Aus dem
INSTITUT FÜR RECHTSMEDIZIN
DES UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF
Direktor Prof. Dr. med. Klaus Püschel

Der Fehler erster Art in randomisierten Studien — Eine Bestandsaufnahme des Jahrgangs 2004 der Zeitschrift *The Lancet*

Dissertation

zur Erlangung des Grades eines Doktors der Medizin
der Medizinischen Fakultät der Universität Hamburg vorgelegt von

Benjamin Siemann

aus Henstedt-Ulzburg

Hamburg 2008

Angenommen vom Fachbereich Medizin der Universität Hamburg am:
25.03.2009

Veröffentlicht mit Genehmigung des
Fachbereiches Medizin der Universität Hamburg

Prüfungsausschuss, der / die Vorsitzende: Prof. Dr. H.-P. Beck-Bornholdt

Prüfungsausschuss, 2. Gutachter: PD Dr. H.-H. Dubben

Prüfungsausschuss, 3. Gutachter: Prof. Dr. Dr. A. Trojan

Inhaltsverzeichnis

1. Fragestellung

Die Zeitschrift *The Lancet* ist eine der angesehensten medizinischen Fachzeitschriften und gehört zu den fünf Journalen mit dem höchsten Impact Factor. Im Jahre 1996 übernahm diese Zeitschrift das Consort Statement zur Verbesserung der Berichterstattung über randomisierte, kontrollierte Studien. Ziel der vorliegenden Arbeit ist eine Bestandsaufnahme der Inflation des Fehlers erster Art in dieser Zeitschrift acht Jahre nach dieser Selbstverpflilchtung. Hierzu werden die methodischen Ungenauigkeiten mit Einfluss auf den Fehler erster Art aller randomisierten, kontrollierten Studien des Jahrgangs 2004 quantitativ erhoben.

2. Einleitung

2.1. Forschung in der Medizin

In der Medizin werden fortlaufend neue Methoden zur Prävention und Therapie von Krankheiten mit dem Ziel erforscht, die Patientenversorgung zu verbessern. In der Regel werden Fortschritte in der Medizin nur in kleinen Schritten erzielt. Bahnbrechende Erfolge in der Entwicklung neuer Behandlungen zur Heilung bislang unheilbarer Erkrankungen waren früher selten und sind auch in Zukunft kaum zu erwarten (82). Dennoch können neue Präventions- und Therapiemethoden auch bei vergleichsweise kleinen Effekten eine klinisch relevante Verbesserung der Patientenversorgung bedeuten.

Wenn eine neue Therapiemethode vorgeschlagen wird, muss im Rahmen einer rationalen Medizin überprüft werden, ob diese eine bessere Aussicht auf erfolgreiche Behandlung des Patienten hat, als eine etablierte, alte Therapie. Ob eine neue Therapiemethode einer alten tatsächlich überlegen ist, lässt sich nur feststellen, indem beide Methoden miteinander verglichen werden. Im Laufe der medizinischen Forschungsgeschichte wurden verschiedene Möglichkeiten für einen solchen Vergleich gefunden und in Bezug auf ihren Nutzen überprüft.

Die älteste Vorgehensweise ist, die klinische Erfahrung von Ärzten aus deren Einzelfallberichten heranzuziehen. Dieses Verfahren wird auch als historischer Vergleich bezeichnet. Zeigt sich die Überlegenheit einer Therapie bei mehreren Fällen eines Arztes, so könnte daraus geschlossen werden, dass die angewandte Therapie tatsächlich besser sei, als eine andernorts oder vorher angewandte Therapie. Diese Vergleichsmethode hat den Vorteil, dass sie sehr anschaulich und für den Kliniker offensichtlich ist. Letztlich hat diese Vergleichsmethode jedoch nicht zu überzeugenden Ergebnissen geführt. So wirkt es sich nachteilig auf die Validität[1] dieser Art von Vergleichen aus, dass Einflussfaktoren aus der Umgebung, z.B. sich verbessernde Hygiene- und Ernährungsstandards in der Bevölkerung oder verbesserte Schutzmaßnahmen

[1] Validität (lat. validus = kräftig, wirksam) beschreibt das argumentative Gewicht bzw. die Gültigkeit einer, zumeist wissenschaftlichen, Feststellung oder Untersuchung.

im Umgang mit gesundheitsschädlichen Arbeitsstoffen, nicht bei der Erklärung von Wirkungsunterschieden der zu vergleichenden Therapien berücksichtigt werden. Außerdem ist der historische Vergleich wenig systematisierbar, da es sich um Einzelfallberichte und nicht um systematisch geplante Untersuchungen handelt. Somit ist er anfällig für zufällige Fehler. Beispielsweise beruhen die Wahl der Therapie und die Bewertung des Krankheitsverlaufs auf der subjektiven Einschätzung des vergleichenden Klinikers. Es handelt sich somit nicht um eine objektive Vergleichsmethode. So wäre es denkbar, dass häufiger Patienten mit einer besseren Ausgangsprognose der vom Kliniker favorisierten, neuen Therapie zugeführt werden, während Patienten mit schlechteren Heilungschancen nach alter Methode therapiert werden. Aus einem solchen systematischen Fehler resultierte eine Verfälschung der Ergebnisse des klinischen Vergleichs, so dass die neue Therapie fälschlicherweise als bessere Alternative gegenüber der alten Therapie bewertet würde. Die Konsequenz aus einer derartigen falschen Annahme ist die Behandlung von Patienten nach Grundsätzen, die auf Irrtümern beruhen und somit im schlimmsten Fall eine Gefährdung von Menschenleben darstellen.

Ohne ein entsprechendes systematisches Vorgehen beim klinischen Vergleich besteht die Gefahr, dass die Erzielung besserer Ergebnisse durch eine neue Therapie bei kleinen Fallzahlen dem Zufall geschuldet ist. Weiterhin könnten Selektionsprozesse beim Einschluss von Patienten in den Vergleich und deren Zuteilung zu den verschiedenen Therapien stattfinden, die zu einer Verfälschung der Ergebnisse führen. Ferner könnte der Kliniker zufällig ein bestimmtes Patientenklientel behandeln, bei dem eine neue Therapie besser abschnitte, z.B. könnten seine Patienten insgesamt gesünder sein oder bestimmte Risikofaktoren könnten fehlen, während unter anderen Umständen die von ihm angewandte neue Therapie nicht besser abschneiden würde als eine etablierte alte Therapie. Wegen der häufig kleinen Fallzahlen und dieser möglichen und ungewollten Patientenselektion lassen sich die Ergebnisse bei dieser Vergleichsmethode nur sehr begrenzt verallgemeinern und auf ein größeres Patientenkollektiv übertragen.

