wird die separate Testung der Subgruppe Nr. 1 durchgeführt. Nur wenn diese ein signifikantes Ergebnis liefert, wird die separate Testung der Subgruppe Nr. 2 durchgeführt, usw. Bei Durchführung mehrerer Subgruppenanalysen ist diese hierarchische Testprozedur entsprechend „verlängert".

Der Nachteil dieser hierarchischen Testprozedur ist, dass Subgruppenanalysen, die in der Hierarchie weiter unten stehen, überhaupt nicht durchgeführt werden, wenn mindestens eine der weiter oben stehenden Hypothesen nicht verworfen wurde.

Eine grundsätzliche Möglichkeit, das Problem des multiplen Testens zu *reduzieren*, besteht darin, nur *wenige* Subgruppenanalysen durchzuführen. Dann ist das multiple Testproblem (im Sinne einer Überschreitung des vorgegebenen Signifikanzniveaus) im allgemeinen weniger stark ausgeprägt als bei Durchführung sehr vieler Subgruppenanalysen. Besonders sinnvoll ist es, sich bei Subgruppenanalysen auf wenige Subgruppen und wenige Zielgrößen, ggf. nur die Hauptzielgröße der Studie, zu beschränken.

Immer dann, wenn Subgruppenanalysen ***konfirmatorischen*** **Charakter** haben sollen, sollte neben der Beschränkung auf wenige Subgruppenanalysen zusätzlich sichergestellt sein, dass das Problem des multiplen Testens ausgeräumt ist und die Testprozedur insgesamt das vorgegebene Niveau einhält. Ist lediglich eine vorsichtige, der Orientierung dienende Interpretation der Ergebnisse von Subgruppenanalysen geplant, so ist es vertretbar, auf Korrekturen für multiples Testen zu verzichten und sich auf wenige Subgruppenanalysen zu beschränken.

3) Die Power von Subgruppenanalysen ist in der Regel gering. Deshalb sind nicht-signifikante Ergebnisse von Subgruppenanalysen nur dann interpretierbar, wenn die Power angeben ist bzw. wenn man sie (näherungsweise) berechnen kann. Idealerweise werden wichtige Subgruppenanalysen in der Fallzahlplanung der Studie berücksichtigt.

Erläuterung:

Die statistische Power von Subgruppenanalysen ist gering, d.h. die Wahrscheinlichkeit, dass ein tatsächlich vorhandener Subgruppeneffekt durch die statistische Auswertung aufgedeckt werden kann, ist gering. Folglich sind nicht-signifikante Ergebnisse von Subgruppenanalysen schwer interpretierbar: Es ist kaum zu entscheiden, ob das nicht-signifikante Ergebnis dadurch entstanden ist, dass kein Subgruppeneffekt vorhanden ist oder aber dadurch, dass die Power zu klein ist.

Beleg der Aussage:

Eine Studie (Überlegenheitsstudie) ist in der Regel so geplant, dass die Hauptauswertung der Studie mit einer vorgegebenen Wahrscheinlichkeit (Power, häufig 80%) ein signifikantes Ergebnis liefern kann, sofern in Wahrheit ein Unterschied zwischen den Therapien vorliegt. In der Fallzahlplanung werden Subgruppenanalysen in der Regel nicht berücksichtigt. Da die Anzahl von Patienten in einer Subgruppe geringer ist als in der Gesamtstudie, ist die Power, in einer *Subgruppe* einen Unterschiedes zwischen den Therapien aufdecken zu können, *kleiner* als die Power in der Hauptauswertung – vorausgesetzt, der tatsächliche Unterschied zwischen den Therapien ist in der Subgruppe genauso groß oder kleiner als in der Gesamtstudie.

Hinzu kommt, dass der Test auf Interaktion (siehe 1) eine nur geringe Power besitzt. Dies haben z.B. Brookes et al. (2001) mittels statistischer Simulationen gezeigt. Sie haben die Power für spezielle Konstellationen quantifiziert.

Das Problem der geringen Power wird noch weiter verschärft, wenn zur Korrektur des multiplen Testens eine Alpha-Adjustierung erforderlich ist (vergleiche Seite 14).

Vorgehen:

oben beschrieben (fett gedruckter Text)

4) Es sollte eine Begründung (physiologische oder biologische Rationale) für die Subgruppenbildung angegeben werden.

Erläuterung:

Damit soll die Plausibilität der Subgruppenbildung und eines möglichen Subgruppeneffektes dargelegt werden. Ferner kann eine solche Begründung als Argumentationshilfe dafür dienlich sein, dass die Subgruppenanalyse nicht datengetrieben erfolgt ist.

Beleg der Aussage:

siehe Seite 19 („Zu 4: ...“)

Vorgehen:

Eine Rationale für die Subgruppenbildung wird möglichst *vor* der Durchführung der Auswertungen dargelegt. Im idealen Fall wird sie im Studienprotokoll niedergelegt.

5) Die Subgruppen-bildende Variable muss eine Baseline-Variable sein. Andernfalls kann die Subgruppenanalyse zu verzerrten Ergebnissen führen.

Erläuterung:

Eine Variable, die keine Baseline-Variable ist, ist eine solche, die *nach* Therapiebeginn erhoben wurde.

Beleg der Aussage:

Werden die Subgruppen anhand einer Variablen gebildet, die keine Baseline-Variable ist, so kann diese Variable von der Therapie beeinflusst sein. Die Therapie könnte also beeinflusst haben, ob ein Patient in eine spezielle Subgruppe gehört oder nicht. Das kann innerhalb der Subgruppen zu Imbalancen und damit zu *nicht vergleichbaren* Therapiegruppen führen, siehe Punkt 6. Werden z.B. die beiden Subgruppen Complier / Non-Complier oder die beiden Subgruppen Responder / Non-Responder untersucht, so sind solche Imbalancen denkbar. *Ein weiterer Beleg dieser Kernaussage ist nicht erforderlich.*

Vorgehen:

Die Ergebnisse von Subgruppenanalysen sind *nicht* interpretierbar, wenn die Subgruppen-bildende Variable kein Baseline-Variable ist.

6) Selbst in randomisierten Studien können in den einzelnen Subgruppen *Imbalancen* zwischen den Therapiegruppen auftreten, besonders in kleinen Subgruppen. Wurde die Randomisierung der Studie stratifiziert nach der Subgruppen-bildenden Variablen durchgeführt, so besteht die Gefahr von Imbalancen praktisch nicht. Wurde in der Studie keine solche stratifizierte Randomisierung durchgeführt, so ist es für die Interpretation hilfreich, die Baseline-Variablen der Therapiegruppen auch für die Subgruppen gegenüberzustellen, um so das Vorhandensein von Imbalancen ausschließen zu können oder bei Bedarf multivariate Analysen durchführen zu können.

Erläuterung:

Der Zweck der Randomisierung ist es gerade, eine Strukturgleichheit zwischen den Therapiegruppen zu erzwingen, d.h. Imbalancen zu vermeiden. Durch die Strukturgleichheit wird sichergestellt, dass die Therapiegruppen „vergleichbar“ sind, so dass die Ergebnisse eines Vergleichs der Therapiegruppen ausschließlich auf die Therapie zurückzuführen sind und nicht auf andere Faktoren. Vorausgesetzt, die Subgruppen-bildende Variable wird nicht von der Therapie beeinflusst (siehe Punkt 5), ist im Prinzip die Strukturgleichheit auch in jeder Subgruppe einer randomisierten Studie wieder vorhanden, denn eine solche Subgruppe ist wieder eine kleine randomisierte Studie. Jedoch können insbesondere in kleinen Subgruppen zufällig Imbalancen auftreten. Sind Imbalancen vorhanden, können die Ergebnisse von Subgruppenanalysen verzerrt sein.

Beleg der Aussage:

siehe Seite 19 („Zu 6: ...“)

Vorgehen:

oben beschrieben (fett gedruckter Text)

7) Post-hoc-Subgruppenanalysen sind ausschließlich explorativ. Sie können der Hypothesen-Generierung dienen, sie erlauben aber keine konfirmatorischen Aussagen.

Erläuterung:

Post-hoc-Subgruppenanalysen können nicht für abschließende Therapieentscheidungen herangezogen werden. Ihre Ergebnisse müssen zuvor durch andere Studien bestätigt oder verworfen werden.

Beleg der Aussage:

siehe Seite 19 („Zu 7+8: ...“)

Vorgehen:

oben beschrieben (fett gedruckter Text)

8) A-priori-Subgruppenanalysen können konfirmatorisch sein, sofern sie korrekt durchgeführt wurden.

Erläuterung:

Im Unterschied zu post-hoc-Subgruppenanalysen können a-priori-Subgruppenanalysen Therapieentscheidungen erlauben. Voraussetzung dafür ist allerdings, dass sie so durchgeführt wurden, dass ihre Ergebnisse keiner Verzerrung unterliegen.

Beleg der Aussage:

siehe Seite 19 („Zu 7+8: ...“)

Vorgehen:

oben beschrieben (fett gedruckter Text)

In Tabelle 2 ist dargestellt, welche dieser 8 Kernaussagen in den einzelnen 38 eingeschlossenen Arbeiten erwähnt sind und ob sie dort belegt bzw. begründet wurden. In der Tabelle sind die Kernaussagen in der selben Reihenfolge angeordnet und mit den selben Nummern versehen wie in der obigen Darstellung.

Das Problem des multiplen Testens (Nr. 2) und die Notwendigkeit des Interaktionstests (Nr. 1) werden in den meisten Arbeiten thematisiert (in 29 bzw. 22 der 38 Arbeiten). Auch das Power-Problem und die Aussage, post-hoc-Subgruppenanalysen seien ausschließlich explorativ, finden sich in vielen Arbeiten (in jeweils 16 der 38 Arbeiten). Die übrigen 4 Kernaussagen sind nur in relativ wenigen Arbeiten erwähnt, das betrifft überraschenderweise auch die Aussage, dass a-priori-Subgruppenanalysen konfirmatorisch sein können.

Belege bzw. Beweise der 8 Kernaussagen finden sich in den Arbeiten kaum. Die Kernaussagen 2 (Multiples Testen), 3 (Power) und 5 (Baseline-Variable) bedürfen, wie oben erläutert, keines Beleges mehr. Für die restlichen 5 Aussagen wurde in den Publikationen folgendes gefunden:

Zu 1: Brookes et al. (2001) haben durch statistische Simulationen nachgewiesen, dass Subgruppen-spezifische Tests anti-konservativ sind. Die Überschreitung des vorgegebenen Signifikanzniveaus kann erhebliche, für die praktische Anwendung inakzeptable Ausmaße annehmen. In der Simulationsstudie wurde bei einem vorgegebenen Signifikanzniveau von 5% ein Fehler 1. Art bis zu ca. 66% beobachtet.

Ferner haben Brookes et al. (2001) in ihren Simulationen gezeigt, dass der **Interaktionstest** das vorgegebene Signifikanzniveau nicht überschreitet, also nicht anti-konservativ ist.

Hierbei haben Brookes et al. jedoch nicht die *Testprozedur* bestehend aus

- Test auf Gesamteffekt und anschließenden Subgruppen-spezifischen Tests bzw.
- Test auf Gesamteffekt und anschließendem Interaktionstest

untersucht, sondern lediglich den isolierten zweiten Schritt der Testprozedur (*nur* Subgruppen-spezifische Tests bzw. *nur* Interaktionstest). Die Tatsache, dass selbst wenn der isolierte zweite Schritt das vorgegebene 5%-Niveau einhält, die Testprozedur insgesamt das 5%-Niveau in vielen Fällen *nicht* einhält bleibt unberücksichtigt. Auf Seite 1 ihrer Arbeit beschreiben Brookes et al. zwar, dass Korrekturen für multiples Testen nicht berücksichtigt werden. Hierbei zielen sie aber auf die Durchführung *mehrerer* Subgruppenanalysen ab. Dass ein multiples Testproblem aber auch schon bei nur einer Subgruppenanalyse auftritt, wird nicht erwähnt.

Tatsächlich hält die Testprozedur bestehend aus Test auf Gesamteffekt und anschließendem Interaktionstest nur dann das vorgegebene Signifikanzniveau ein, wenn der Interaktionstest *nur* bei signifikantem Test auf Gesamteffekt durchgeführt wird.

Die in Punkt 1 (Seite 12) beschriebene Testprozedur, einen Interaktionstest durchzuführen und nur bei signifikantem Ergebnis die Subgruppen separat auszuwerten, wurde in der Simulationsstudie von Brookes et al. nicht untersucht. Deshalb liegen aus dieser Arbeit zum Fehler 1. Art der Testprozedur keine Informationen vor. Wie oben (Seite 14) erläutert, kann aber etwa durch die dort beschriebene hierarchische Testprozedur mit a-priori geordneten Hypothesen dafür gesorgt werden, dass die Testprozedur das Niveau einhält.

Zu 4: *In keiner der Arbeiten findet sich ein Beleg* für die Notwendigkeit einer **Rationale** für die Subgruppenbildung.

Dieser Aussage liegt der Gedanke zugrunde, dass ein beobachteter Subgruppeneffekt glaubwürdiger ist, wenn er sich plausibel begründen lässt. Besonderer Wert wird darauf gelegt, die Rationale *vor* Durchführung der Auswertungen festzulegen, da vermutlich davon ausgegangen wird, dass sich eine *nachträgliche* Erklärung zu praktisch jedem beobachteten Effekt angeben lässt.

Zu 6: Cui et al. (2002) zeigen anhand mathematischer Betrachtungen, dass in Subgruppen randomisierter Studien – besonders in *kleinen* Subgruppen – **Imbalancen** zwischen den Therapiegruppen auftreten können. Außerdem zeigen Lee et al. (1980) anhand einer durch Simulation generierten randomisierten Studie die Möglichkeit von Imbalancen in den einzelnen Subgruppen; eine nähere Erläuterung ist in Tabelle 4 angegeben.

Zu 7+8: Für die Aussagen zu **post-hoc- und a-priori-Subgruppenanalysen** ist *in keiner der Arbeiten eine Begründung* angeführt.

Den beiden Aussagen liegt der Gedanke zugrunde, dass nur geplante Experimente konfirmatorisch sein können. Ungeplante Experimente können immer nur der Hypothesengenerierung dienen. Ein Grund für diesen Gedanken dürfte sein, dass bei post-hoc-Subgruppenanalysen nicht nachprüfbar ist, ob sie z.B. tatsächlich nur die angegebenen Auswertungen umfassen oder ob z.B. die berichteten Ergebnisse einem umfangreichen „Data-Dredging“ entstammen. Dies ist in post-hoc-Subgruppenanalysen nur dann nachprüfbar, wenn es einen vor Auswertungsbeginn verabschiedeten Analysenplan gibt; in der Regel ist das nicht der Fall.

Eine detaillierte Darstellung der Aussagen zur Aussagekraft von Subgruppenanalysen, die aus den einzelnen eingeschlossenen Arbeiten extrahiert wurden, findet sich in Tabelle 2 bis Tabelle 6.

Tabelle 1: Kernaussagen der eingeschlossenen Arbeiten

Publikation: Erstautor (Jahr)	KERNAUSSAGEN															
	1) Interaktionstest		**2) Multiples Testen**		**3) Geringe Power**		**4) Rationale** der Subgruppenbildung		**5)** Nur **Baseline-Variablen** zur Subgruppenbildung		**6)** Möglichkeit von **Imbalancen**		**7)** Post-hoc-Sub gruppenanalysen nur ***explorativ***		**8)** A-priori-Subgru-penanalysen kön nen ***konfirmato-risch*** sein	
	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**
Abramson (1992)	×		×	entfällt		entfällt	×		×	entfällt			×			
Adams (1998)			×	entfällt		entfällt	×		×	entfällt	×		×			
Altman (1991)	×		×	entfällt		entfällt				entfällt			×			
Altman (1996)			×	entfällt		entfällt	×			entfällt						
Assmann (2000)	×		×	entfällt	×	entfällt	×		×	entfällt			×			
Brookes (2001)	×	×	×	entfällt	×	entfällt	×			entfällt			×			
Brookes (2003)				entfällt	×	entfällt				entfällt						
Bulpitt (1988)			×	entfällt	×	entfällt	×			entfällt	×					
Buyse (1989)	×		×	entfällt	×	entfällt				entfällt						
Consort Statement	×			entfällt		entfällt				entfällt						
Cui (2002)			×	entfällt	×	entfällt				entfällt	×	×				
Cuzick (1999)			×	entfällt		entfällt				entfällt			×		×	
Detsky (1995)	×		×	entfällt	×	entfällt				entfällt			×			
Freemantle (2001)	×			entfällt		entfällt				entfällt						
Furberg (1983)			×	entfällt	×	entfällt				entfällt						
ICH-E9	×			entfällt		entfällt				entfällt			×		×	
ISIS-2 (1988)				entfällt		entfällt				entfällt			×			
Kernan (1999)			×	entfällt	×	entfällt				entfällt	×					
Lee (1980)			×	entfällt		entfällt				entfällt	×	×				
Matthews (1996a)	×			entfällt		entfällt				entfällt						
Matthews (1996b)	×			entfällt		entfällt				entfällt						
Moreira (2001)	×		×	entfällt	×	entfällt				entfällt						
Moreira (2002)	×		×	entfällt	×	entfällt				entfällt						
Moyé (2001)			×	entfällt	×	entfällt				entfällt					×	
Nahler (1992)			×	entfällt		entfällt				entfällt	×		×		×	
Oxman (1992)	×		×	entfällt		entfällt	×		×	entfällt			×			
Parker (2000)	×		×	entfällt		entfällt	×			entfällt						
Parker (2002)				entfällt		entfällt				entfällt						
Peto (1995)				entfällt		entfällt				entfällt						
Pocock (1987)	×		×	entfällt		entfällt				entfällt						

Publikation: Erstautor (Jahr)	KERNAUSSAGEN															
	1) Interaktionstest		**2) Multiples Testen**		**3) Geringe Power**		**4) Rationale** der Subgruppenbildung		**5)** Nur **Baseline-Variablen** zur Subgruppenbildung		**6)** Möglichkeit von **Imbalancen**		**7)** Post-hoc-Subgruppenanalysen nur ***explorativ***		**8)** A-priori-Subgruppenanalysen können ***konfirmatorisch*** sein	
	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**	enthalten?*	Beleg?**
Pocock (1990)	×		×	entfällt		entfällt				entfällt			×			
Pocock (2002)	×		×	entfällt		entfällt				entfällt						
Simon (1982)	×		×	entfällt	×	entfällt			×	entfällt			×		×	
Simon (1988)	×		×	entfällt	×	entfällt				entfällt			×		×	
van Gijn (1999)				entfällt		entfällt				entfällt						
White (2003)	×			entfällt		entfällt				entfällt						
Yusuf (1990)			×	entfällt	×	entfällt				entfällt			×			
Yusuf (1991)	×		×	entfällt	×	entfällt	×		×	entfällt			×		×	

* ×: Die Publikation enthält eine Aussage zu diesem Aspekt.
leere Zelle: Die Publikation enthält *keine* Aussage zu diesem Aspekt.

** ×: Die Publikation enthält einen Beleg/ Beweis für diese Aussage.
leere Zelle: Die Publikation enthält *keinen* Beleg/ Beweis für diese Aussage.
entfällt: Ein Beleg/Beweis dieser Aussage ist nicht erforderlich, da die Aussage bereits bewiesen ist, siehe oben.

Hinweis: Eine genauere Darstellung der 8 Kernaussagen findet sich über der Tabelle (Seite 12 ff).

6 Diskussion

Subgruppenanalysen sind sehr gebräuchlich und finden sich in fast jeder Publikation einer klinischen Studie. Man kann 2 Typen von Subgruppenanalysen unterscheiden: a-priori- und post-hoc-Subgruppenanalysen.

Unter „a-priori-Subgruppenanalysen“ werden hier Subgruppenanalysen verstanden, die vor Studienbeginn festgelegt und üblicherweise im Studienprotokoll dargelegt wurden, sowie Subgruppenanalysen, die zwar nach Studienbeginn, aber ohne Einsicht in die bereits erhobenen Daten definiert und (üblicherweise in einem Amendment zum Studienprotokoll) festgehalten wurden. Alle anderen Subgruppenanalysen werden hier als „post-hoc-Subgruppenanalysen“ bezeichnet; dies sind in aller Regel Subgruppenanalysen, die nach Sicht auf die Daten (*datengetrieben*) durchgeführt wurden.

Im vorliegenden Gutachten wird die Aussagekraft von a-priori-Subgruppenanalysen mit der Aussagekraft von post-hoc-Subgruppenanalysen verglichen. Dazu wurde zunächst ausgehend von einer systematischen Recherche und Aufbereitung der gefundenen Literatur die Aussagekraft beider Typen von Subgruppenanalysen untersucht und anschließend ein Vergleich angestellt.

Stand der Literatur:

38 Arbeiten wurden identifiziert, die Informationen zur Aussagekraft von Subgruppenanalysen enthalten. Diese Arbeiten enthalten **8 Kernaussagen**, die im vorangegangenen Abschnitt aufgeführt wurden. Für die meisten dieser Kernaussagen werden in den 38 Arbeiten keine Belege bzw. Beweise angegeben. Dies überrascht nicht. Belege sind für Aussagen über die Eignung statistischer Verfahren zu erwarten und hierfür liegen sie auch weitgehend vor. Hingegen sind für Aussagen, wie etwa die Notwendigkeit einer Rationale, Belege bzw. Beweise kaum vorstellbar.

Überraschend ist jedoch, dass nur 7 der 38 eingeschlossenen Arbeiten die Aussage enthalten, dass **a-priori-Subgruppenanalysen konfirmatorisch** sein können, sofern sie adäquat durchgeführt wurden. Eine mögliche Erklärung hierfür könnte sein, dass diese Aussage breit akzeptiert ist und deshalb nur selten explizit formuliert wird. In den 38 Arbeiten finden sich kaum Äußerungen, die der Aussage, a-priori-Subgruppenanalysen könnten konfirmatorisch sein, widersprechen. Lediglich Yusuf et al. (1990) empfehlen, beobachtete Subgruppeneffekte, selbst wenn sie aus a-priori-Subgruppenanalysen stammen, in weiteren Studien zu prüfen.

Anforderungen an Subgruppenanalysen:

Aus den 8 Kernaussagen leiten sich unmittelbar folgende **Anforderungen an Subgruppenanalysen** ab:

- multiples Testproblem liegt nicht vor
- Interaktionstest durchgeführt
- Rationale der Subgruppenbildung angegeben
- Subgruppenbildung durch Baseline-Variable
- keine Imbalancen

Die Kernaussagen Nr. 3, 7 und 8 sind in dieser Auflistung noch nicht aufgegangen; aus ihnen ergibt sich sofort folgendes:

- idealerweise werden Subgruppenanalysen in Fallzahlplanung berücksichtigt, ansonsten Interpretation nicht-signifikanter Ergebnisse nur zusammen mit Power-Überlegungen
- post-hoc-Subgruppenanalysen nur explorativ
- a-priori-Subgruppenanalysen unter Umständen konfirmatorisch

Aus den letzten beiden Aussagen folgt insbesondere die Antwort auf die Hauptfrage des Gutachtens: **A-priori-Subgruppenanalysen besitzen höhere Aussagekraft als post-hoc-Subgruppenanalysen** – vorausgesetzt, es handelt sich um adäquat durchgeführte Subgruppenanalysen.

Neben den 8 Kernaussagen finden sich in den eingeschlossenen Arbeiten noch einige weitere Informationen zur Aussagekraft von Subgruppenanalysen:

Einige Autoren weisen darauf hin, dass bei **nicht-signifikantem Gesamteffekt** der Studie die Ergebnisse von Subgruppenanalysen mit besonderer Vorsicht zu interpretieren seien (Assmann, 2000; Cuzick, 1999; Pocock, 1990; Freemantle, 2001; Simon, 1982). Eine Begründung hierfür wird jedoch nicht angegeben.

Plausibel erscheint die Begründung, dass Autoren dann, wenn der Gesamteffekt der Studie nicht signifikant ist (d.h. die Überlegenheit der zu prüfenden Therapie konnte nicht gezeigt werden), tendenziell eine höhere Motivation haben, Patientengruppen mit einem positiven Effekt zu identifizieren und dass deshalb die Gefahr eines „Data-Dredging“ größer ist als bei einem signifikanten Gesamteffekt.

Ein weiterer Grund für diesen Hinweis könnte sein, dass eine *hierarchische Testprozedur* wie die auf Seite 14 beschriebene, bei der Subgruppenanalysen nur bei *signifikantem* Gesamteffekt durchgeführt werden, das vorgegebene Niveau α einhalten. Wird *kein* solches hierarchisches Vorgehen gewählt, ist streng genommen – und insbesondere dann, wenn die Subgruppenanalysen konfirmatorischen Charakter haben sollen – eine Alpha-Adjustierung erforderlich. Da also eine hierarchische Testprozedur nicht die einzige Möglichkeit ist, sicherzustellen, dass die Testprozedur (Test auf Gesamteffekt und Subgruppenanalysen) das vorgegebene Niveau einhält, wird hier das Vorliegen eines signifikanten Gesamteffektes *nicht* als Voraussetzung für aussagekräftige Subgruppenanalysen beurteilt und folglich *nicht* in die Liste der Anforderungen an Subgruppenanalysen aufgenommen.

Ein weiterer Grund für diese Entscheidung ist, dass eine strikte Einhaltung des Signifikanzniveaus in Subgruppenanalysen nur dann als zwingend notwendig beurteilt wird, wenn die Subgruppenanalysen konfirmatorischen Charakter haben (siehe Seite 15).

Ein weiterer Aspekt, der in den eingeschlossenen Arbeiten nicht explizit erwähnt wurde, ist zu beachten: **Subgruppenanalysen aus inadäquaten Studien** (Studien mit erheblichen Mängeln) sind selbstverständlich *nicht* aussagekräftig. Dies betrifft sowohl a-priori- als auch post-hoc-Subgruppenanalysen. Dieser Aspekt ist in die Liste der Anforderungen aufzunehmen.

Nachprüfbarkeit der Anforderungen:

Um die Aussagekraft einer Subgruppenanalyse zu beurteilen, ist es u.a. erforderlich, zu beurteilen, ob die oben genannten Anforderungen erfüllt sind. Dies wird in vielen Fällen schwierig oder unmöglich sein, denn empirische Untersuchungen haben gezeigt, dass in vielen Publikationen wichtige Informationen zu Subgruppenanalysen, wie z.B. die Anzahl durchge-

führter Tests, nicht angegeben sind (Assmann et al. (2000), Parker et al. (2000), Moreira et al. (2001)).

Ohne die Angabe der Anzahl insgesamt durchgeführter Analysen kann in der Regel nicht beurteilt werden, ob das Problem des multiplen Testens in relevantem Maße vorliegt – es sei denn, es wurde für multiples Testen korrigiert bzw. eine hierarchische Testprozedur verwendet. Hierfür ist es notwendig, die Anzahl sämtlicher Subgruppenanalysen – inklusive derjenigen, die nicht in die Publikation aufgenommen wurden – zu kennen.

Fehlt in einer Studienpublikation die Angabe eines Interaktionstests, so kann dieser Test in aller Regel *nicht* anhand der in der Publikation angegebenen Daten durchgeführt werden, da hierzu die Daten der einzelnen Patienten erforderlich wären.

Die Interpretation eines nicht-signifikanten Ergebnisses einer Subgruppenanalyse setzt die Kenntnis der zugehörigen Power voraus. Wurde die Subgruppenanalyse in der Fallzahlplanung der Studie berücksichtigt – dies ist in der Regel nicht der Fall –, so kann die Power aus der Beschreibung der Fallzahlplanung abgelesen werden. Andernfalls ist die angemessene Interpretation dann möglich, wenn in der Studienpublikation zumindest grobe Power-Überlegungen angestellt wurden. Auch dies ist jedoch in der Regel nicht der Fall. Liegen keine Power-Angaben vor, so kann versucht werden, eigene grobe Power-Berechnungen durchzuführen. Ist auch das nicht möglich – das wird meist der Fall sein –, so ist keine oder höchstens eine vorsichtige, ausschließlich orientierende Interpretation der nicht-signifikanten Ergebnisse möglich.

Limitationen:

Die hier durchgeführte Recherche kann nur eingeschränkt einen Anspruch auf Vollständigkeit erfüllen. Informationen über die Aussagekraft von Subgruppenanalysen können sich in sehr vielfältigen Publikationen finden, beispielsweise auch in Arbeiten, die sich allgemein mit der Methodik klinischer Studien befassen. Ist in solchen Arbeiten der Begriff „Subgruppenanalyse“ (oder ähnlicher Begriff oder auf englisch, vergleiche Suchstrategie in 9.2) weder im Abstract, noch im Titel, noch in den Schlagworten enthalten, so werden diese Arbeiten durch die computerbasierte Recherche nicht gefunden. Um dieses Problem zu beseitigen, hätte die computerbasierte Recherche extrem breit durchgeführt werden müssen, wodurch sich der Bearbeitungsaufwand unangemessen erhöht hätte. Mit dem Ziel, die Recherche auf effizienterem Wege zu verbessern, wurden die Literaturverzeichnisse der gefundenen Arbeiten auf weitere relevante Quellen hin geprüft.

7 Fazit

Zusammenfassend lassen sich auf der Basis der internationalen methodischen Literatur die folgenden Aussagen schlussfolgern:

- **Subgruppenanalysen sind nur dann aussagekräftig und ihre Ergebnisse nur dann interpretierbar, wenn jedes der folgenden 4 Kriterien erfüllt ist:**

  (1) Die Subgruppenanalyse gehört zu einer adäquat durchgeführten, aussagekräftigen Studie.

  (2) Es wurden nur wenige Subgruppenanalysen durchgeführt und möglichst eine Korrektur für multiples Testen vorgenommen (bzw. eine hierarchische Testprozedur, die das Niveau einhält, angewendet).

  (3) Eine adäquate statistische Methodik (insbesondere Interaktionstest, kein alleiniges separates Auswerten der Subgruppen) wurde angewendet.

  (4) Die Subgruppen-bildende Variable ist eine Baseline-Variable.