Von den gewählten Behandlungsmethoden hängen der Krankheitsverlauf eines Patienten und im Extremfall sein Leben ab. Deshalb erfordert der Vergleich der Therapiemethoden in der Medizin eine besondere Beachtung. Entscheidend ist, dass der Vergleich eine möglichst genaue und effektive Schlussfolgerung darüber erlaubt, welche der untersuchten Therapiemethoden für das gegebene Problem am besten geeignet ist. Der Einfluss des Zufalls auf das Ergebnis des Therapievergleichs muss möglichst gering gehalten werden. Aus diesem Grund wird in der heutigen Medizin die Durchführung von möglichst groß angelegten, methodisch abgesicherten und zu evaluierenden klinischen Studien dem historischen Vergleich vorgezogen, da in ersteren die Auswirkungen des Zufalls aufgrund der Patientenzahl und der systematischeren Methodik geringer gehalten werden.

2.2. Randomisierte, kontrollierte Studien

Als eine Forschungsmethode, die diesem Ziel sehr nahe kommt, hat sich die randomisierte, kontrollierte Studie entwickelt. Sie gilt heute als bestes Studiendesign in der medizinischen Forschung und wird deshalb auch als „Goldstandard“ bezeichnet. Die erste randomisierte, kontrollierte Studie war die „Streptomycin-Studie“, die vom British Medical Research Council im Jahre 1947 zur Behandlung der pulmonalen Tuberkulose durchgeführt wurden (28; 87). Der variable Verlauf der Erkrankung und die begrenzte Verfügbarkeit des neuen Medikaments Streptomycin erhöhten die Anforderungen an einen Wirksamkeitsnachweis. Da die Krankheit sehr unterschiedlich verlief wurde es als notwendig angesehen, eine Kontrollgruppe mit einzubeziehen, die die Standardbehandlung, nämlich Bettruhe, erhielt. Außerdem wurden die Patienten randomisiert, das heißt zufällig zur Streptomycin- oder zur Kontrollgruppe zugeteilt. Sowohl das Mitführen einer Kontrollgruppe, als auch die Randomisierung der Patienten war für die damalige Zeit revolutionär. Das Prinzip der Randomisierung wurde von R.A. Fisher in den 1920er Jahren in landwirtschaftlichen Versuchen eingeführt (36); es ist auf Sir Austin Bradford Hill zurückzuführen, dass dieses Prinzip in klinische Versuche Einzug hielt (39; 48).

In einer wie oben beschriebenen randomisierten, kontrollierten Studie werden zwei oder mehr Patientengruppen, auch Studiengruppen genannt, miteinander verglichen, die jeweils eine unterschiedliche Therapie erhalten. „Randomisierung“ bedeutet dabei, dass jeder Patient[1], der in die Studie eingeschlossen wird, eine vorgegebene, bekannte Wahrscheinlichkeit hat, jede der Behandlungen zu erhalten, die konkrete Behandlungszuteilung für den einzelnen Patienten aber nicht vorhergesagt werden kann (7). Hierdurch wird der bei dem historischen Vergleich mögliche, oben erwähnte Selektionsprozess, das heißt eine systematische falsche Zuordnung von Patienten mit besonders guter oder besonders schlechter Prognose zu den einzelnen Behandlungsgruppen, maximal reduziert.

Der Begriff „kontrolliert“ bedeutet in diesem Zusammenhang, dass neben der Gruppe von Patienten, auf welche die neue, zu untersuchende Therapiemethode angewendet wird, noch mindestens eine zweite Gruppe existiert. Diese zweite Gruppe wird entweder mit einem etablierten Therapieregime oder mit einem Placebo[2] behandelt. Ziel der Etablierung einer Kontrollgruppe ist es, den therapeutischen Effekt der traditionellen Therapie bzw. den Placebo-Effekt unter gleichen Versuchsbedingungen aufzuzeichnen, um ihn mit dem Effekt der zu untersuchenden, neuen Therapiemethode vergleichen zu können.

2.3. Evidenzbasierte Medizin (EbM)

Die Ergebnisse von randomisierten, kontrollierten Studien bilden eine wesentliche Grundlage der evidenzbasierten Medizin (Evidence based

[1] Die Teilnehmer an einer klinischen Studie sind nicht notwendigerweise auch Patienten, es können auch gesunde Probanden teilnehmen. Der Einfachheit halber wird in dieser Arbeit stets der Terminus „Patient“ verwendet, auch wenn es sich bei den Teilnehmern einer Studie möglicherweise um gesunde Probanden handelt.

[2] Ein Placebo (lat. „ich werde gefallen“) ist ein medizinisches Präparat, das keinen pharmazeutischen Wirkstoff enthält und somit per Definition keine pharmazeutische Wirkung verursachen kann. Es wird in der Forschung eingesetzt, um die unspezifischen, potentiell den Krankheitsverlauf beeinflussenden Effekte wie die vermehrte Aufmerksamkeit des Personals und der Ärzte oder die Teilnahme als Patient an einer Studie selbst auszuschalten. Die Wirkung eines Placebos, die sich aus diesen unspezifischen Wirkfaktoren ergibt, wird als Placebo-Effekt bezeichnet.

Medicine; EbM). Eine klassische Definition von evidenzbasierter Medizin geht auf Sackett et. al. zurück: „Evidenzbasierte Medizin ist der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen, wissenschaftlichen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten. Die Praxis der evidenzbasierten Medizin bedeutet die Integration individueller klinischer Expertise mit der bestverfügbaren externen Evidenz aus systematischer Forschung“ (78).

Danach soll in der evidenzbasierten Medizin neben der klinischen Erfahrung, die bisher hauptsächlich den Therapieablauf bestimmte, die Wahl der geeigneten Therapie wesentlich von den Ergebnissen wissenschaftlicher Forschung mitbestimmt werden. Entgegen dahingehenden Annahmen wurde dies in der modernen Medizin keineswegs schon immer praktiziert. Das therapeutische Vorgehen und die Integration vorhandener Evidenz unterscheidet sich teilweise erheblich bei verschiedenen Klinikern in Bezug auf Patienten mit vergleichbaren Krankheitsbildern (78; 97).