- **Darüber hinaus wird die Aussagekraft von Subgruppenanalysen von den folgenden Kriterien beeinflusst:**

  - Die Subgruppenanalyse ist *nicht* das Ergebnis eines umfangreichen „Data-Dredging“, sondern wurde zumindest *vor* Sicht auf die Daten *geplant*. Optimal ist eine *a-priori-Subgruppenanalyse*, die bereits im Studienprotokoll oder in einem Amendment zum Studienprotokoll festgelegt wurde.
  - Eine Rationale für die Subgruppenbildung (biologische, physiologische, medizinische Plausibilität) wurde vor der Durchführung der Auswertungen dargelegt.
  - Es gibt keine Anhaltspunkte für Imbalancen zwischen den Therapiegruppen innerhalb der einzelnen Subgruppen. Optimal ist es, wenn die Randomisierung der Studie *stratifiziert* nach der Subgruppen-bildenden Variable durchgeführt wurde.

  Je mehr dieser Kriterien erfüllt sind, umso größer ist die Aussagekraft der Subgruppenanalyse – vorausgesetzt, die oben genannten Kriterien (1) - (4) sind erfüllt.

- Für einige der 7 genannten Kriterien wird häufig anhand der Studienpublikation *nicht* beurteilt werden können, ob sie erfüllt sind.

- Nicht-signifikante Ergebnisse von Subgruppenanalysen sind immer nur in Verbindung mit **Power**-Überlegungen interpretierbar – es sei denn (und das ist der optimale Fall), die Studie wurde so geplant, dass sie auch für die betreffende Subgruppenanalyse ausreichende Power besitzt.

- In der methodischen Literatur besteht weitgehend Konsens darüber, dass

  - post-hoc-Subgruppenanalysen ausschließlich **explorativer Natur** sind und deshalb lediglich der Hypothesengenerierung dienen können, aber keine Therapieentscheidungen erlauben.
  - a-priori-Subgruppenanalysen **konfirmatorisch** sein können, d.h. dass aus ihren Ergebnissen Therapieentscheidungen abgeleitet werden können. Die Voraussetzung hierfür ist, dass die obigen Kriterien (1) - (4) erfüllt sind und dass ferner die gesamte

Testprozedur (bestehend aus Test auf Gesamteffekt *und* Subgruppenanalysen) das vorgegebene Signifikanzniveau einhält.

- **Insbesondere besteht also in der methodischen Literatur Konsens darüber, dass a-priori-Subgruppenanalysen höhere Aussagekraft besitzen als post-hoc-Subgruppenanalysen – vorausgesetzt, die Subgruppenanalysen wurden korrekt durchgeführt** (Kriterien (1) - (4) erfüllt).

- Natürlich können auch a-priori-Subgruppenanalysen Mängel aufweisen. Dann ist ihre Aussagekraft nicht zwingend höher als die von post-hoc-Subgruppenanalysen.

- Auch wenn die Aussagekraft von post-hoc-Subgruppenanalysen geringer ist als die von a-priori-Subgruppenanalysen (bei adäquater Durchführung), können post-hoc-Subgruppenanalysen dennoch sinnvoll und nützlich sein.

8 Review

Das Gutachten wurde einem Review-Prozess unterzogen, der Reviewer war Herr PD Dr. med. S. Lange (Abteilung Medizinische Informatik, Biometrie und Epidemiologie, Ruhr Universität Bochum). Es wurde schriftlich dokumentiert und begründet, welche Änderungen sich aus den einzelnen Kommentaren des Reviewers ergaben. Dieses Schriftstück wird zusammen mit dem Review beim Fachbereich Evidenz-basierte Medizin des MDS aufbewahrt.

9 Anhang

9.1 Die aus den eingeschlossenen Arbeiten extrahierten Aussagen

Tabelle 2: Eingeschlossene Arbeiten, die empirische Untersuchungen enthalten

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
Assmann (2000)	**empirische Untersuchung** an 50 RCTs aus jeder Publikation wurde extrahiert: ob Subgruppenanalysen berichtet wurden, für wie viele Subgruppen-bildende Faktoren, für wie viele Zielgrößen, wie viele Subgruppenanalysen insgesamt berichtet wurden, welche statistische Methodik (nur deskriptiv, P-Werte für Subgruppen, Interaktionstest), ob Subgruppenunterschiede behauptet wurden, ob solche Behauptungen in Abstract oder Fazit standen. • Subgruppenbildung nur durch Baseline-Variablen • **Interaktionstest**; inadäquat ist das alleinige Berichten der P-Werte in den einzelnen Subgruppen (kann ungerechtfertigte Effekte produzieren) • Selbst bei signifikantem Interaktionstest besteht das multiple Testproblem • **multiples Testproblem**; oft ist unklar, wie viele Subgruppenanalysen durchgeführt wurden – vermutlich werden häufig nur diejenigen Subgruppenanalysen mit interessantem Ergebnis berichtet • dienen der Hypothesengenerierung (die in zukünftigen Studien geprüft werden können), nicht dazu, Handlungsempfehlungen / Therapieentscheidungen o.ä. abzuleiten • **Power-Mangel** in meisten Studien, um Subgruppen-Effekte aufzudecken (z.B. um Subgruppen zu identifizieren, die besonders von Therapie profitieren) • Ergebnisse aus Subgruppenanalysen sind explorativ und sollten nur in Ausnahmefällen die Schlussfolgerungen einer Studie beeinflussen • besondere Vorsicht bei Interpretation eines Subgruppen-Effektes, **wenn Gesamteffekt nicht signifikant** (Subgruppen-Effekt ist dann oft ungerechtfertigt, es sei denn die Evidenz ist statistisch überzeugend und klinisch sinnvoll) • Glaubwürdigkeit von Subgruppenanalysen ist verbessert, falls auf Hauptzielgröße beschränkt, auf wenige prädefinierte Subgruppen und auf der Basis biologisch plausibler Hypothesen • Die **Aussagekraft hängt ab von: adäquate Auswertung (Interaktionstest), biologische Plausibilität, Anzahl durchgeführter Subgruppenanalysen, Präspezifizierung der Subgruppenanalysen.**
Moreira (2001)	Es geht um Subgruppenanalysen ***in RCTs***. **Empirische Untersuchung** an 32 RCTs. Diese Studienpublikationen wurden daraufhin durchgesehen, wie häufig spezielle Aspekte von Design und Auswertung von Subgruppenanalysen berichtet werden. Aus jeder Publikation wurde extrahiert: - beschrieben, ob a-priori- oder post-hoc-Subgruppenanalyse - beschrieben, wie viele Subgruppen analysiert wurden - Begründung für Subgruppen-Definition - **Beschrieben, ob Subgruppen vor oder nach Randomisierung definiert wurden** - Statistische Methoden für Subgruppenanalysen beschrieben - Info zur Power (Fallzahlplanung oder Größe der aufdeckbaren Differenz) - Info zur klinischen Bedeutung der Interaktion - Infos zum Gesamteffekt der Studie Ergebnis: Darstellung von Subgruppenanalysen in RCTs ist schlecht und muss verbessert werden. Als Hilfe kann eine Liste ähnlich der obigen Aufzählung dienen. Als besonders wichtig wird bewertet: - Beschreiben, ob *a-priori oder post-hoc* - *Anzahl Subgruppen* darlegen, da sonst Problem des ***multiplen Testens*** nicht beurteilbar. ***Adäquate Auswertung nur in wenigen Studien*** *(Interaktionstest* anstatt separate P-Werte)

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
Parker (2000)	Subgruppenanalysen ***von RCTs*** **Empirische Untersuchung** an 67 großen RCTs in kardiovaskulärer Pharmakotherapie Ziel: Untersuchen, wie in solchen großen RCTs Subgruppenanalysen durchgeführt bzw. berichtet wurden. • Subgruppenanalysen können sowohl falsch-positive Ergebnisse (unwirksame oder schädliche Therapie wird für Subgruppen von Patienten als von Nutzen behauptet) als auch falsch-negative Ergebnisse (wirksame Therapie wird Subgruppe von Patienten vorenthalten) erzeugen. • Um diese beiden Fehlermöglichkeiten zu reduzieren, empfehlen Methodiker: – **wenige Analysen** – **vorab festlegen** – **Interaktionstest** • Meist wird Rationale (biol.) nicht dargestellt. • Oft post-hoc definiert • oft Anzahl nicht angegeben • meist kein Interaktionstest berichtet • Wenn Overall-Effekt nahe legt, dass Therapie unwirksam, dann müssen Interaktionen besonders vorsichtig interpretiert werden → nur als Hypothesengenerierung, möglicherweise neue Studie bei den Patienten mit dem möglichen Benefit • meist univariat • meist keine Angabe, ob post-hoc oder a-priori geplant • Empfehlung: Subgruppenanalysen ignorieren, wenn Interaktionstest nicht-signifikant oder nicht durchgeführt • Empfehlung: **univariate Auswertung nur, wenn pathophysiologische Rationale, sonst multivariat**
Pocock (1987)	**Empirische Untersuchung** an 45 Studienpublikationen (vergleichende Studien, sowohl randomisierte als auch nicht-randomisierte); Ergebnisse: • mindestens 1 Subgruppenanalyse: 23/45 (51%) Publikationen • beschriebene Methodik: 16/23 P-Werte der separaten Subgruppen (SCHLECHT), 4/23 deskriptive Statistik, 3/23 Interaktionstests • dargelegt, dass Subgruppen vorab definiert wurden: 0/23 • mehr als 1 progn. Faktor in Subgruppenanalysen betrachtet: 10/45 • Hinweise: auch bei Durchführung vieler Interaktionstests tritt multiples Testproblem auf; Probleme: nicht vorab festgelegte Subgruppen + nur separate P-Werte
Pocock (2002)	**selbe empirische Untersuchung wie bei ASSMANN (2000)**; Ergebnisse: • zu oft verwendet • oft überinterpretiert • meist keine adäquate statistische Methodik (**Interaktionstest**) • meist **zu viele Analysen** • in Publikation meist **unklar, ob vorab definiert oder post-hoc**

Tabelle 3: Eingeschlossene Arbeiten, die statistische Simulationsstudien enthalten

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
Brookes (2001)	Es geht um Subgruppenanalysen in RCTs. Es wurden statistische Simulationen durchgeführt, Ziel der Simulationsstudie: Untersuchung der beiden statistischen Methoden 1) separates Auswerten der Subgruppen (subgruppen-spezifische Tests) 2) Test auf Interaktion hinsichtlich Fehler 1. Art und Power • Subgruppen-spezifische Tests (Tests für die Subgruppen separat) sind anti-konservativ. Das Ausmaß der Niveauüberschreitung ist in Situationen mit signifikantem Overall-Effekt besonders ausgeprägt. Auch bei nicht-signifikantem Overall-Effekt sind diese Tests anti-konservativ. • Subgruppen-spezifische Effekte (Subgruppen separat ausgewertet) können hilfreich sein bei der Interpretation, aber sie sollten nur NACH einem **Interaktionstest** angewendet werden. • Die Power des Interaktionstests ist gering. • Idealerweise beschränken sich Subgruppenanalysen auf **vorab geplante**, die von klinischem Interesse sind, um „Data Dredging" zu vermeiden. • Es sollte klar unterschieden werden zwischen vorab festgelegten Subgruppenanalysen und post-hoc-Subgruppenanalysen. • Die Ergebnisse von Subgruppenanalysen sollten nicht überinterpretiert / überbewertet werden. Sie werden am besten als Mittel zur Hypothesen-Generierung betrachtet – außer wenn es eine starke a-priori-Hypothese für einen Subgruppeneffekt gibt. • Eine nicht-signifikante Subgruppenanalyse (nicht-signifikanter Interaktionstest) sollte nicht als Abwesenheit eines Subgruppeneffektes interpretiert werden. Der Grund hierfür ist die nicht ausreichender **Power** – es sei denn, die Studie wurde entsprechend geplant (Fallzahlplanung bzgl. Subgruppenanalyse). Idealerweise werden die Studien so geplant, dass mögliche Subgruppeneffekte mit ausreichender Power aufgedeckt werden können. Häufig ist dies aber nicht realisierbar, da es die Fallzahl extrem erhöhen würde. • **Interaktionstest** unerlässlich • Fälschlicherweise behauptete Subgruppeneffekte können bei Subgruppen-spezifischen Tests (separat) sehr häufig sein, besonders bei signifikantem Overall-Effekt. • Subgruppen-Effekte sollten nur mit Blick auf die **biologische Plausibilität** und **Untermauerung durch andere Studien** o.ä. (→ Literatur) interpretiert werden. • Korrektur für **multiples Testen** (wird im Bericht nicht berücksichtigt) • möglichst Beschränkung auf Hauptzielgröße, **wenige** vorab definierte Subgruppen auf Basis biologisch **plausibler** Hypothesen. • Ergebnisse sind als explorativ zu betrachten und können nur in Ausnahmefällen die Schlussfolgerungen einer Studie beeinflussen
Brookes (2003) [Tagungs-Abstract]	• **Power von Interaktionstests ist sehr gering**. Sie ist nur adäquat, wenn die Interaktion 2-mal so groß ist wie der Overall-Effekt oder größer. Für realistischere Interaktionen liegt die Power unter 50%. (Beleg der Aussage offenbar durch statistische Simulationen)

Tabelle 4: Eingeschlossene Arbeit, die eine durch Simulation generierte Studie **enthält**

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
Lee (1980)	Simulation: Es wurde eine RCT simuliert, wobei der wahre Therapieeffekt in dieser RCT null ist. • Ziel der Untersuchung ist es, die Probleme von Subgruppenanalysen zu illustrieren • **Auf der Basis echter Patientendaten wurde eine RCT simuliert**: Es wurde eine Datenbank von 1073 KHK-Patienten, die eine mindestens 75%-ige Okklusion an einer oder mehreren Koronararterien haben und zwischen 08/1969 und 01/1977 an dem „Duke University Medical Center" behandelt wurden, verwendet. Diese Patienten wurden auf 2 Gruppen randomisiert, die beiden so erhaltenen Gruppen wurden als „2 Therapiearme" angesehen – in Wahrheit

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
	wurden die Patienten natürlich keiner Therapie zugewiesen. Die Randomisierung erfolgte stratifiziert nach 2 Faktoren. Zielgröße war die Überlebenszeit, diese war in der verwendeten Datenbank enthalten. • Ergebnisse: Erwartungsgemäß wurde beim Vergleich der beiden Gruppen hinsichtlich Überlebenszeit kein Unterschied gefunden. Subgruppenanalysen wurden durchgeführt anhand der Stratifizierungsvariablen der Randomisierung. Dabei wurde in 1 Subgruppe ein signifikanter Unterschied in der Überlebenszeit zwischen den beiden Gruppen gefunden (**falsch-positives** Ergebnis). Gleichzeitig wurden in dieser Subgruppe auch Unterschiede in einigen Baseline-Variablen gefunden. • Fazit: Subgruppenanalysen können falsch-positive Ergebnisse liefern. Ursachen sind **multiples Testen** und **Imbalancen** in Baseline-Variablen innerhalb der Subgruppe(n). Um mit Imbalancen adäquat umzugehen, sind multivariate Methoden univariaten vorzuziehen.

Tabelle 5: Eingeschlossene Arbeiten, die mathematische Begründungen für ihre Aussagen enthalten

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
Cui (2002)	• Bei Subgruppenanalysen in RCTs kann es passieren, dass die Therapiegruppen innerhalb der Subgruppen nicht vergleichbar sind (**Imbalancen**, Selektionsbias). Besonders dann, wenn die Randomisierung nicht stratifiziert nach dem Subgruppenfaktor durchgeführt wurde oder wenn die Subgruppen klein sind, können solche Imbalancen auftreten. Beleg: mathematische Berechnungen • Problem des **multiplen Testens**: Bei im Protokoll präspezifizierten Subgruppenanalysen kann eine Adjustierung gemacht werden, sofern die Anzahl der Analysen bekannt ist. Bei post-hoc-Subgruppenanalysen ist eine Adjustierung i.a. nicht möglich, da unbekannt ist, wie viele Analysen durchgeführt wurden. • **Power-Problem**: In Subgruppenanalysen ist durch die geringe Patientenzahl häufig die Schätzgenauigkeit bzw. die Power gering. Ausnahme: wenn bei Studienplanung auch für die Subgruppenanalysen eine Fallzahlplanung gemacht wurde. • Wichtigstes dieser 3 Probleme: Imbalancen • Auch Subgruppenanalysen, die im Protokoll vorab spezifiziert wurden, müssen nicht valide sein. Wichtig: stratifizierte Randomisierung, ausreichende Patientenzahl, Korrektur für multiples Testen, ausreichende Power • Konsistenz von Subgruppenergebnissen ist keine Garantie für valide Ergebnisse.

Tabelle 6: Eingeschlossene Arbeiten, die keine Belege bzw. Begründungen für ihre Aussagen enthalten

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
Abramson (1992)	Es geht um ***post-hoc-Subgruppenanalysen*** • wichtig: physiologische **Rationale** für Subgruppenbildung • Subgruppenbildung durch Baseline-Variablen ist OK. Subgruppenbildung durch Variablen, die nach Randomisierung erhoben wurden, ist kaum interpretierbar und kann extrem verzerrte Ergebnisse liefern, da die Therapie diese Variable beeinflusst haben könnte. • sollten als post-hoc-Subgruppenanalysen gekennzeichnet werden • geeignete Auswertung wichtig (**Korrektur für multiples Testen**, **Test auf Interaktion**) • Adäquate post-hoc-Subgruppenanalysen sind sinnvoll, um Hypothesen zu generieren, die in nachfolgenden Studien geprüft werden.
Adams	Es geht um ***post-hoc-Subgruppenanalysen***

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
(1998)	• multiples Testproblem • **Imbalancen** in den Subgruppen • pathophysiologische / biologische **Rationale** der Subgruppenbildung wichtig • prospektiv geplante Subgruppenanalysen sind wünschenswert • Subgruppenbildung durch **Baseline**-Variablen ist OK • dienen der Generierung von Hypothesen, die in zukünftigen Studien geprüft werden können • erlauben keine abschließende Entscheidung
Altman (1991) [Buch]	• Es ist vernünftig, eine **kleine Anzahl** von Subgruppenanalysen durchzuführen, sofern sie **im Studienprotokoll spezifiziert** wurden. (Seite 466) • Keinesfalls sollten die Daten auf viele verschiedene Arten ausgewertet werden in der Hoffnung, einen signifikanten Unterschied zu finden. (Seite 466) • Es ist keine valide Auswertungsmethode, die Subgruppen separat auszuwerten und anschließend die P-Werte oder Konfidenzintervalle der Subgruppen zu vergleichen. Der korrekte Ansatz ist, die **Interaktion** (zwischen der Therapie und der Variable, die die Subgruppen definiert) zu untersuchen. (Seite 467, 486) • Subgruppenanalysen sind Analysen von Beobachtungsstudien ähnlich, deshalb kann aus ihnen **keine *Kausalität*** eines Zusammenhanges geschlossen werden. (Seite 467) • Subgruppenanalysen können zu irreführenden Ergebnissen kommen. (Seite 472)
Altman (1996)	• Es ist vernünftig, nicht jede mögliche Interaktion zu untersuchen. • Wenn es hingegen eine spezielle a-priori-Vermutung für eine Interaktion gibt, sollte diese auch untersucht werden. • Es ist wichtig, immer anzugeben, **wann und warum Subgruppen gewählt wurden**. Studien, die Subgruppenanalysen ohne diese Angaben berichten, sind schwer interpretierbar. • Studien, bei denen die Subgruppen-Definitionen **datengetrieben** erfolgten, sollten für multiple Tests korrigieren. Selbst dann sollten sie noch mit Skepsis behandelt werden, solange die Ergebnisse nicht in weiteren Studien bestätigt sind.
Bulpitt (1988)	• Es geht um Subgruppenanalysen in RCTs • Die Interpretation von Subgruppenanalysen ist stark davon abhängig, ob die Analysen vorab festgelegt und die Studie entsprechend stratifiziert randomisiert wurde. • Plausibilität der Subgruppenbildung wichtig • multiples Testproblem • geringe Power wegen geringer Patientenzahl • Es werden 4 Kriterien angegeben, wann Subgruppenanalysen durchgeführt werden sollten: 1. wenn sie **vor Studienbeginn festgelegt** wurden 2. wenn kein **Bias** vorliegt (Bias kann z.B. dann vorliegen, wenn die Subgruppen-Variable vom Outcome abhängig ist (z.B. Responder)) 3. wenn die Subgruppenanalyse **biologisch plausibel** ist 4. wenn der **Gesamteffekt der Studie positiv ist**. Begründungen für diese 4 Kriterien werden nicht gegeben.
Buyse (1989)	• geringe Power von Interaktionstests • P-Wert-Vergleiche sind irreführend. • Das Problem des multiplen Testens tritt in Subgruppenanalysen auf.
Cuzick (1999)	• Korrektur für **multiples Testen** erforderlich • **Geplante Subgruppenanalysen sind aussagekräftiger als post-hoc-Subgruppenanalysen.** • Man sollte bei Subgruppen-Ergebnissen besonders skeptisch sein, wenn der **Gesamteffekt nicht signifikant** ist.
Detsky (1995)	• Der **Vergleich der P-Werte** aus den separaten Analysen der beiden komplementären Subgruppen ist inadäquat. Ein Test auf Interaktion ist adäquat. • Subgruppenanalysen besitzen bedingt durch die begrenzte Anzahl Patienten häufig **geringe Power**. • Die Subgruppen-Variablen müssen *vor* Studienbeginn festgelegt werden.

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
	• Post-hoc-Subgruppenanalysen (Analysen nach Sicht auf die Daten) können der Hypothesengenerierung für zukünftige Studien dienen, aber nicht der Hypothesen-Testung. • Das Problem des multiplen Testens tritt in Subgruppenanalysen auf, in vorab spezifizierten ebenso wie in post-hoc-Subgruppenanalysen.
Freemantle (2001)	• Adäquate statistische Methodik für Subgruppenanalysen: Schätzer mit KI für Differenz der Schätzer der beiden Subgruppen (Pocock & Hughes, 1990) • Für individuelle Patienten können Subgruppenanalysen u.U. die besten verfügbaren Ergebnisse sein. Für Gesundheitssystem-Entscheidungen ist es anders. • Aus puristischer Sicht sollten Subgruppenanalysen dann, **wenn der Gesamteffekt *nicht* positiv ist**, lediglich als hypothesen-generierend angesehen werden. • Subgruppenanalysen sollten möglichst präspezifiziert und nicht datengetrieben sein.
Furberg (1983)	zur Illustration wird ein Studien-Beispiel (BHAT-Studie) verwendet, kein allgemeingültiger Beleg angegeben • Die Subgruppen sollten möglichst **vorab festgelegt** werden. • Problem des **multiplen Testens** • Power-Problem • **Besonders die Ergebnisse kleiner Subgruppen sind mit Vorsicht zu interpretieren.** • Die Ergebnisse von Subgruppenanalysen sind glaubwürdiger, wenn – der Effekt plausibel ist – der Effekt **konsistent** ist mit anderen in der selben Studie beobachteten Effekten – der Effekt durch andere Studien bestätigt ist/wird.
ICH-E9	• In einigen Fällen werden Subgruppeneffekte erwartet oder sind von besonderem Interesse, dann sind die entsprechenden Analysen Teil der **geplanten konfirmatorischen** Analyse. • In den meisten Fällen sind Subgruppenanalysen **explorativ** und sollten entsprechend gekennzeichnet werden. • zuerst **Interaktionstest**, dann Analyse der relevanten Subgruppen • Explorative Subgruppenanalysen sollten mit Vorsicht interpretiert werden. • **Explorative Subgruppenanalysen alleine erlauben i.R. keine Schlussfolgerungen zur Wirksamkeit oder Sicherheit.**
ISIS-2 Collaborative Group (1988)	Am Beispiel der ISIS-2 Studie wird demonstriert: • **Selbst in sehr großen Studien** wie ISIS-2 (insgesamt 17187 Patienten) ist es **unwahrscheinlich, dass *zuverlässig*** Subgruppen von Patienten identifiziert werden können, die besonders von der Therapie profitieren oder bei denen die Therapie unwirksam ist. • Werden in einer Studie mit deutlichem positivem Gesamteffekt viele Subgruppenanalysen durchgeführt, so sind in einigen Subgruppen falsch-negative Ergebnisse zu erwarten. Dies wird anhand der Sternzeichen-Subgruppenanalyse der ISIS-2-Studie (siehe Darstellung bei Peto (1995)) demonstriert. • Subgruppenanalysen können irreführend sein.
Kernan (1999)	• Subgruppenanalysen haben oft schlechte Power und können dadurch einen vorhandenen Therapieeffekt übersehen. (**hoher Fehler 2. Art** bzw. geringe Power) • Subgruppenanalysen können zu einem **erhöhten Fehler 1. Art** führen (multiples Testproblem). • Sie sind wichtig zur Hypothesengenerierung und zum Management individueller Patienten. • **Power von Subgruppenanalysen sollte berichtet werden**, um so dem Power-Problem Rechnung zu tragen. • Richtig durchgeführt und interpretiert können Subgruppenanalysen nützlich sein. • Subgruppenanalysen **sollten vor Studienbeginn definiert** werden. Dadurch wird die Möglichkeit, dass ein gefundener Subgruppeneffekt nur eine Folge eines Fehler 1. Art ist, reduziert. Außerdem hat der Leser dadurch mehr Vertrauen, dass die Subgruppen auf der Basis biologischer Gründe oder Vorbeobachtungen definiert wurden. • Bei **stratifizierter Randomisierung** können für die entsprechenden Subgruppen (Stratifizie-

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
	rungsvariable) valide Schlüsse aus Subgruppenanalysen gezogen werden
Matthews (1996a)	• Es werden 2 Beispiele dafür gegeben, dass der Vergleich der beiden P-Werte separater Analysen der beiden Subgruppen irreführend sein kann. Dieses Vorgehen (**P-Wert-Vergleich**) ist nicht korrekt.
Matthews (1996b)	• Interaktionstest wird an den 2 Beispielen aus Matthews (1996a) demonstriert • ebenso Differenz der beiden Therapieeffekte und zugehöriges Konfidenzintervall
Moreira (2002)	• Ergebnisse von Subgruppenanalysen in aussagekräftigen RCTs müssen nicht valide sein. • **Ergebnisse nur interpretierbar, wenn eine Reihe von Infos vorliegen (s. oben), diese fehlen jedoch in den meisten Publikationen.** • Subgruppenanalysen sind grundsätzlich sinnvoll und notwendig. • Separate Auswertungen extrem irreführend, Interaktionstest erforderlich • **Multiples Testproblem** • **Power-Mangel**
Moyé (2001)	• **Es wird eine Methode für prospektiv geplante Subgruppenanalysen vorgeschlagen**; diese Subgruppenanalysen erlauben **konfirmatorische** Aussagen. • Methode: Neben der Gesamt-Studienpopulation wird eine Subgruppe untersucht. (Die komplementäre Subgruppe wird nicht untersucht → Unterschied zum Interaktions-Ansatz) Vorgehen: ➢ Das Signifikanzniveau wird auf die Hauptanalyse und diese Subgruppenanalyse verteilt. (Verteilung muss nicht gleichmäßig sein – nur so, dass Fehler 1. Art insgesamt stimmt.) ➢ Fallzahlplanung für Gesamtpopulation und auch für die Subgruppe. Dabei sind für die Fallzahlplanung der Subgruppenanalysen andere Annahmen als für Gesamtgruppe möglich, z.B. höhere Ereignis-Rate, andere Zielgröße, geringere Varianz. ➢ Es handelt sich praktisch um eine in die Gesamtstudie eingebettete prosp. geplante Subgruppenanalyse. • Besonderheit: Subgruppenanalyse kann auch dann positiven Effekt zeigen, wenn Gesamteffekt nicht positiv ist • Interaktionstest grundsätzlich nicht zwingend erforderlich, prospektive Analyse einer einzelnen Subgruppe auch möglich • Eine oder 2 Subgruppen können auf die o.g. Art prospektiv geplant und in die Studie eingebettet werden, die restlichen Subgruppen von Interesse können auf die übliche Art durchgeführt und entsprechend als explorative Analysen interpretiert werden.
Nahler (1992)	***Post-hoc-Subgruppenanalysen*** • dürfen nur zur Hypothesengenerierung herangezogen werden. „Therapeutische Aussagen“ können nur aus a-priori geplanten Subgruppenanalysen, die bei der Planung der Studie berücksichtigt wurden, abgeleitet werden. • **Multiples Testproblem** • heben die Randomisierung auf (Dies ist eine Fehleinschätzung des Autors.)