Erfordert die Definition von Sacket et al. die beste externe, wissenschaftliche Evidenz, so setzt dies logisch die Möglichkeit der hierarchischen Einordnung der Evidenzen voraus. Eine gebräuchliche Hierarchie wurde von der US Agency for Health Care Policy and Research veröffentlicht (2). Darin gibt es vier Evidenzstufen, wobei die Stufen I und II weiter in a und b untergliedert sind, so dass insgesamt sechs verschiedene Evidenzstufen unterschieden werden können. Die Evidenzstufe Ia ist als die beste externe Evidenz anzusehen und besteht aus Meta-Analysen[1] randomisierter, kontrollierter Studien in systematischen Übersichtsarbeiten. Die einzelne randomisierte, kontrollierte Studie folgt direkt danach auf der Stufe Ib. Auf den weiteren Stufen finden sich kontrollierte Studien ohne Randomisierung, quasi-experimentelle Studien, nicht experimentelle, deskriptive Studien und als niedrigste Evidenz die Meinung von Experten und Konsensuskonferenzen.

[1] Eine Meta-Analyse ist eine nach bestimmten Regeln durchgeführte zusammenfassende Bewertung der Ergebnisse möglichst vieler randomisierter kontrollierter Studien zur gleichen bzw. ähnlichen Fragestellung

Das Ergebnis einer randomisierten, kontrollierten Studie stellt also für sich genommen schon die zweitbeste Evidenz dar und bildet zugleich den wesentlichen Bestandteil für die Meta-Analyse mehrerer randomisierter, kontrollierter Studien als den bestmöglichen Evidenzgrad. Das bedeutet, dass in der täglichen Versorgung von unzähligen Patienten die Therapie auf der Basis von Ergebnissen aus klinischen, randomisierten Studien ausgewählt wird. Diese zentrale Bedeutung der randomisierten, kontrollierten Studie für die Weiterentwicklung und Verbesserung der modernen Medizin erfordert die Beachtung und Überprüfung der Qualität dieser Studien, um eine bestmögliche medizinische Qualität zu gewährleisten.

Eine randomisierte kontrollierte Studie wird zu einer inhaltlichen Fragestellung durchgeführt, die in Form eines Hypothesenpaares vor der Datenerhebung formuliert und anhand eines statistischen Tests überprüft wird. Das Hypothesenpaar besteht aus einer zu überprüfenden Nullhypothese (H_0) und einer Alternativhypothese (H_1), die sich gegenseitig ausschließen. Die Nullhypothese gilt es mittels des statistischen Tests gegenüber der Alternativhypothese zu bestätigen oder zu verwerfen. Um die Nullhypothese zu überprüfen muss eine Variable untersucht werden, auf deren Veränderung im Rahmen der Studie geachtet wird. Die Variable, anhand derer entschieden wird, welche Hypothese man den Vorzug gibt, wird auch als primärer Endpunkt bezeichnet. Der primäre Endpunkt ist das Kriterium, anhand dessen über Wirksamkeit und Unwirksamkeit der zu untersuchenden Intervention entschieden wird.

> Beispiel: Es soll in einer randomisierten kontrollierten Studie überprüft werden, ob der Genuss von Alkohol die Reaktionsfähigkeit von Probanden einschränkt. Hierzu wird folgende Nullhypothese formuliert: H_0= „der Genuss von Alkohol schränkt die Reaktionsfähigkeit von Probanden nicht ein“ mit der dazugehörigen Alternativhypothese H_1= „der Genuss von Alkohol schränkt die Reaktionsfähigkeit von Probanden ein“. Es werden zwei Studiengruppen mit beispielsweise je 100 Probanden gebildet, von denen eine eine definierte Menge (z.B. 1 l) alkoholhaltiges Bier (4,5% Alkoholgehalt) und die andere alkoholfreies

Bier (0% Alkoholgehalt) erhält; idealerweise weisen beide Getränke dabei keine geschmacklichen Unterschiede auf. Jeweils vor und nach dem Konsum wird die Reaktionsfähigkeit zu einem festgelegten Zeitpunkt gemessen, etwa indem die Zeit zwischen einem akustischen Signal und der Betätigung eines Knopfes gemessen und aufgezeichnet wird; dies dient der Überprüfung der Hypothese, so dass die Zeitspanne den primären Endpunkt der Studie bildet. Das Ergebnis könnte im gegebenen Beispiel sein, dass die Zeitspanne zwischen Signal und Knopfdruck in der Gruppe, die 1 l alkoholhaltiges Bier zu sich genommen hat, signifikant[1] länger ist, als in der Gruppe, die alkoholfreies Bier erhielt.

Was ließe sich aus einem solchen Ergebnis schließen? Eine Studie wird mit dem Ziel unternommen, eine Antwort auf die klinische Fragestellung zu erhalten (im obigen Beispiel: „Hat Alkohol Einfluss auf die Reaktionsfähigkeit?“). Bevor sie die Frage beantwortet, wirft eine Studie jedoch eine entscheidende Frage auf: Wie sicher ist das Ergebnis der Studie? Bildet das Ergebnis die Wirklichkeit ab oder basiert es lediglich auf Zufall?

Die Zuverlässigkeit eines gewonnenen Ergebnisses hängt davon ab, wie wahrscheinlich es ist, dass das Ergebnis zufällig zustande gekommen sein kann. Bei der randomisierten Zuteilung von Probanden in die Studiengruppen kann das Ergebnis wie oben dargelegt stets zu einem bestimmten Teil dem Zufall geschuldet sein und dadurch falsche Schlussfolgerungen bedingen. Die Wahrscheinlichkeit für ein solches zufälliges Ergebnis wird Signifikanz genannt und in der Statistik mit einem so genannten P-Wert bezeichnet, der zwischen 0 und 1 liegt. Ein P-Wert von P=0 bedeutet dabei, dass es sich nicht um ein zufälliges Ergebnis handeln kann (0% Wahrscheinlichkeit), während P=1 die Aussage enthält, dass es sich mit Sicherheit (100%) um ein zufälliges Ergebnis handelt. Der P-Wert wurde von R.A. Fisher (36) als Hilfsmittel für den Forscher bei der Beurteilung von Studienergebnissen entwickelt. Die Beurteilung der Glaubwürdigkeit von Studienergebnissen anhand des P-Wertes hat sich in der medizinischen Forschung durchgesetzt. Sie erfolgt ferner gemeinhin im Sinne