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
Oxman (1992)	Es geht um Subgruppenanalysen ***von RCTs***. • **Guidelines (Kriterien) für die Validität der Ergebnisse von Subgruppenanalysen:** Die Kriterien gelten unter der Voraussetzung, dass die Interaktion – sofern sie tatsächlich vorliegt – auch klinisch bedeutsam ist: 1) **Ausmaß der Interaktion / des Subgruppeneffektes ist klinisch relevant** 2) **Interaktion/Subgruppeneffekt ist statistisch signifikant** (Inadäquate Auswertungen sind: separate Auswertungen der Subgruppen oder gar nur das Ergebnis derjenigen Subgruppen großem Therapieeffekt berichten. Sie können zu einer Überschätzung der Signifikanz und zu einer Schätzung des Subgruppeneffektes / Therapieeffektes führen) 3) **Hypothese *vor* der Analyse** (Glaubwürdigkeit von Interaktionen, die in post-hoc Explorationen der Daten entdeckt wurden, ist fragwürdig.) 4) Es wurden **nur wenige Subgruppenanalysen** durchgeführt. (multiples Testen) 5) Es gibt *indirekte Evidenz (**biologische Plausibilität**)* für den Subgruppeneffekt. • Post-hoc-Subgruppenanalysen sollten als Mittel zur Hypothesengenerierung angesehen werden, nicht als Mittel zur Hypothesentestung. • Bei Post-hoc-Subgruppenanalysen ist es wahrscheinlicher, dass die Entscheidung, welche Analysen durchgeführt und welche berichtet werden, datengetrieben erfolgt. • Wurde die Hypothese eindeutig aus *anderen* Daten erhalten, dann gewinnt die Subgruppenanalyse hypothesentestenden Charakter. • In Post-hoc-Subgruppenanalysen ist die Gefahr von Bias und damit die Gefahr irreführender Ergebnisse größer. • Ist die Subgruppen-Variable eine, die erst nach Baseline gemessen wurde, dann sind die Ergebnisse der Subgruppenanalysen extrem unglaubwürdig (da die Therapie beeinflussen könnte, ob ein Pat. in eine spezielle Subgruppe kommt oder nicht). Solche Analysen sind selbst dann unglaubwürdig, wenn sie vorab geplant wurden. • Selbst vorab geplante Subgruppenanalysen sind nicht aussagekräftig, wenn sehr viele Analysen (ohne entsprechende Korrektur) gemacht wurden (**multiples Testen**). • Leider weiß der Leser häufig nicht, wie viele Subgruppenanalysen durchgeführt wurden. Er kann so leicht irregeführt werden. • Subgruppenanalysen spiegeln häufig eher die individuelle klinische Situation wider.
Parker (2002)	• **Univariate Subgruppenanalysen** besitzen nur eingeschränkte Aussagekraft, **multivariate** Analysen können nützlich sein, zusätzlich oder als Alternative zu univariaten Analysen. • Dies wird am Beispiel von 2 gepoolten Studien (Enalapril bei Patienten mit linksventrikulärer Dysfunktion, SOLVD-Studien) demonstriert
Peto (1995)	• Datengetriebene Subgruppenanalysen können zu massivem Bias führen. • **Beispiel für Subgruppenanalyse mit falschem Ergebnis:** In der ISIS-2 Studie zur Therapie des akuten Herzinfarktes mit Aspirin, Streptokinase bzw. einer Kombination aus beiden (placebokontrollierte RCT mit insgesamt 17187 Patienten) wurde beim Vergleich von Aspirin mit Placebo ein signifikante Reduktion der 1-Monats-Mortalität beobachtet. Eine (absurde) Subgruppenanalyse, bei der die Patienten nach ihrem astrologischen Sternzeichen eingeteilt wurden, ergab folgendes: Bei Patienten mit dem Sternzeichen Steinbock ist Aspirin besonders effektiv, während es bei Patienten mit dem Sternzeichen Waage oder Zwilling keinerlei Vorteil gegenüber Placebo zeigt. (Beispiel für falsch-negatives Ergebnis einer Subgruppenanalyse)
Pocock (1990)	• Probleme bei Subgruppenanalysen: 1. kein **Interaktionstest** 2. **post-hoc Definition** der Subgruppen bzw. post-hoc Betonung interessanterer Subgruppen 3. Tendenz von Studienmachern, positive Subgruppeneffekte zu suchen, wenn der Gesamteffekt nicht positiv war • Korrekte statistische Methodik für Subgruppenanalysen: **Differenz zwischen den Effektschätzern in den beiden Subgruppen bilden und zugehöriges Konfidenzintervall be-**

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
	rechnen • Schätzer und Konfidenzintervalle für die Subgruppen separat zu berechnen ist nicht adäquat, wenn es keine Evidenz für Interaktion gibt. • Eine nachgewiesene Interaktion alleine genügt nicht. Es ist auch wichtig, zu wissen, ob mehrere (Subgruppen)analysen durchgeführt wurden und ob bei Bedarf für **multiples Testen** korrigiert wurde. • Das nachträgliche Auswählen der „signifikantesten" Subgruppenanalysen ist ein Problem. • Verschiedene Autoren haben Bayes-Ansätze für Subgruppenanalysen vorgeschlagen. Diese berücksichtigen die Probleme von Subgruppenanalysen (multiples Testen, Interaktion) und bewirken ein „shrinkage" der Effektschätzer und Konfidenzintervalle der Subgruppen in Richtung Gesamteffekt. • Subgruppenanalysen sind ein Beispiel explorativer Datenanalyse. Formal korrekte Schätzer können nicht post-hoc erhalten werden. Deshalb sind insbesondere große Subgruppeneffekte mit Vorsicht zu betrachten und sie erfordern idealerweise eine Bestätigung in einer nachfolgenden Studie.
Revised Consort Statement (Altman (2001))	Verweis auf 5 Quellen (sind alle im vorliegenden Gutachten berücksichtigt), sonst keine Begründungen • Methode für Subgruppenanalysen müssen klar dargestellt sein. • beste Methode: Interaktionstest • Eine gebräuchliche, aber unterlegene Methode ist es, P-Werte separater Analysen auf TE in einzelnen Subgruppen zu vergleichen. Es ist nicht korrekt, aus einem signifikanten und einem nicht-signifikanten P-Wert auf einen Subgruppen-Effekt zu schließen. Solche Schlussfolgerungen haben einen hohen Fehler 1. Art (hohe Wahrscheinlichkeit einer falsch-positiven Schlussfolgerung) • **Post-hoc-Subgruppenanalysen werden hier definiert als Analysen, die nach Sicht auf die Daten durchgeführt werden.** Sie sind besonders anfällig für falsche Entscheidungen und haben deshalb keine hohe Glaubwürdigkeit.
Simon (1982)	ohne Begründungen (lediglich mathematische Erläuterung des multiplen Testproblems): • **multiples Testproblem** mit mathematischer Erläuterung Umgang mit dem multiplen Testproblem: **entweder nur wenige Analysen durchführen (≤ 10) und diese vorab festlegen oder Alpha-Adjustierung** (0.05 / Anzahl Tests). Ersteres ist vorzuziehen. • Wenn der **Gesamteffekt positiv** ist und der Interaktionstest signifikant ist, dann werden separate Analysen der Subgruppen durchgeführt. Aber auch wenn der **Gesamteffekt *nicht* positiv**, aber der Interaktionstest signifikant ist, können die Subgruppen separat analysiert werden – auch wenn einige Autoren hiervon abraten. Allerdings sollten dann keine harten Schlussfolgerungen aus den Subgruppenanalysen abgeleitet werden. • **Konfidenzintervalle** sind für das separate Auswerten der Subgruppen besser geeignet als Tests. • **Separate Analysen** der Subgruppen ohne Interaktionstest können irreführend sein. • Wenn die **Subgruppen-bildende Variable keine Baseline-Variable** ist, sondern eine Variable, die erst nach Therapiebeginn erhoben wird (z.B. gute / schlechte Complier), dann besteht die Gefahr von **Imbalancen** und damit von *nicht* vergleichbaren Therapiegruppen innerhalb einer Subgruppe. • Ist die Studie so geplant, dass die Subgruppen groß genug sind, dann können Subgruppenanalysen **konfirmatorisch** sein. Meist können sie lediglich **Hypothesen generieren**, die in späteren Studien zu prüfen sind.
Simon (1988)	• vorab eine oder sehr wenige Subgruppen-Hypothesen spezifizieren • Problem des Data-Dredging • **Die Bestätigung vorab spezifizierter Hypothesen ist glaubwürdiger als die Bestätigung von Hypothesen, die aus den Daten generiert wurden.** • Falls die Haupt-Subgruppenhypothesen vorab spezifiziert werden, ist das Ausmaß des multiplen Testproblems bekannt und es kann, wenn gewünscht, dafür adjustiert werden. Andern-

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
	falls muss man versuchen, **retrospektiv zu bestimmen, wie viele Subgruppen untersucht wurden UND wie viele untersucht worden wären, wenn die „interessanten" Ergebnisse zu einem anderen Zeitpunkt gefunden worden wären** • Power-Problem bei Subgruppenanalysen → Falls Subgruppen-Hypothesen von primärem Interesse sind, sollten sie die Fallzahlplanung der Studie beeinflussen. • Wenn mehr als nur sehr wenige präspezifizierte Subgruppenanalysen durchgeführt werden sollen, ist die Erstellung eines **PROTOKOLL**s nützlich (festlegen, wie die Analysen durchgeführt werden und welche) → Dadurch ist reliable Bestimmung des Ausmaßes des multiplen Testproblems möglich und es kann dafür korrigiert werden. → Idealerweise Protokoll Teil des Studienprotokolls. Auch Hierarchie der Hypothesen benennen. • **geringe Power des Interaktionstests**, da Studien nicht so geplant sind, dass sie für Subgruppenanalysen groß genug sind • In Fällen sehr wichtiger Subgruppen und ausreichender Patientenzahlen sollte ein signifikanter Interaktionstest nicht als zwingende Voraussetzung für separate Analysen der Subgruppen gesehen werden. • Wenn kein vorab Festlegen einiger weniger Subgruppenanalysen erfolgt, kann alternativ ein „**sample splitting**" angewendet werden: Die Stichprobe wird in 2 Teile aufgeteilt. In einem Teil der Stichprobe werden Subgruppenanalysen durchgeführt und daraus Subgruppen-Hypothesen generiert. Im 2. Teil der Stichrobe werden diese Hypothesen geprüft. • Insgesamt wird die **Power**, Subgruppen-Effekte zu entdecken, klein sein – egal, welche Methodik angewendet wird.
van Gijn (1999)	• **Datengetriebene**, nicht a-priori definierte, Subgruppenanalysen können irreführend und gefährlich sein.
White (2003) [Tagungs-Poster]	offenbar Untersuchung am Beispiel, keine allgemeingültige Untersuchung • Die **Wahl des Effektmaßes** (Risikodifferenz, Relatives Risiko, Odds Ratio) kann zu unterschiedlichen Ergebnissen des Interaktionstests führen. Dies ist besonders wahrscheinlich, wenn die Ereignisraten sich zwischen den Subgruppen deutlich unterscheiden. • Deshalb sollte das Effektmaß, auf dessen Basis der Interaktionstest durchgeführt wird, vorab festgelegt werden – besonders dann, wenn die Ereignisraten zwischen den Subgruppen deutlich variieren. • Die beste Wahl des Effektmaßes für den Interaktionstest ist das Maß, für das Interaktionen am unwahrscheinlichsten sind. • Ein Bayes-Ansatz scheint robuster gegen die Wahl des Effektmaßes zu sein.
Yusuf (1990)	• Wenn Subgruppenanalysen in der einen Subgruppe einen Effekt zugunsten der Therapie und in der anderen Subgruppe zugunsten der Kontrolle zeigen, dann werden diese Ergebnisse meist in späteren Studien *nicht* bestätigt und sind also offenbar falsch. Dies illustrieren die Autoren an 2 Beispielen, eine allgemeingültige Begründung wird nicht angegeben. • niedrige **Power** • **multiples Testproblem** • Beobachtete Subgruppeneffekte sollten – ob aus a-priori- oder aus post-hoc-Subgruppenanalysen – in anderen Studien geprüft werden.
Yusuf (1991)	• Zur Subgruppenbildung sollten nur **Baseline-Variablen** verwendet werden. Variablen, die nach der Randomisierung gemessen werden, sollten nicht verwendet werden, denn die Analyse solcher Subgruppen (z.B. Responder / Nonresponder; Complier / Noncomplier) kann irreführend sein, da der Therapieeffekt die Subgruppenbildung beeinflusst haben könnte. Folglich ist unklar, ob der beobachtete Effekt in der Subgruppe tatsächlich durch die Therapie oder aber durch einen anderen Faktor bedingt ist. • In RCTs stellt jede Subgruppe wieder eine kleine RCT dar (d.h. die **Randomisierung bleibt erhalten)** – sofern die Subgruppen-bildende Variable eine Baseline-Variable ist. • **Die Autoren empfehlen, bei der Bewertung des Therapieeffektes in einer Subgruppe eher auf den Gesamteffekt der Studie zu vertrauen als auf das Ergebnis der Subgruppenanalyse.** Dies begründen sie einerseits mit dem Power-Problem in Subgruppenanalysen, das dazu führt, dass negative Ergebnisse für Subgruppen kaum interpretierbar sind und andererseits mit dem Problem des multiplen Testens, das bei inadäquater statistischer Methodik zu falsch-positiven Subgruppenergebnissen führen kann.

Erstautor (Jahr)	Extrahierte Aussagen zur Aussagekraft von Subgruppenanalysen
	• Für das **Power-Problem** in Subgruppenanalysen (bedingt durch die geringe Patientenzahl) werden 2 Beispiele angeführt: 2 Studien (GISSI, ISIS-2) mit positivem Gesamteffekt und negativem (falsch-negativem) Ergebnis der Subgruppenanalyse • multiples Testproblem • Eine **Alpha-Adjustierung** der Subgruppenanalysen wird **besonders bei Studien mit *kleinem* Gesamteffekt** empfohlen. • A-priori-Subgruppenanalysen sollten eine **Begründung** für die Subgruppenbildung enthalten. • Durch das vorab Festlegen der Subgruppen werden in der Regel weniger Subgruppen betrachtet und dadurch die Probleme von Subgruppenanalysen reduziert. • **A-priori-Subgruppenanalysen sind aussagekräftiger als post-hoc-**Subgruppenanalysen: A-priori-Subgruppenanalysen können **Hypothesen testen**, post-hoc-Subgruppenanalysen können lediglich **Hypothesen generieren**. Datengetriebene Subgruppenanalysen sollten mit großer Vorsicht interpretiert werden. • Medizinisch interessante, datengetriebene Subgruppenanalysen sollten **als post-hoc Analysen gekennzeichnet** dargestellt werden. Hierbei sollten Schätzer und **Konfidenzintervalle** angegeben werden. • **Bei datengetriebenen Subgruppenanalysen werden stringente statistische Methoden** empfohlen (Test auf Interaktion, Alpha-Adjustierung). Dies führt allerdings häufig dazu, dass Effekte nicht aufgedeckt werden können. Deshalb sollten Studien ausreichend groß geplant werden, wenn großes Interesse an einem speziellen Subgruppeneffekt besteht. • Subgruppeneffekte sollten im Kontext mit anderen verfügbaren Informationen zur Krankheit und zur Wirkungsweise der Therapie interpretiert werden.

9.2 Recherche

Recherche in DIMDI

Datum	11.09.03			
Datenbank(en)	Medline, Medline Alert, Oldmedline, Cancerlit, Ethmed, Euroethics, Gerolit, Medikat, AnimAlt-ZEBET, Kluwer-Verlagsdatenbank f. Volltexte, Springer-Verlagsdatenbank f. Volltexte, Springer PrePrint, Thieme-Verlagsdatenbank f. Volltexte, Toxline			
Anzahl der Treffer	130	Abstracts:	**129**	Volltextbeschaffung: **28**
Suchstrategie	1 2 3 4 5 6	subgroup analys? OR subgruppen#analys? OR untergruppen#analys? interpret? OR relevance? #1 AND (LA=ENGLISH OR LA=GERMAN) #2 AND (LA=ENGLISH OR LA=GERMAN) #3 AND #4 #5 AND PY=1990 to 2003		

Recherche in der Cochrane Library

Datum	28.11.03			
Datenbank(en)	Cochrane Library, Issue 4, “Cochrane Database of Methodology Reviews“ und „Cochrane Methodology Register“			
Anzahl der Treffer	26	Abstracts:	**8**	Volltextbeschaffung: **8**
Suchstrategie	1 2 3 4	subgroup analys* OR subgruppenanalyse* OR subgruppen analyse* OR untergruppenanalyse* OR untergruppen analyse* interpret* OR relevan* #1 AND #2 #1 AND #2 (1990 to current date)		

Recherche in Google

Datum	28.11.03			
Datenbank(en)	Google Suchmaschine			
Anzahl der Treffer	2	Abstracts:	**1**	Volltextbeschaffung: **0**
Suchstrategie	1	(subgruppenanalyse OR subgroupanalysis OR subgroupanalyses) AND (interpret* OR relevan*)		

9.3 Ausgeschlossene Literatur

Literatur	Ausschlussgrund
Armitage (1974)	keine Information zur Aussagekraft von Subgruppenanalysen
Bailey (1987)	keine Information zur Aussagekraft von Subgruppenanalysen
Barker (2002)	nicht zum Thema gehörend
Baujat (1999)	keine Information zur Aussagekraft von Subgruppenanalysen
Baujat (2002)	nicht zum Thema gehörend
Behrens (1995)	nicht zum Thema gehörend
Bhuta (1997)	nicht zum Thema gehörend
Blumsohn (2001)	nicht zum Thema gehörend
Borgaonkar (2000)	nicht zum Thema gehörend
Borghede (1997)	nicht zum Thema gehörend
Bristol (1997)	keine Information zur Aussagekraft von Subgruppenanalysen (statistische Methodik zur P-Wert-Adjustierung)
Büla (1999)	nicht zum Thema gehörend
Burns (1999)	nicht zum Thema gehörend
Byar (1985)	keine Information zur Aussagekraft von Subgruppenanalysen
Carolei (1996)	nicht zum Thema gehörend
Chen (2000)	nicht zum Thema gehörend
Chiou (2003)	nicht zum Thema gehörend
Christiansen (2002)	nicht zum Thema gehörend
Clarke (1997)	keine Information zur Aussagekraft von Subgruppenanalysen
Clayton (2003)	andere Fragestellung
Cleophas (2000)	nicht zum Thema gehörend
Corvò (2000)	nicht zum Thema gehörend
Counsell (1994)	keine für Einzelstudien relevanten Informationen zu Subgruppenanalysen (es geht um Subgruppenanalysen in systematischen Reviews)
de Vet (2001)	nicht zum Thema gehörend
Direct Thrombin Inhibitor Trialists' Collaborative Group (2002)	nicht zum Thema gehörend
Disch (1994)	nicht zum Thema gehörend
Dixon (1991)	keine Information zur Aussagekraft von Subgruppenanalysen (Bayes-Methode für Subgruppenanalysen wird vorgestellt)
Dixon (1992)	keine Information zur Aussagekraft von Subgruppenanalysen
Early Breast Cancer Trialists' Collaborative Group (1996)	nicht zum Thema gehörend
Early Breast Cancer Trialists' Collaborative Group (2000a)	nicht zum Thema gehörend
Early Breast Cancer Trialists' Collaborative Group (2000b)	nicht zum Thema gehörend
Elbourne (2001)	nicht zum Thema gehörend
Elmore (1998)	nicht zum Thema gehörend
Evers (2003)	nicht zum Thema gehörend
Fisher (1994)	keine Information zur Aussagekraft von Subgruppenanalysen (Beispiel für post-hoc, datengetriebene Subgruppenanalyse mit positivem Ergebnis, das in nachfolgender Studie nicht bestätigt werden konnte)
Flenady (2000)	nicht zum Thema gehörend
Flenady (2002)	nicht zum Thema gehörend
Flotte (1999)	keine Information zur Aussagekraft von Subgruppenanalysen

Literatur	Ausschlussgrund
Forbes (2001)	nicht zum Thema gehörend
Fragmin during Instability in Coronary Artery Disease (FRISC) study group (1996)	nicht zum Thema gehörend
Friedman (1992)	nicht zum Thema gehörend
Friedman (1993)	nicht zum Thema gehörend
Friedman (1994)	nicht zum Thema gehörend
Gelber (1987)	keine für Einzelstudien relevanten Informationen zu Subgruppenanalysen (es geht um Subgruppenanalysen in systematischen Reviews)
Gelber (1992a)	Quellenangaben nicht korrekt; Artikel nicht auffindbar
Gelber (1992b)	keine Information zur Aussagekraft von Subgruppenanalysen in Einzelstudien (es werden Subgruppenanalysen in systematischen Reviews kurz angesprochen)
Gershlick (1997)	nicht zum Thema gehörend
Goldberger (2001)	keine neuen Informationen (Beispiel für Subgruppenanalyse mit negativem Ergebnis, das wegen geringer Power nicht interpretierbar ist)
Goldhirsch (1992)	nicht zum Thema gehörend
Hahn (1999)	keine Information zur Aussagekraft von Subgruppenanalysen
Hahn (2000)	anderes Thema
Hall (1994)	nicht zum Thema gehörend
Hampton (1997)	nicht zum Thema gehörend
Hasenclever (2001)	nicht zum Thema gehörend
Henderson-Smart (2000)	nicht zum Thema gehörend
Higgins (2002)	keine Information zur Aussagekraft von Subgruppenanalysen (empirische Untersuchung an 28 systematischen Reviews mit dem Ziel, dazulegen, wie Subgruppenanalysen in systematischen Reviews durchgeführt und berichtet werden)
Hogan (2001)	nicht zum Thema gehörend
Hombrink (2000)	nicht zum Thema gehörend
Huenerbein (1998)	nicht zum Thema gehörend
Imrey (1992)	nicht zum Thema gehörend
Ioannidis (1998)	keine Information zur Aussagekraft von Subgruppenanalysen (vorgeschlagenes Verfahren bietet Alternative zur klassischen Subgruppenanalyse)
ISIS-1 Collaborative Group (1986)	keine Information zur Aussagekraft von Subgruppenanalysen (Beispiel-Studie)
Jenkinson (1996)	nicht zum Thema gehörend
Julian (1995)	nicht zum Thema gehörend
Kautzky-Willer (2001)	nicht zum Thema gehörend
Keavney (2000)	nicht zum Thema gehörend
Keeley (2003)	nicht zum Thema gehörend
Khoury (1992)	nicht zum Thema gehörend
Kors (1998)	nicht zum Thema gehörend
Kurth (2003)	nicht zum Thema gehörend
Lamarche (2001)	nicht zum Thema gehörend
Laule (1999)	nicht zum Thema gehörend
Linn (1995)	nicht zum Thema gehörend
Lloyd (2003)	nicht zum Thema gehörend
Lommel (2002)	nicht zum Thema gehörend
Loviscach (2000)	nicht zum Thema gehörend

Literatur	Ausschlussgrund
Magnesium in Coronaries (MAGIC) Trial Investigators(2002)	nicht zum Thema gehörend
Mahé (2001)	nicht zum Thema gehörend
Matsuo (2001)	nicht zum Thema gehörend
McLaughlin (1994)	keine Information zur Aussagekraft von Subgruppenanalysen
Midgette (1993)	nicht zum Thema gehörend
Moyé (1994)	nicht zum Thema gehörend
Moyé (2001)	nicht zum Thema gehörend
Multiple Risk Factor Intervention Trial Research Group (1982)	keine Information zur Aussagekraft von Subgruppenanalysen (Beispiel-Studie)
Nadareishvili (2002)	nicht zum Thema gehörend
Naglie (2002)	nicht zum Thema gehörend
Nicholl (2003)	keine neuen Informationen zur Aussagekraft von Subgruppenanalysen
no author (1996)	nicht zum Thema gehörend
no author (2002)	nicht zum Thema gehörend
no author (2003)	nicht zum Thema gehörend
Norman (1991)	keine neuen Informationen (Beispiel für Überinterpretation einer Subgruppenanalyse mit negativem Ergebnis)
Ober (1999)	nicht zum Thema gehörend
Osborn (2002)	nicht zum Thema gehörend
Ottenbacher (1998)	nicht zum Thema gehörend
Owens (1996)	nicht zum Thema gehörend
Packer (1996)	keine Information zur Aussagekraft von Subgruppenanalysen (schönes Beispiel für eine sauber durchgeführte und adäquat interpretierte Subgruppenanalyse, PRAISE-Studie)
Pankow (2003)	keine Information zur Aussagekraft von Subgruppenanalysen
Perez-Jimenez (2002)	nicht zum Thema gehörend
Peto (1990)	keine neuen Informationen (3 Beispiele für irreführende, inadäquate, datengetriebene Subgruppenanalysen)
Piantadosi (1993)	keine Information zur Aussagekraft von Subgruppenanalysen (Vergleich zweier statistischer Tests auf qualitative Interaktion)
Pinelli (2000)	nicht zum Thema gehörend
Pinelli (2001)	nicht zum Thema gehörend
Pocock (1979)	keine Information zur Aussagekraft von Subgruppenanalysen
Pocock (1993)	keine Information zur Aussagekraft von Subgruppenanalysen
Pollack (2002)	nicht zum Thema gehörend
PORT Meta-analysis Trialists Group (1998)	nicht zum Thema gehörend
Potischman(1992)	nicht zum Thema gehörend
Premji (2003)	nicht zum Thema gehörend
Przuntek (1999)	nicht zum Thema gehörend
Raaijmakers (1999)	nicht zum Thema gehörend
Randazzo (2003)	nicht zum Thema gehörend
Rickenbacher (1995)	nicht zum Thema gehörend
Ridker (1997)	nicht zum Thema gehörend
Rochon (1998)	nicht zum Thema gehörend
Salvesen (1999)	nicht zum Thema gehörend
Saunders (1997)	nicht zum Thema gehörend
Schmidt (1999)	nicht zum Thema gehörend

Literatur	Ausschlussgrund
Schneeweiss (2002)	nicht zum Thema gehörend
Severson (1993)	nicht zum Thema gehörend
Shuster (1983)	keine Information zur Aussagekraft von Subgruppenanalysen
Sienel (2003)	nicht zum Thema gehörend
Simon (2002)	keine Information zur Aussagekraft von Subgruppenanalysen
Skovlund (1996)	nicht zum Thema gehörend
Steiner (2000)	nicht zum Thema gehörend
Stephens (2001)	keine neuen Informationen
Sylvester (1998)	nicht zum Thema gehörend
Takakuwa (2000)	nicht zum Thema gehörend
Tonkin (2001)	nicht zum Thema gehörend
Tsai (1997)	nicht zum Thema gehörend
Valencia-Flores (1996)	nicht zum Thema gehörend
Varonen (2001)	keine Information zur Aussagekraft von Subgruppenanalysen
Waitzinger (1995)	nicht zum Thema gehörend
Wallack (1996)	keine Information zur Aussagekraft von Subgruppenanalysen (Beispiel für RCT mit Subgruppenanalyse und negativem Gesamtergebnis)
Wehrmann (1993)	nicht zum Thema gehörend
Wheatley (2003)	nicht zum Thema gehörend
Whitehead (2003)	keine Information zur Aussagekraft von Subgruppenanalysen
Williams (1998)	nicht zum Thema gehörend

10 Literaturverzeichnis

Abramson, N.S., Kelsey, S.F., Safar, P., Sutton-Tyrrell, K. Simpson's paradox and clinical trials: what you find is not necessarily what you prove. Ann Emerg Med, 1992; 21: 1480-1482

Adams, K.F.J. Post hoc subgroup analysis and the truth of a clinical trial. Am Heart J, 1998; 136: 753-758

Altman D.G. Practical statistics for medical research. Boca Raton, London, New York, Washington, D.C.: Chapman & Hall/CRC. 1991; S. 466, 467, 472, 486

Altman, D., Matthews, J.N.S. Statistics Notes: Interaction 1: heterogeneity of effects. Br Med J, 1996; 313: 486

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzsche, P.C., Lang, T. The revised CONSORT statement for reporting randomized trials: explanation and Elaboration. Ann Intern Med, 2001; 134: 663-694

Armitage, P., Gehan, E.A. Statistical methods for the identification and use of prognostic factors. Int J Cancer, 1974; 13: 16-36

Assmann, S.F., Pocock, S.J., Enos, L.E., Kasten, L.E. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet, 2000; 355: 1064-1069

Bailey, K.R. Inter-study differences: how should they influence the interpretation and analysis of results? Stat Med, 1987; 6: 351-358

Barker, B., Garcia, F.A., Warner, J., Lozerski, J., Hatch, K. Baseline inaccuracy rates for the comparison of cervical biopsy to loop electrosurgical excision histopathologic diagnoses. Am J Obstet Gynecol, 2002; 187: 349-352

Bauer, P. Multiple testing in clinical trials. Stat Med, 1991; 10: 871-890

Baujat, B., Mahé, C., Guerin, S., Pignon, J.P. Detection of outlying and influential trials: a graphical method to explore heterogeneity in meta-analysis. 7th Annual Cochrane Colloquium Abstracts, October 1999 in Rome

Baujat, B., Mahé, C., Pignon, J.P., Hill, C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. Stat Med, 2002; 21: 2641-2652

Behrens, S., Galecka, M., Brüggemann, T., Ehlers, C., Willich, S.N., Ziss, W., Dissmann, R., Adresen, D. Circadian variation of sustained ventricular tachyarrhythmias terminated by appropriate shocks in patients with an implantable cardioverter defibrillator. Am Heart J, 1995; 130: 79-84

Bhuta, T., Henderson-Smart, D.J. Elective high-frequency oscillatory ventilation versus conventional ventilation in preterm infants with pulmonary dysfunction: systematic review and meta-analyses. Pediatrics, 1997; 100: E6

Blumsohn, A., McAllion, S.J., Paterson, C.R. Excess paternal age in apparently sporadic osteogenesis imperfecta. Am J Med Genet, 2001; 100: 280-286

Borgaonkar, M., MacIntosh, D., Fardy, J., Simms, L. Anti-tuberculous therapy for maintaining remission of Crohn's disease. Cochrane Library, 2000; Issue 2

Borghede, G., Karlsson, J., Sullivan, M. Quality of life in patients with prostatic cancer: results from a Swedish population study. J Urol, 1997; 158: 1477-1485

Bristol, D.R. p-value adjustments for subgroup analyses. J Biopharm Stat, 1997; 7: 313-321

Brookes, S., Peters, T., Whitley, E., Smith, G.D., Egger, M. Subgroup analyses in randomized trials: assessing the power of the interaction test and planning sample size. Control Clin Trials, 2003; 24: 80S

Brookes, S.T., Whitley, E., Peters, T.J., Mulheran, P.A., Egger, M., Smith, G.D. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess, 2001; 5

Büla, C.J., Bérod, A.C., Stuck, A.E., Alessi, C.A., Aronow, H.U., Santos-Eggimann, B., Rubenstein, L.Z., Beck, J.C. Effectiveness of preventive in-home geriatric assessment in well functioning, community-dwelling older people: secondary analysis of a randomized trial. J Am Geriatr Soc, 1999; 47: 389-395

Bulpitt, C.J. Subgroup analysis. Lancet, 1988; July 2: 31-34

Burns, T., Creed, F., Fahy, T., Thompson, S., Tyrer, P., White, I. Intensive versus standard case management for severe psychotic illness: a randomised trial. UK 700 Group. Lancet, 1999; 353: 2185-2189

Buyse, M.E. Analysis of clinical trial outcomes: some comments on subgroup analyses. Control Clin Trials, 1989; 10: 187S-194S

Byar, D.P. Assessing apparent treatment--covariate interactions in randomized clinical trials. Stat Med, 1985; 4: 255-263

Carolei, A., Marini, C., De Matteis, G. History of migraine and risk of cerebral ischaemia in young adults. The Italian National Research Council Study Group on Stroke in the Yound. Lancet, 1996; 347: 1503-1506

Chen, C.H., Hu, H.H., Lin, Y.P., Chern, C.M., Hsu, T.L., Ding, P.Y. Increased arterial wave reflection may predispose syncopal attacks. Clin Cardiol, 2000; 23: 825-830