[1] Als signifikant (lat.: *significans* bezeichnend, anschaulich) wird in der Statistik ein Ergebnis bezeichnet, bei dem es unwahrscheinlich ist, dass es zufällig zustande kam.

einer dichotomen Aussage, so dass die Ergebnisse einer Studie entweder als glaubwürdig oder als unglaubwürdig gelten, während die Abstufungen dazwischen selten benannt werden. Als Grenze für die Glaubwürdigkeit wird das so genannte Signifikanzniveau (α) gewählt. Dieses Signifikanzniveau darf der P-Wert des statistischen Tests nicht überschreiten, damit das Ergebnis einer Studie als „auf dem gegebenen Signifikanzniveau statistisch signifikant“ gelten darf. Regelmäßig werden in der medizinischen Forschung dabei Werte von α<0.05 oder seltener α<0.01 als Grenze gewählt (82). Danach sind im ersten Fall 5% und im zweiten Fall 1% der Studienergebnisse als zufällig akzeptiert.

2.4. Der Fehler erster Art (α-Fehler)

Da diese zufälligen Ergebnisse unabhängig von der Genauigkeit in der Durchführung einer Studie auftreten, kommen sie auch bei methodisch perfekt durchgeführten Studien vor. Tritt ein solches zufälliges Ergebnis auf, so lehnt der Forscher die Nullhypothese aufgrund der zufälligen Daten ab und gibt fälschlicherweise der Alternativhypothese den Vorzug. Bei einer einzelnen Studie ist es nicht möglich, einen solchen Umstand festzustellen.

Das Auftreten eines zufälligen Ergebnisses wird auch als falsch positives Ergebnis oder als Fehler erster Art (α-Fehler) bezeichnet. Auch in dem oben erwähnten Beispiel („Hat Alkohol Einfluss auf die Reaktionsfähigkeit?“) besteht die Möglichkeit, dass zufällig vermehrt Probanden an der Studie teilnehmen, welche trotz des Konsums der definierten Menge Alkohol keine deutliche Einschränkung ihrer Reaktionsfähigkeit erleiden, etwa dann, wenn sich Gewöhnungseffekte einstellten. Das Ergebnis der Studie ließe dann allenfalls den Schluss zu, dass Alkohol gemäß der Nullhypothese keinen Einfluss auf die Reaktionsfähigkeit hat. Das Ergebnis wäre mithin falsch positiv, das heißt man würde einen Fehler erster Art begehen.

Besteht ein solcher Fehler erster Art, wovon bei einer von 20 Studien mit einem 5%-Signifikanzniveau ausgegangen werden muss, so hat dieses falsch positive

Ergebnis nach Publikation in entsprechenden Fachzeitschriften der Medizin Einfluss auf die Behandlung von Patienten. Es ist zu befürchten, das in einem solchen Falle die Patientenversorgung schlechter wird, da sie auf falschen Ergebnissen beruht. Die Wahrscheinlichkeit für das Auftreten von falsch positiven Ergebnissen mit einer Häufigkeit gemäß Signifikanzniveau einer Studie lässt sich nicht verringern. Andererseits kann die Wahrscheinlichkeit für ein falsch positives Ergebnis einer Studie unter bestimmten Umständen über das angegebene Signifikanzniveau hinaus ansteigen. Man spricht dann von einer Inflation des Fehlers erster Art, bei der das effektive Signifikanzniveau größer ist als das für die Studie angegebene Niveau. Als effektives Signifikanzniveau bezeichnet man das tatsächliche Signifikanzniveau der gesamten Studie unter Berücksichtigung von Faktoren, die das Signifikanzniveau erhöhen, wie z.B. multiple Testung, was im Folgenden erklärt wird.

2.5. Grundprinzipien statistischer Tests

Grundsätzlich beginnt jeder statistische Test mit *einer* inhaltlichen Fragestellung, die sich in *einem* Hypothesenpaar formulieren lässt und mit *einem* Experiment anhand *eines* primären Endpunkts überprüft wird. Die Reihenfolge dieser Schritte ist einzuhalten, insbesondere müssen Fragestellung, Hypothesenpaar, Aufbau des Experiments und Definition des primären Endpunkts *vor* der Datenerhebung festgelegt werden (82). Für diesen Fall gilt, dass das angegebene Signifikanzniveau auch tatsächlich die Wahrscheinlichkeit angibt mit der das Ergebnis zufällig sein könnte. Wird eines der Elemente erst nach Datenerhebung festgelegt, so besteht die Gefahr der unbewussten Beeinflussung der Formulierung durch bereits erhobene Daten und es ist nicht mehr gewährleistet, dass die Wahrscheinlichkeit eines zufälligen Ergebnisses nicht über dem angegebenen Signifikanzniveau liegt.

2.6. Multiple Testung

Forscher und deren Sponsoren haben das Ziel, aus den aufwendigen Untersuchungen einer klinischen Studie so viele Informationen wie möglich zu gewinnen. Der Versuch einer möglichst umfangreichen Auswertung der erhobenen Daten liegt dabei nahe. Dies kann geschehen, indem mehr als ein statistischer Test durchgeführt wird, beispielsweise dadurch, dass mehr als ein primärer Endpunkt untersucht wird, bereits vor Abschluss der Datenerhebung Auswertungen, so genannte Interimanalysen durchgeführt oder mehr als zwei Studiengruppen miteinander verglichen werden. Bei der statistischen Absicherung der so gewonnen Erkenntnisse ergäbe sich jedoch schnell die Schwierigkeit, dass das Signifikanzniveau einer Studie für genau *einen* statistischen Test bei der Überprüfung genau *eines* primären Endpunkts (siehe oben) gilt. Sobald ein weiterer Endpunkt untersucht oder ein primärer Endpunkt mehrfach überprüft und somit weitere statistische Tests durchgeführt werden, besteht indessen erneut die Möglichkeit eines zufälligen Zustandekommens der Testergebnisse mit einer Wahrscheinlichkeit gemäß dem angegebenen Signifikanzniveau.

Die Durchführung mehr als eines statistischen Tests in Bezug auf das primäre Ergebnis einer Studie wird dies als multiple Testung bezeichnet, wobei mehrere Arten unterschieden werden. Ist eine multiple Testung vorhanden, ist das effektive Signifikanzniveau der Studie in annähernd gleich der Summe der Signifikanzniveaus der einzelnen Tests, vorausgesetzt, dass keine sonstigen Einflüsse auf das Signifikanzniveau bestehen.

2.7. Mehr als ein primärer Endpunkt

Eine offensichtlich zu erkennende Form der multiplen Testung liegt vor, wenn in einer Studie mehr als ein primärer Endpunkt untersucht wird. Dies kann zwar durchaus legitim sein, doch sind dann Vorkehrungen zu treffen, um eine Inflation des Fehlers erster Art unter solchen Bedingungen zu vermeiden. Werden mehrere primäre Endpunkte untersucht, so ist für jeden Endpunkt die Wahrscheinlichkeit, dass ein zufälliges Ergebnis auftritt so groß wie das

angegebene Signifikanzniveau für diesen Test. Werden also zwei primäre Endpunkte mit einem Signifikanzniveau von P=0,05 untersucht, so ist in beiden Fällen die Wahrscheinlichkeit für ein zufälliges Ergebnis unabhängig voneinander 5%. Das bedeutet, dass das effektive Signifikanzniveau, also die Wahrscheinlichkeit, dass mindestens einer von beiden statistischen Tests ein zufälliges Ergebnis liefert, näherungsweise 10% entspricht[1]. Die Wahrscheinlichkeit, dass diese Studie ein falsch positives Ergebnis produziert liegt demzufolge bei ca. 10% statt der üblicherweise geforderten 5%. Für die Patientenversorgung im Rahmen der evidenzbasierten Medizin verdoppelt sich mithin die Wahrscheinlichkeit, eine Therapiemethode ohne tatsächliche Überlegenheit fälschlicherweise als besser zu erachten.

Es gibt jedoch Möglichkeiten, eine Studie mit zwei primären Endpunkten durchzuführen und bei einem effektiven Signifikanzniveau von 5% zu bleiben, beispielsweise könnten hierfür die Signifikanzniveaus der beiden einzelnen Tests bei 2,5% gehalten werden. Dadurch kann das übliche Signifikanzniveau von 5% in etwa erreicht werden. Werden mehr als zwei primäre Endpunkte getestet, so müssen die Einzelwahrscheinlichkeiten gemäß der Anzahl der Tests reduziert werden[2]. Wird ein solches Vorgehen angewendet, spricht man auch von einer α-Adjustierung oder davon, dass für die multiple Testung „korrigiert“ wurde. Die oben beschriebene Methode ist das Vorgehen nach Bonferroni, welches das bekannteste, aber auch gröbste Verfahren ist: es wurde der Anschaulichkeit halber als Beispiel gewählt. Darüber hinaus gibt es andere, differenziertere Verfahren (82). Werden mehr als ein primärer Endpunkt getestet ohne korrigierend einzugreifen, so verliert die Studie weit mehr an Aussagekraft als es auf den ersten Blick erscheint. Die Wahrscheinlichkeit, einen Fehler erster Art zu begehen und damit negativ auf die Behandlungsstrategien für die untersuchten Krankheitsbilder einzuwirken, ist somit größer als in der Publikation angegeben.

[1] Allgemein ausgedrückt beträgt das effektive Signifikanzniveau bei *n* unabhängigen Tests auf einem 5%-Signifikanzniveau $p = 1 – (0,95)^n$.

[2] Es gilt allgemein $\alpha = 0,05/n$ mit α = Signifikanzniveau für den einzelnen Test und n = Anzahl der primären Endpunkte.

2.8. Multiple Testung, wenn nur ein primärer Endpunkt vorhanden ist

Auch wenn nur ein primärer Endpunkt in einer Studie untersucht wird, besteht die Möglichkeit der multiplen Testung. Sie führt zu einer Inflation des Fehlers erster Art, wenn korrigierend eingegriffen wird. Sobald ein zweiter statistischer Test in Bezug auf den primären Endpunkt durchgeführt wird, handelt es sich um multiple Testung mit der oben beschriebenen Problematik des erhöhten α-Fehlers. Typische Vertreter dieser Form der multiplen Testung sind die Durchführung von Interim- und Subgruppenanalysen, sowie der Vergleich von mehr als zwei Studiengruppen untereinander.

2.9. Interimanalysen

Bei der Interimanalyse werden vorab, d.h. vor Ablauf der festgelegten Studiendauer, Auswertungen der gesammelten Daten durchgeführt. Dies geschieht regelmäßig, um im Falle einer deutlichen Überlegenheit bzw. Schädlichkeit einer der beiden Therapieoptionen die Studie vorzeitig abbrechen zu können. Eine Interimanalyse stellt eine multiple Testung bezüglich des primären Endpunktes dar (12). Ein Einfluss auf das effektive Signifikanzniveau lässt sich lediglich dann vermeiden, wenn die Interimanalyse von einem von der Studienleitung unabhängigen Komitee durchgeführt wird. Die Auswertung darf nur dann Einfluss auf den Verlauf der Studie haben, wenn vorher definierte Abbruchkriterien erfüllt sind; ansonsten darf die Studienleitung die Ergebnisse der Interimanalyse im Verlauf der Studie nicht erfahren. Erlangt die Studienleitung Kenntnis von den Ergebnissen der Interimanalyse währen der Durchführung, so liegt eine multiple Testung vor und es muss hierfür durch eine Anpassung des Signifikanzniveaus korrigiert werden, damit es nicht zu einer Inflation des Fehlers erster Art kommt. Hierfür dienen so genannte gruppensequentielle Verfahren (72, 73).

Zur Veranschaulichung des Einflusses von Interimanalysen diene folgendes Beispiel:

> Ziel einer Studie sei der Vergleich eines neuen blutdrucksenkenden Medikaments mit einem etablierten Medikament. Der primäre Endpunkt

sei die Reduktion des systolischen Blutdruckes der Probanden nach vierzehntägiger Behandlung und verglichen mit dem systolischen Blutdruck vor der Einnahme der Medikamente. Würde nun im Verlauf der Studie eine Interimanalyse durchgeführt, indem schon nach sieben Tagen die Blutdruckwerte der Behandlungsgruppen verglichen würde, besteht die Möglichkeit, dass die Werte in einer Behandlungsgruppe im Durchschnitt niedriger sind als in der anderen Gruppe, beispielsweise, weil eines der Medikamente seine Wirkung schneller entfaltet. Würde nach 14 Tagen gemessen, könnten die durchschnittlichen Blutdruckwerte der anderen Behandlungsgruppe niedriger sein, weil das Medikament dieser Gruppe insgesamt den besseren Effekt hat, auch wenn es erst später anfängt zu wirken. Außerdem könnten die Unterschiede zufälliger Natur sein. Bestehen keine Unterschiede zwischen den Medikamentenwirkungen und basieren die gemessenen Unterschiede auf Zufall, so beginge man einen Fehler erster Art, wenn das Medikament der Gruppe mit dem niedrigeren Blutdruck dem anderen vorgezogen würde. Wegen der zweimaligen Durchführung dieser Prüfung bei einer Interimanalyse, besteht zweimal die Wahrscheinlichkeit, ein zufälliges Ergebnis und in der Folge einen Fehler erster Art zu erhalten.