Chiou, C.F., Hay, J.W., Wallace, J.F., Bloom, B.S., Neumann, P.J., Sullivan, S.D., Yu, H.T., Keeler, E.B., Henning, J.M., Ofman, J.J. Development and validation of a grading system for the quality of cost-effectiveness studies. Med Care, 2003; 41: 32-44

Christiansen, M.S., Hommel, E., Magid, E., Feldt-Rasmussen, B. Orosomucoid in urine predicts cardiovascular and over-all mortality in patients with type II diabetes. Diabetologia, 2002; 45: 115-120

Clarke, M., Stewart, L. Individual patient data or published meta-analysis: a systematic review. 2nd International Conference. Scientific Basis of Health Services & 5th Annual Cochrane Colloquium, October 1997 in Amsterdam

Clayton, T., Pocock, S.J. Problems in the reporting of trials and (mis)interpretation of results: lessons from the RITA-3 Trial. Control Clin Trials, 2003; 24: 224S-225S

Cleophas, T.J., Kalmansohn, R.B. Relevance of correlation between treatment responses in clinical trials. Int J Clin Pharmacol Ther, 2000; 38: 373-380

Corvò, R., Paoli, G., Giaretti, W., Sanguineti, G., Geido, E., Benasso, M., Margarino, G., Vitale, V. Evidence of cell kinetics as predictive factor of response to radiotherapy alone or chemoradiotherapy in patients with advanced head and neck cancer. Int J Radiat Oncol Biol Phys, 2000; 47: 57-63

Counsell, C.E., Clarke, M.J., Slattery, J., Sandercock, P.A.G. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? Br Med J, 1994; 309: 1677-1681

Cui, L., Hung, H.M.J., Wang, S.J., Tsong, Y. Issues related to subgroup analysis in clinical trials. J Biopharm Stat, 2002; 12: 347-358

Cuzick, J. Interaction, subgroup analysis and sample size. In: Metabolic Polymophisms and Suspectibility to Cancer. IARC Publications No. 148. Ryder, W. (Eds.), Lyon: International Agency for Research on Cancer. 1999; 109-121

de Vet, H.C., van der Weijden, T., Muris, J.W., Heyrman, J., Buntinx, F., Knotterus, J.A. Systematic reviews of diagnostic research. Considerations about assessment and incorporation of methodological quality. Eur J Epidemiol, 2001; 17: 301-306

Detsky, A.S., Naglie, I.G. Subgroup analyses: primary and secondary. ACP J Club, 1995; May/June: A-12-A-14

Direct Thrombin Inhibitor Trialists' Collaborative Group. Direct thrombin inhibitors in acute coronary syndromes: principal results of a meta-analysis based on individual patients' data. Lancet, 2002; 359: 294-302

Disch, D.L., O'Connor, G.T., Birkmeyer, J.D., Olmstead, E.M., Levy, D.G., Plume, S.K. Changes in patients undergoing coronary artery bypass grafting: 1987-1990. Northern New England Cardiovascular Disease Study Group. Ann Thorac Surg, 1994; 57: 416-423

Dixon, D.O., Simon, R. Bayesian subset analysis in a colorectal cancer clinical trial. Stat Med, 1992; 11: 13-22

Dixon, D.O., Simon, R. Bayesian subset analysis. Biometrics, 1991; 47: 871-881

Early Breast Cancer Trialists' Collaborative Group. Ovarian ablation for early breast cancer. Cochrane Library, 2000; Issue 2

Early Breast Cancer Trialists' Collaborative Group. Ovarian ablation for early breast cancer. Cochrane Library, 2000; Issue 3

Early Breast Cancer Trialists' Collaborative Group. Ovarian ablation in early breast cancer: oberview of the randomised trials. Lancet, 1996; 348: 1189-1196

Elbourne, D.R., Prendiville, W.J., Carroli, G., Wood, J., McDonald, S. Prophylactic use of oxytocin in the third stage of labour. Cochrane Library, 2001; Issue 4

Elmore, J.R., Franklin, D.P., Youkey, J.R., Oren, J.W., Frey, C.M. Normothermia is protective during infrarenal aortic surgery. J Vasc Surg, 1998; 28: 984-992

Evers, J.L., Collins, J.A. Assessment of efficacy of varicocele repair for male subfertility: a systematic review. Lancet, 2003; 361: 1849-1852

Fisher, C.j., Dhainaut, J.F., Pribble, J., Knaus, W., IL-1ra Phase III Sepsis Syndrome Study Group. A study evaluating the efficacy of human recombinant interleukin-1 receptor antagonist (IL-1ra) in the treatment of patients with sepsis syndrome: preliminary results from a phase III multicenter trial. 1994

Flenady, V.J., Gray, P.H. Chest physiotherapy for preventing morbidity in babies being extubated from mechanical ventilation. Cochrane Library, 2000; Issue 2

Flenady, V.J., Gray, P.H. Chest physiotherapy for preventing morbidity in babies being extubated from mechanical ventilation. Cochrane Library, 2002; Issue 2

Flotte, T.J., Duncan, L.M., Lerner, L.H., Mihm, M.C.j. Tools to the trade: statistical analysis in dermatopathology articles. J Cutan Pathol, 1999; 26: 265-268

Forbes, C., Shirran, L., Bagnall, A.M., Duffy, S., ter Riet, G. A rapid and systematic review of the clinical effectiveness and cost-effectiveness of topotecan for ovarian cancer. Health Technol Assess, 2001; 5: 1-110

Fragmin during Instability in Coronary Artery Disease (FRISC) study group. Low-molecular-weight heparin during instability in coronary artery disease. Lancet, 1996; 347: 561-568

Freemantle, N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? Br Med J, 2001; 322: 989-991

Friedman, R.H. Computer based adherence interventions in clinical trial. Crisp Data Base National Institutes of Health, 1994

Friedman, R.H. Computer-based adherence interventions in clinical trials. Crisp Data Base National Institutes of Health, 1992

Friedman, R.H. Computer-based adherence interventions in clinical trials. Crisp Data Base National Institutes of Health, 1993

Furberg, C.D., Byington, R.P. What do subgroup analyses reveal about differential response to beta-blocker therapy? Circulation, 1983; 67: I-98-I-101

Gelber, R.D., Coates, A.S., Goldhirsch, A. Meta-analysis: the fashion of summing-up evidence. Part II: Interpretations and uses. Ann Oncol, 1992; 3: 683-691

Gelber, R.D., Goldhirsch, A. Interpretation of results from subset analyses within overviews of randomized clinical trials. Stat Med, 1987; 6: 371-378

Gelber, R.D., Goldhirsch, A. Reporting and interpreting adjuvant therapy clinical trials. International Breast Cancer Study Group (formerly Ludwig Group). J Natl Cancer Inst Monogr, 1992; 11: 59-69

Gershlick, A.H. Treating the non-electrical risks of atrial fibrillation. Eur Heart J, 1997; 18 suppl. C: C19-C26

Goldberger, J.J., Parker, M.A., Kadish, A.H., Hallstrom, A. Over-AVID subgroup analysis. J Am Coll Cardiol, 2001; 38: 1586-1587

Goldhirsch, A., Castiglione, M., Gelber, R.D. A single perioperative adjuvant chemotherapy course for node-negative breast cancer: five-year results of trial v. J Natl Cancer Inst Monogr, 1992; 11: 89-96

Hahn, S., Garner, P., Williams, P. Investigating heterogeneity across studies - a review of the systematic reviews in the infectious diseases module of the Cochrane Library. In: 1999; 7th Annual Cochrane Colloquium Abstracts, October 1999 in Rome:

Hahn, S., Williamson, P.R., Hutton, J.L. Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to a local research ethics committee. 8th Annual Cochrane Colloquium Abstracts, October 2000 in Cape Town

Hall, A.S., Ball, S.G. ACE-inhibitor therapy after myocardial infarction -- a new treatment strategy. Z Kardiol, 1994; 83 suppl 4: 57-62

Hampton, J.R., van Veldhuisen, D.J., Kleber, F.X., Cowley, A.J., Ardia, A., Block, P., Cortina, A., Cserhalmi, L., Follath, F., Jensen, G., Kayanakis, J., Lie, K.I., Mancia, G., Skene, A.M. Randomised study of effect of ibopamine

on survival in patients with advanced severe heart failure. Second Prospective Randomised Study of Ibopamine on Mortality and Efficacy (PRIME II) Investigators. Lancet, 1997; 349: 971-977

Hasenclever, D., Brosteanu, O., Gerike, T., Loeffler, M. Modelling of chemotherapy: the effective dose approach. Ann Hematol, 2001; 80: B89-B94

Henderson-Smart, D.J., Davis, P.G. Prophylactic doxapram for the prevention of morbidity and mortality in preterm infants undergoing endotracheal extubation. Cochrane Library, 2000; Issue 3

Higgins, J., Thompson, S., Deeks, J., Altman, D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Health Serv Res Policy, 2002; 7: 51-61

Hogan, D.B., MacDonald, F.A., Betts, J., Bricker, S., Ebly, E.M., Delarue, B., Fung, T.S., Harbridge, C., Hunter, M., Maxwell, C.J., Metcalf, B. A randomized controlled trial of a community-based consultation service to prevent falls. CMAJ, 2001; 165: 537-543

Hombrink, J., Fröhlich, D., Glatzel, M., Krauss, A., Thiel, H.J., Meier, J., Hamann, D., Mücke, R., Glaser, F.H., Köst, S. Prophylaxe der strahleninduzierten Diarrhö durch Smektit. Ergebnisse einer doppelblind randomisierten, plazebokontrollierten Multicenterstudie. Strahlenther Onkol, 2000; 176: 173-179

Huenerbein, M., Rau, B., Hohenberger, P., Schlag, P.M. The role of staging laparoscopy for multimodal therapy of gastrointestinal cancer. Surg Endosc, 1998; 12: 921-925

Imrey, P.B., Chilton, N.W. Design and analytic concepts for periodontal clinical trials. J Periodontol, 1992; 63: 1124-1140

Ioannidis, J.P.A., Lau, J. Heterogeneity of the baseline risk within patient populations of clinical trials. Am J Epidemiol, 1998; 148: 1117-1126

ISIS-1 Collaborative Group. Randomised trial of intravenous atenolol among 16027 cases of suspected acute myocardial infarction: ISIS-1. Lancet, 1986; July: 57-66

ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17 187 cases of suspected acute myocardial infarction: ISIS-2. Lancet, 1988; August: 349-360

Jenkinson, C., Layte, R., Coulter, A., Wright, L. Evidence for the sensitivity of the SF-36 health status measure to inequalities in health: results from the Oxford healthy lifestyles survey. J Epidemiol Community Health, 1996; 50: 377-380

Julian, D.G. Treatment for survivors of acute myocardial infarction: what have we learned from large intervention trials? Cardiovasc Drugs Ther, 1995; 9 suppl 3: 495-502

Kautzky-Willer, A., Pacini, G., Tura, A., Bieglmayer, C., Schneider, B., Ludvik, B., Prager, R., Waldhäusl, W. Increased plasma leptin in gestational diabetes. Diabetologia, 2001; 44: 164-172

Keavney, B. Genetic association studies in complex diseases. J Hum Hypertens, 2000; 14: 361-367

Keeley, E.C., Boura, J.A., Grines, C.L. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials. Lancet, 2003; 361: 13-20

Kernan, W.N., Viscoli, C.M., Makuch, R.W., Brass, L.M., Horwitz, R.I. Stratified randomization for clinical trials. J Clin Epidemiol, 1999; 52: 19-26

Khoury, M.J., James, L.M., Flanders, W.D., Erickson, J.D. Interpretation of recurring weak associations obtained from epidemiologic studies of suspected human teratogens. Teratology, 1992; 46: 69-77

Koch, G.G., Gansky, S.A. Statistical considerations for multiplicity in confirmatory protocols. Drug Inf J, 1996; 30: 523-534

Kors, J.A., de Bruyne, M.C., Hoes, A.W., van Herpen, G., Hofman, A., van Bemmel, J.H., Grobbee, D.E. T axis as an indicator of risk of cardiac events in elderly people. Lancet, 1998; 352: 601-605

Kurth, T., Glynn, R.J., Walker, A.M., Chan, K.A., Buring, J.E., Hennekens, C.H., Gaziano, J.M. Inhibition of clinical benefits of aspirin on first myocardial infarction by nonsteroidal antiinflammatory drugs. Circulation, 2003; 108: 1191-1195

Lamarche, B., St-Pierre, A.C., Ruel, I.L., Cantin, B., Dagenais, G.R., Després, J.P. A prospective, population-based study of low density lipoprotein particle size as a risk factor for ischemic heart disease in men. Can J Cardiol, 2001; 17: 859-865

Laule, M., Cascorbi, I., Stangl, V., Bielecke, C., Wernecke, K.D., Mrozikiewicz, P.M., Felix, S.B., Roots, I., Baumann, G., Stangl, K. A1/A2 polymorphism of glycoprotein IIIa and association with excess procedural risk for coronary catheter interventions: a case-controlled study. Lancet, 1999; 353: 708-712

Lee, K.L., McNeer, J.F., Starmer, C.F., Harris, P.J., Rosati, R.A. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation, 1980; 61: 508-515

Linn, S.C., Giaccone, G., van Diest, P.J., Blokhuis, W.M., van der Valk, P., van Kalken, C.K., Kuiper, C.M., Pinedo, H.M., Baak, J.P. Prognostic relevance of P-glycoprotein expression in breast cancer. Ann Oncol, 1995; 6: 679-685

Lloyd, J., Askie, L., Smith, J., Tarnow-Mordi, W. Supplemental oxygen for the treatment of prethreshold retinopathy of prematurity. Cochrane Library, 2003; Issue 2

Lommel, A., Dengler, D., Janssen, U., Fertmann, R., Hentschel, S., Wessel, M. Bleibelastung durch Trinkwasser - Teil I: Einfluss auf den Blutbleispiegel junger Frauen. Bundesgesundheitsblatt, 2002; 45: 605-612

Loviscach, M., Rehman, N., Carter, L., Mudaliar, S., Mohadeen, P., Ciaraldi, T.P., Veerkamp, J.H., Henry, R.R. Distribution of peroxisome proliferator-activated receptors (PRARs) in human skeletal muscle and adipose tissue: relation to insulin action. Diabetologia, 2000; 43: 304-311

Magnesium in Coronaries (MAGIC) Trial Investigators. Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. Lancet, 2002; 360: 1189-1196

Mahé, I., Meune, C., Diemer, M., Caulin, C., Bergmann, J.F. Interaction between aspirin and ACE inhibitors in patients with heart failure. Drug Saf, 2001; 24: 167-182

Marcus, R., Peritz, E., Gabriel, K.R. On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 1976; 63: 655-660

Matsuo, K., Hamajima, N., Suzuki, R., Nakamura, S., Seto, M., Morishima, Y., Tajima, K. No substantial difference in genotype frequencies of interleukin and myeloperoxidase polymorphisms between malignant lymphoma patients and non-cancer controls. Haematologica, 2001; 86: 602-608

Matthews, J.N.S., Altman, D.G. Statistics Notes: Interaction 2: compare effect sizes not P values. Br Med J, 1996**a**; 313: 808

Matthews, J.N.S., Altman, D.G. Statistics Notes: Interaction 3: How to examine heterogeneity. Br Med J, 1996**b**; 313: 862