2.10. Subgruppenanalysen

Während der Analyse einer Studie werden häufig Fragen aufgeworfen, die nicht im ursprünglichen Studiendesign berücksichtigt wurden, wie etwa der Vergleich der Therapien in bestimmten durch prognostische Faktoren definierten Untergruppen von Patienten (Subgruppen). Besonders dann, wenn das erhoffte Ergebnis – in der Regel die Überlegenheit einer neuen Therapiemethode gegenüber einer Standardmethode – ausbleibt, kann es hilfreich erscheinen Subgruppenanalysen mit dem Ziel durchzuführen, die Überlegenheit der neuen Therapiemethode zumindest in Bezug auf eine definierte Untergruppe von Patienten, z.B. besonders alte oder junge Patienten oder solche in einem bestimmten Krankheitsstadium, zu zeigen. Bei dieser Vorgehensweise wird die

Grenze für die Wahrscheinlichkeit fälschlicherweise auf einen Wirkungsunterschied zwischen den Therapiegruppen zu schließen nicht mehr eingehalten, da mehrere statistische Tests durchgeführt werden. Geht man davon aus, dass die einzelnen Subgruppen sich nicht überlappen und auf einem 5%-Signifikanzniveau getestet wird, so ist die Wahrscheinlichkeit für mindestens eine falsch positive Entscheidung bei zwei Subgruppen 10%, bei fünf Subgruppen 23% und bei 100 Subgruppen 99%[1] (82).

Die Analyse von Subgruppen kann durchaus berechtigt sein, um aus aufwendigen, teuren Studien die größtmögliche Erkenntnis zu ziehen. Dabei muss jedoch beachtet werden, dass Ergebnisse aus nachträglich durchgeführten Subgruppenanalysen als solche im Sinne eines explorativen Ergebnisses kenntlich gemacht und durch spätere Studien mit einer auf das gefundene Ergebnis zugeschnittenen Fragestellung überprüft werden müssen. Dieses Vorgehen wird jedoch oft nicht befolgt (74).

2.11. Vergleich von mehr als zwei Studiengruppen untereinander

Es kann sich als günstig erweisen, im Rahmen einer Studie mehr als zwei Studiengruppen zu führen. So etwa, wenn es zwei etablierte Medikamente für eine Erkrankung gibt und ein neues Medikament in seiner Wirksamkeit mit diesen beiden verglichen werden soll. Bei einem solchen Studiendesign ist zu beachten, dass es sich um multiple Testung mit der Notwendigkeit der Korrektur handelt, wenn die Studiengruppen untereinander verglichen werden, da z.B. das neue Medikament erst gegen das eine alte getestet wird und dann gegen das andere. Da es sich um zwei statistische Tests handelt, besteht zweimal die Möglichkeit, dass ein zufälliges Ergebnis gemäß dem gewählten Signifikanzniveau entsteht.

[1] Die Formel zur exakten Berechnung ist dieselbe wie zur Berechnung des effektiven Signifikanzniveaus bei mehr als einem primären Endpunkt (siehe oben).

2.12. Darstellung von mehr als einer Studie im Rahmen einer Publikation

Üblicherweise werden in einer Publikation einer randomisierten, kontrollierten Studie die Ergebnisse genau *einer* Studie veröffentlicht. Es kommt jedoch auch vor, dass die Ergebnisse von zwei unabhängigen randomisierten, kontrollierten Studien in einer Publikation veröffentlicht werden. Dies ist insofern problematisch, als dass dies zu werten ist wie das Vorhandensein zweier primärer Endpunkte einer Studie, da auch in diesem Fall die Chance auf ein falsch positives Ergebnis der Publikation bei ca. 10% liegt, wenn von einem 5%-Signifikanzniveau der einzelnen Tests ausgegangen wird und keine Korrektur für die multiple Testung stattgefunden hat.

2.13. Intention-to-treat Analyse

Eine weitere Quelle für eine Inflation des Fehlers erster Art liegt im Umgang mit der Randomisierung und dem anschließenden Einbezug der Patienten in die Auswertung der Ergebnisse. Die Randomisierung schafft neben der Verblindung[1] die Basis für einen unverfälschten Behandlungsvergleich, da durch die Randomisierung keine Beeinflussung der Zuteilung von Patienten zu den zu vergleichenden Studiengruppen möglich ist.

Es kann passieren, dass der Eindruck entsteht, als müssten Patienten nach erfolgter Randomisierung die Studiengruppe noch wechseln. Wenn tatsächlich Patienten im Verlauf der Studie einer anderen Gruppe zugeteilt werden als der, der sie ursprünglich zufällig zugewiesen wurden, so ist das Prinzip des *unverfälschten* Behandlungsvergleichs bereits gefährdet. Hierzu folgendes Beispiel:

> Zwei Medikamente sollen bezüglich ihrer Wirksamkeit bei der Behandlung von Bluthochdruck verglichen werden. Etliche Patienten, die

[1] „Verblindung“ ist ein Prinzip bei der Durchführung randomisierter, kontrollierter Studien, bei dem die Patienten nicht wissen, ob sie den echten Wirkstoff oder ein Placebo bekommen. Bei einer *Doppel*blindstudie wissen weder die Patienten noch die Behandelnden, welche Patienten den Wirkstoff und welche ein Placebo bekommen. Ziel dieser Methode ist es, unbewusste und unspezifische Einflüsse von Seiten der Patienten und Behandelnden auf die Ergebnisse der Studie auszuschließen.

den neuen Wirkstoff erhielten, brechen die Therapie ab. Bei genauerem Hinsehen findet sich, dass dies auf die für zu stark befundenen Nebenwirkungen des neuen Medikaments zurückzuführen ist. Würden die Patienten nicht in ihrer ursprünglichen Studiengruppe, also in der Gruppe mit dem neuen Medikament, ausgewertet, so käme es zu einer Verfälschung, da nur solche Patienten in der Gruppe blieben, die das Medikamente besonders gut vertragen und sich somit in einem wesentlichen Punkt von den anderen Patienten unterscheiden. Die *unverfälschte* Vergleichbarkeit wäre nicht mehr gegeben. Die Wahrscheinlichkeit, eines falsch positiven Ergebnisses ist größer als das angegebene Signifikanzniveau, weil die Patienten aus der Gruppe mit dem neuen Medikament aus der Auswertung herausgenommen worden wären, bei denen die Wirksamkeit nicht gut gewesen wäre z.B. aufgrund von mangelnder Compliance[1] bei zu starken Nebenwirkungen.