Maurer, W., Hothorn, L.A., Lehmacher, W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In: Biometrie in der chemisch-pharmazeutischen Industrie. Vol. 6. Vollmar, J. (Hrsg.), Stuttgart: Fischer. 1995; 3-18

McLaughlin, J.K. Formaldehyde and cancer: a critical review. Int Arch Occup Environ Health, 1994; 66: 295-301

Midgette, A.S., Stukel, T.A., Littenberg, B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. Med Decis Making, 1993; 13: 253-257

Moreira, E.D., Stein, Z., Susser, E. Reporting on methods of subgroup analysis in clinical trials: a survey of four scientific journals. Braz J Med Biol Res, 2001; 34: 1441-1446

Moreira, E.D., Susser, E. Guidelines on how to assess the validity of results presented in subgroup analysis of clinical trials. Rev Hosp Clín Fac Med S Paulo, 2002; 57: 83-88

Moyé, L.A., Deswal, A. Trials within trials: confirmatory subgroup analyses in controlled clinical experiments. Control Clin Trials, 2001; 22: 605-619

Moyé, L.A., Pfeffer, M.A., Wun, C.C., Davis, B.R., Geltman, E., Hayes, D., Farnham, D.J., Randall, O.S., Dinh, H., Arnold, J.M. Uniformity of captopril benefit in the SAVE Study: subgroup analysis. Eur Heart J, 1994; 15 suppl B: 2-8

Moyé, L.A., Powell, J.H. Evaluation of ethnic minorities and gender effects in clinical trials: opportunities lost and rediscovered. J Natl Med Assoc, 2001; 93: 29S-34S

Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. Risk factor changes and mortality results. JAMA, 1982; 248: 1465-1477

Nadareishvili, Z.G., Beletsky, V., Black, S.E., Fremes, S.E., Freedman, M., Kurzman, D., Leach, L., Norris, J.W. Is cerebral microembolism in mechanical prosthetic heart valves clinically relevant? J Neuroimaging, 2002; 12: 310-315

Naglie, G., Tansey, C., Kirkland, J.L., Ogilvie-Harris, D.J., Detsky, A.S., Etchells, E., Tomlinson, G., O'Rourke, K., Goldlist, B. Interdisciplinary inpatient care for elderly people with hip fracture: a randomized controlled trial. CMAJ, 2002; 167: 25-32

Nahler, G. Methodische Mängel klinischer Studien. Fortschr Med, 1992; 110: 511-514

Nicholl, J. Observational data and randomised trials. Adding value. 2003

no author. Drotrecogin alfa: new preparation. For some cases of severe sepsis? Prescrire Int, 2003; 12: 55-57

no author. Styrene. 2002; 437-550

no author. West of Scotland Coronary Prevention Study: identification of high-risk groups and comparison with other cardiovascular intervention trials. Lancet, 1996; 348: 1339-1342

Norman, S.A., Berlin, J.A., Talbott, E.O., Klassen, A.C., Celentano, D.D., Brookmeyer, R. Protective effects of cervical screening. J Clin Epidemiol, 1991; 44: 457-458

Ober, C., Karrison, T., Odem, R.R., Barnes, R.B., Branch, D.W., Stephenson, M.D., Baron, B., Walker, M.A., Scott, J.R., Schreiber, J.R. Mononuclear-cell immunisation in prevention of recurrent miscarriages: a randomised trial. Lancet, 1999; 354: 365-369

Osborn, D.A., Cole, M.J., Jeffery, H.E. Opiate treatment for opiate withdrawal in newborn infants. Cochrane Library, 2002; Issue 3

Ottenbacher, K.J. Quantitative evaluation of multiplicity in epidemiology and public health research. Am J Epidemiol, 1998; 147: 615-619

Owens, D.K., Holodniy, M., McDonald, T.W., Scott, J., Sonnad, S. A meta-analytic evaluation of the polymerase chain reaction for the diagnosis of HIV infection in infants. JAMA, 1996; 275: 1342-1348

Oxman, A.D., Guyatt, G.H. A consumer's guide to subgroup analyses. Ann Intern Med, 1992; 116: 78-84

Packer, M., O'Connor, C.M., Ghali, J.K., Pressler, M.L., Carson, P.E., Belkin, R.N., Miller, A.B., Neuberg, G.W., Frid, D., Wertheimer, J.H., Cropp, A.B., DeMets, D.L. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. N Engl J Med, 1996; 335: 1107-1114

Pankow, W., Lies, A., Nabe, B., Becker, H.F., Ploch, T., Lohmann, F.W. Continuous positive airway pressure lowers blood pressure in hypertensive patients with obstructive sleep apnea. Somnologie, 2003; 7: 17

Parker, A.B., Naylor, C.D. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J, 2000; 139: 952-961

Parker, A.B., Yusuf, S., Naylor, C.D. The relevance of subgroup-specific treatment effects: the studies of left ventricular dysfunction (SOLVD) revisited. Am Heart J, 2002; 144: 941-947

Perez-Jimenez, F., Lopez-Miranda, J., Gomez, P., Velasco, M.J., Marin, C., Perez-Martinez, P., Moreno, J.A., Paniagua, J.A. The SstI polymorphism of the apo C-III gene is associated with insulin sensitivity in young men. Diabetologia, 2002; 45: 1196-1200

Peto, R., Collins, R., Gray, R. Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol, 1995; 48: 23-40

Peto, R., Mauri, F., Gasparini, M., Barbonaglia, L., Santoro, E., Franzosi, M.G., Tognoni, G., Rovelli, F. Misleading subgroup analyses in GISSI. Am J Cardiol, 1990; 66: 771-772

Piantadosi, S., Gail, M.H. A comparison of the power of two tests for qualitative interactions. Stat Med, 1993; 12: 1239-1248

Pinelli, J., Symington, A. Non-nutritive sucking for promoting physiologic stability and nutrition in preterm infants. Cochrane Library, 2000; Issue 2

Pinelli, J., Symington, A. Non-nutritive sucking for promoting physiologic stability and nutrition in preterm infants. Cochrane Library, 2001; Issue 3

Pocock, S.J. Allocation of patients to treatment in clinical trials. Biometrics, 1979; 35: 183-197

Pocock, S.J. Meta-analysis. Stat Methods Med Res, 1993; 2: 117-119

Pocock, S.J., Assmann, S.E., Enos, L.E., Kasten, L.E. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med, 2002; 21: 2917-2930

Pocock, S.J., Hughes, M.D. Estimation issues in clinical trials and overviews. Stat Med, 1990; 9: 657-671

Pocock, S.J., Hughes, M.D., Lee, R.J. Statistical problems in the reporting of clinical trials. N Engl J Med, 1987; 317: 426-432

Pollack, A., Kuban, D.A., Zagars, G.K. Impact of androgen deprivation therapy on survival in men treated with radiation for prostate cancer. Urology, 2002; 60 (3 suppl 1): 22-30

PORT Meta-analysis Trialists Group. Postoperative radiotherapy in non-small-cell lung cancer: systematic review and meta-analysis of individual patient data from nine randomised controlled trials. Lancet, 1998; 352: 257-263

Potischman, N., Byers, T., Houghton, L., Root, M., Nemoto, T., Campbell, T.C. Effects of breast cancer treatments on plasma nutrient levels: implications for epidemiological studies. Cancer Epidemiol Biomarkers Prev, 1992; 1: 555-559

Premji, S., Chessell, L. Continuous nasogastric milk feeding versus intermittent bolus milk feeding for premature infants less than 1500 grams. Cochrane Library, 2003; Issue 1

Przuntek, H., Conrad, B., Dichgans, J., Kraus, P.H., Krauseneck, P., Pergande, G., Rinne, U., Schimrigk, K., Schnitker, J., Vogel, H.P. SELEDO: a 5-year long-term trial on the effect of selegiline in early Parkinsonian patients treated with levodopa. Eur J Neurol, 1999; 6: 141-150

Raaijmakers, E., Faes, T.J., Scholten, R.J., Goovaerts, H.G., Heethaar, R.M. A meta-analysis of three decades of validating thoracic impedance cardiography. Crit Care Med, 1999; 27: 1203-1213

Randazzo, M.R., Snoey, E.R., Levitt, M.A., Binder, K. Accuracy of emergency physician assessment of left ventricular ejection fraction and central venous pressure using echocardiography. Acad Emerg Med, 2003; 10: 973-977

Rickenbacher, P.R., Pinto, F.J., Lewis, N.P., Hunt, S.A., Gamberg, P., Alderman, E.L., Schroeder, J.S., Valantine, H.A. Correlation of donor characteristics with transplant coronary artery disease as assessed by intracoronary ultrasound and coronary angiography. Am J Cardiol, 1995; 76: 340-345

Ridker, P.M., Hennekens, C.H., Schmitz, C., Stampfer, M.J., Lindpaintner, K. PIA1/A2 polymorphism of platelet glycoprotein IIIa and risks of myocardial infarction, stroke, and venous thrombosis. Lancet, 1997; 349: 385-388

Rochon, P.A., Clark, J.P., Binns, M.A., Patel, V., Gurwitz, J.H. Reporting of gender-related information in clinical trials of drug therapy for myocardial infarction. CMAJ, 1998; 159: 321-327

Salvesen, K.A., Eik-Nes, S.H. Ultrasound during pregnancy and subsequent childhood non-right handedness: a meta-analysis. Ultrasound Obstet Gynecol, 1999; 13: 241-246

Saunders, M., Dische, S., Barrett, A., Harvey, A., Gibson, D., Parmar, M. Continuous hyperfractionated accelerated radiotherapy (CHART) versus conventional radiotherapy in non-small-cell lung cancer: a randomised multicentre trial. CHART Steering Committee. Lancet, 1997; 350: 161-165

Schmidt, M.I., Duncan, B.B., Sharrett, A.R., Lindberg, G., Savage, P.J., Offenbacher, S., Azambuja, M.I., Tracy, R.P., Heiss, G. Markers of inflammation and prediction of diabetes mellitus in adults (Atherosclerosis Risk in Communities study): a cohort study. Lancet, 1999; 353: 1649-1652

Schneeweiss, S., Maclure, M., Soumerai, S.B., Walker, A.M., Glynn, R.J. Quasi-experimental longitudinal designs to evaluate drug benefit policy changes with low policy compliance. J Clin Epidemiol, 2002; 55: 833-841

Severson, R.K., Buckley, J.D., Woods, W.G., Benjamin, D., Robison, L.L. Cigarette smoking and alcohol consumption by parents of children with acute myeloid leukemia: an analysis within morphological subgroups -- a report from the Childrens Cancer Group. Cancer Epidemiol Biomarkers Prev, 1993; 2: 433-439

Shuster, J., van Eys, J. Interaction between prognostic factors and treatment. Control Clin Trials, 1983; 4: 209-214

Sienel, W., Dango, S., Woelfle, U., Morresi-Hauf, A., Wagener, C., Brümmer, J., Mutschler, W., Passlick, B., Pantel, K. Elevated expression of carcinoembryonic antigen-related cell adhesion molecule 1 promotes progression on non-small cell lung cancer. Clin Cancer Res, 2003; 9: 2260-2266

Simon, R. Bayesian subset analysis: application to studying treatment-by-gender interactions. Stat Med, 2002; 21: 2909-2916

Simon, R. Patient subsets and variation in therapeutic efficacy. Br J Clin Pharmacol, 1982; 14: 473-482

Simon, R. Statistical tools for subset analysis in clinical trials. Recent Results Cancer Res, 1988; 111: 55-66

Skovlund, E. Controlled clinical trials in cancer research. Acta Oncol, 1996; 35 suppl.8: 27-33

Steiner, G. Lipid intervention trials in diabetes. Diabetes Care, 2000; 23 suppl. 2: B49-B53

Stephens, R. The dangers of subgroup analysis. Lancet Oncol, 2001; 2: 9

Sylvester, R.J., Denis, L., de Voogt, H. The importance of prognostic factors in the interpretation of two EORTC metastatic prostate cancer trials. European Organization for Research and Treatment of Cancer (EORTC) Genito-Urinary Tract Cancer Cooperative Group. Eur Urol, 1998; 33: 134-143

Takakuwa, K.M., Ernst, A.A., Weiss, S.J., Nick, T.G. Breast cancer knowledge and preventive behaviors: an urban emergency department-based survey. Acad Emerg Med, 2000; 7: 1393-1398

Tonkin, A.M. Clinical relevance of statins: their role in secondary prevention. Atheroscler Suppl, 2001; 2: 21-25

Tsai, Y.Y., Petersen, G.M., Booker, S.V., Bacon, J.A., Hamilton, S.R., Giardiello, F.M. Evidence against genetic anticipation in familial colorectal cancer. Genet Epidemiol, 1997; 14: 435-446

U.S.Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for industry. E9 statistical principles for clinical trials. ICH (Eds.); 1998; E9

Valencia-Flores, M., Bliwise, D.L., Guilleminault, C., Cilveti, R., Clerk, A. Cognitive function in patients with sleep apnea after acute nocturnal nasal continous positive airway pressure (CPAP) treatment: sleepiness and hypoxemia effects. J Clin Exp Neuropsychol, 1996; 18: 197-210

van Gijn, J. From therapeutic trials to current practice. Rev Neurol (Paris), 1999; 155: 708-712

Varonen, H., Kunnamo, I., Stancliffe, R. Grading strength of evidence from A to D for EBM guidelines statements: strength of evidence in summaries of Cochrane reviews. 9th Annual Cochrane Colloquium Abstracts, October 2001 in Lyon

Waitzinger, J., Lenders, H., Pabst, G., Reh, C., Ulbrich, E. Three explorative studies on the efficacy of the antihistamine mebhydroline (Omeril). Int J Clin Pharmacol Ther, 1995; 33: 373-383

Wallack, M.K., Muthukumaran Sivanandham, Whooley, B., Ditaranto, K., Bartolucci, A.A. Favorable clinical responses in subsets of patients from a randomized, multi-institutional melanoma vaccine trial. Ann Surg Oncol, 1996; 3: 110-117

Wehrmann, T., Hurst, A., Lembcke, B., Jung, M., Caspary, W. Biliary lithotripsy with a new electromagnetic shock wave source. A 2-year clinical experience. Dig Dis Sci, 1993; 38: 2113-2120

Wheatley, K., Ives, N., Hancock, B., Gore, M., Eggermont, A., Suciu, S. Does adjuvant interferon-alpha for high-risk melanoma provide a worthwhile benefit? A meta-analysis of the randomised trials. Cancer Treat Rev, 2003; 29: 241-252

White, I., Elbourne, D.R. Subgroup analysis with binary outcomes: which effect measure should be used? 2003

Whitehead, J., Matsushita, T. Stopping clinical trials because of treatment ineffectiveness: a comparison of a *futility design* with a method of stochastic curtailment. Stat Med, 2003; 22: 677-687

Williams, C., Crossland, L., Finnerty, J., Crane, J., Holgate, S., Pearce, N., Beasley, R. Case-control study of salmeterol and near-fatal attacks of asthma. Thorax, 1998; 53: 7-13

Yusuf, S., Wittes, J., Probstfield, J. Evaluating effects of treatment in subgroups of patients within a clinical trial: the case of non-q-wave myocardial infarction and beta blockers. Am J Cardiol, 1990; 66: 220-222

Yusuf, S., Wittes, J., Probstfield, J., Tyroler, H.A. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA, 1991; 266: 93-98