Diesem Problem kann mit der so genannten Intention-to-treat[2] Analyse begegnet werden. Darunter versteht man eine Methode der Auswertung randomisierter, kontrollierter Studien. Hiernach müssen (a) alle in die Studie eingeschlossenen und randomisierten Patienten in die Analyse eingehen, und zwar (b) in der Gruppe, zu der sie randomisiert wurden, unabhängig davon, was nach der Randomisierung mit ihnen geschieht. Allerdings wird dieses Verfahren zum Teil unterschiedlich strikt angewendet (49). Streng genommen dürfte kein einziger Patient nach erfolgter Randomisierung die Gruppe tauschen oder aus der Auswertung herausgenommen werden.

In der Praxis ergeben sich häufig Probleme, weil bei einzelnen Patienten z.B. später eindeutig festgestellt wird, dass sie von vornherein die Einschlusskriterien nicht erfüllt hatten und folglich nicht in die Gruppe hätten randomisiert werden dürfen. Für Einzelfälle wie diesen gibt es keine einheitliche Einigung bezüglich der Anwendung der Intention-to-treat Analyse (49).

[1] Compliance (englisch Befolgung) beschreibt die Mitarbeit des Patienten bei dessen Genesung sowie die Befolgung ärztlicher Anordnungen durch den Patienten.
[2] In etwa aus dem Englischen: Analyse nach Behandlungsabsicht (*Übers. des Autors).*

Grundsätzlich ist festzuhalten, dass bei Ausschluss von Patienten aus der Auswertung in der Gruppe, in die sie randomisiert wurden, ein falsch positives Ergebnis durch mögliche Selektionsprozesse wahrscheinlicher wird und es somit zu einer Inflation des Fehlers erster Art kommen kann.

2.14. Welches Signifikanzniveau wurde gewählt?

Das Signifikanzniveau, welches für eine randomisierte kontrollierte Studie gewählt wird, beeinflusst entscheidend, wie wahrscheinlich es ist, dass ein falsch positives Ergebnis zustande kommt. Je höher das Signifikanzniveau ist, desto wahrscheinlicher ist ein falsch positives Ergebnis. Je niedriger das Signifikanzniveau hingegen ist, desto wahrscheinlicher beruhen die Ergebnisse auf einem realen Effekt und nicht auf dem Zufall. Deshalb ist die Kenntnis des gewählten Signifikanzniveaus für die kritische Beurteilung und die Einschätzung der Verlässlichkeit der Ergebnisse einer Publikation durch den Leser notwendig.

Das Signifikanzniveau sollte *vor* Durchführung der Studie gewählt und explizit in der Publikation erwähnt werden. Würde das Signifikanzniveau nach Abschluss der Studie gewählt, so könnte eine Anpassung des Signifikanzniveaus an die tatsächlichen Ergebnisse nicht ausgeschlossen werden. Wird das gewählte Signifikanzniveau nicht explizit erwähnt, so kann ebenfalls nicht mit Sicherheit beurteilt werden, auf welchem Signifikanzniveau die Studienleiter die Testung ursprünglich durchführen wollten. So könnten beispielsweise Ergebnisse auf einem 5% Signifikanzniveau als signifikant dargestellt werden, obwohl ursprünglich eine Testung auf einem 1% Signifikanzniveau intendiert war. Hingewiesen sei hier auf die Möglichkeit unbewusster Einflussnahme auf die Darstellung von Ergebnissen, wenn die Modalitäten der Durchführung und Auswertung von Studienergebnissen nicht explizit vor Erhebung und Auswertung der Daten festgelegt wird.

2.15. Bedeutung der Qualität von Publikationen randomisierter, kontrollierter Studien

Die Ergebnisse von randomisierten, kontrollierten Studien nehmen wie oben dargestellt im Rahmen der evidenzbasierten Medizin deutlichen Einfluss auf die Behandlung von Patienten. Die Veröffentlichung der Ergebnisse erfolgt in Form von Publikationen in medizinischen Fachzeitschriften. Ein Rezipient, der die Ergebnisse einer randomisierten, kontrollierten Studie für seine Arbeit nutzen möchte, erhält Informationen über die Studie, d.h. über das Studiendesign, die Durchführung und die Ergebnisse der Studie aus der Publikation. Anhand dieser urteilt er auch über die Qualität und Glaubwürdigkeit einer Studie. Deshalb ist es wichtig, dass in der Publikation einer randomisierten, kontrollierten Studie die Informationen für eine solche Evaluation möglichst eindeutig dargestellt werden. Trotz der Bemühungen zur Verbesserung der Berichterstattung über randomisierte kontrollierte Studien während einiger Jahrzehnte sind viele Publikationen noch nicht adäquat in Bezug auf die Vollständigkeit der dargebotenen Informationen (67).

2.16. Das Consort-Statement

Mitte der 90er Jahre mündeten zwei voneinander unabhängige Initiativen zur Qualitätsverbesserung der Beschreibung von randomisierten kontrollierten Studien in der Publikation des CONSORT Statement (Consolidated Standards of Reporting Trials) (15). Das CONSORT Statement wird unterstützt von einer zunehmend größeren Zahl medizinischer Zeitschriften (6; 6; 6; 38; 38; 53; 81), von Herausgebervereinigungen wie dem “International Committee of Medical Journal Editors (ICMJE, auch bekannt als die Vancouver Group) (25) und von der World Association of Medical Editors (WAME).

Das Consort-Statement besteht aus einer Checkliste und einem Flussdiagramm zur Beschreibung von randomisierten kontrollierten Studien. Beides zusammen wird der Einfachheit halber Consort genannt. Consort ist in erster Linie dazu bestimmt, beim Verfassen, Beurteilen und Auswerten von Berichten über

randomisierte kontrollierte Studien mit zwei parallelen Gruppen behilflich zu sein.

Consort wurde durch kontinuierliche Verbesserungsvorschläge, z.B. von Meinert (65) modifiziert, bis bei einem Treffen der 13 Consort Mitglieder im Jahr 1999 beschlossen wurde, eine revidierte Fassung zu erstellen (67). Sowohl die Checkliste als auch das Flussdiagramm wurden überarbeitet und die einzelnen Items geändert und spezifiziert. Außerdem wurde eine mit Erläuterungen der einzelnen Items versehene Langfassung geschrieben (8).

Der Gebrauch von Consort scheint die Anzahl unzureichender Berichte über randomisierte kontrollierte Studien zu verringern (29; 66). Moher et al. verglichen die Qualität von Berichten über randomisierte kontrollierte Studien vor (1994) und nach (1998) Einführung des Consort Statement in vier führenden medizinischen Zeitschriften (British Medical Journal, Journal of the American Medical Association, The Lancet, and The New England Journal of Medicine), von denen alle mit Ausnahme des The New England Journal of Medicine das Consort Statement übernahmen. Es zeigte sich eine Verbesserung der Berichterstattung für alle vier Zeitschriften über die Zeit, die jedoch lediglich für die Zeitschriften, die Consort übernahmen, signifikant auf einem 1%-Signifikanzniveau waren. Es fanden sich in dieser Arbeit jedoch auch weiterhin Mängel in der Berichterstattung in Bezug auf die Checklisten-Items und die Autoren heben hervor, dass die Berichterstattung über randomisierte kontrollierte Studien noch immer der Verbesserung bedarf.

2.17. Ziel der vorliegenden Arbeit

Aus den oben dargelegten Gründen spielt eine Inflation des Fehlers 1. Art eine große Rolle für die evidenzbasierte Medizin und wirkt sich potentiell negativ auf die medizinische Versorgungsqualität von Patienten aus. Eine Inflation des Fehlers 1. Art kann der Rezipient einer Studie lediglich aus der Publikation heraus erkennen und nur dann, wenn diese die Informationen offen legt, die dafür notwendig sind.

Ziel der vorliegenden Arbeit ist eine Bestandsaufnahme der Inflation des Fehlers 1. Art in randomisierten, kontrollierten Studien in einer der führenden biomedizinischen Zeitschriften acht Jahre nachdem die Zeitschrift das Consort Statement zur Verbesserung der Berichterstattung über randomisierte, kontrollierte Studien übernahm. Die methodischen Ungenauigkeiten mit Einfluss auf den Fehler erster Art werden anhand der Publikationen einer der angesehensten internationalen medizinischen Fachzeitschriften quantifiziert. Geprüft wird der Jahrgang 2004 der Zeitschrift *The Lancet*, die zu den fünf Zeitschriften mit dem höchsten Impact Factor[1] gehört.

[1] Der Impact Factor einer Fachzeitschrift misst, wie oft in Arbeiten anderer Fachzeitschriften aus ihr zitiert wird in Relation zur Gesamtzahl der dort veröffentlichten Artikel. Er wird häufig herangezogen um Bedeutung und Ansehen einer biomedizinischen Fachzeitschrift zu beurteilen.

3. Material und Methoden

3.1. Methoden der Studienauswahl

In dieser Arbeit werden die Publikationen der randomisierten, klinischen Studien eines kompletten Jahrgangs (2004) einer englischsprachigen, biomedizinischen Fachzeitschrift (The Lancet) hinsichtlich des Ausmaßes des Fehlers erster Art untersucht.

Aus dem Jahrgang 2004 wurde jede Publikation in die Auswertung eingeschlossen, die in Titel oder Zusammenfassung als „randomisiert" beschrieben wird und menschliche Probanden untersucht. Berücksichtigt wurden Originalpublikation klinischer Studien, die in der Rubrik „original research"[1] veröffentlicht wurden.

Die Identifizierung der zu untersuchenden Publikationen wurde mit der erweiterten Suchfunktion auf der Internetseite der Zeitschrift The Lancet (http:// www.thelancet.com/search/advanced) vorgenommen.

Eine Suchoperation mit folgenden Suchparametern lieferte eine Vorauswahl von 113 Publikationen. Die Suchparameter waren: „Search for text: 'randomised' in 'Summary' or 'Title' restricted by 'original research' (später 'primary research') with 'Date range' January 2004 to December 2004".

Auf die gefundenen 113 Publikationen wurde zur Überprüfung eine Volltextsuche nach dem Suchbegriff „randomised" mit der Software Adobe® Reader® in der bei Benutzung aktuellsten Version 6.0.218.05.2004 angewendet.

25 Publikationen enthalten den Suchbegriff nicht. Acht Publikationen enthalten den Suchbegriff im Volltext, jedoch nicht in 'Title' oder 'Summary'. Hiervon liegen bei drei Publikationen die Treffer im Literaturverzeichnis. Drei

[1] Die Rubrik „original research" wurde im Rahmen einer Umgestaltung des Designs und Aufbaus des Lancet im Juli 2004 beginnend mit dem Issue 9428, Volume 364 in „primary research" umbenannt. Das Einschlusskriterium „Publikation in 'original research'" wurde beginnend mit diesem Zeitpunkt in „Publikation in 'primary research'" geändert.

Publikationen enthalten den Suchbegriff in einem Verweis auf randomisierte Studien. Zwei Publikationen, die den Suchbegriff „randomised“ im Methodenteil enthalten, scheiden aus, da es sich um ein systematisches Review und um eine Studie an Ratten handelt.

Bei den verbliebenen 80 Publikationen, die tatsächlich den Suchbegriff „randomised“ in 'Title' oder 'Summary' enthalten, wurde der Kontext hinzugezogen, in dem der Suchbegriff steht.

Acht Publikationen wurden ausgeschlossen, weil aus dem Kontext ersichtlich ist, dass es sich um Übersichtsarbeiten, so genannte „Reviews“, handelt oder mit dem Begriff „randomised“ lediglich Bezug auf andere Publikation genommen wurde. Die verbleibenden 72 Publikationen wurden in die Arbeit eingeschlossen.

Um den Auswahlprozess und dessen Interraterreliabilität zu überprüfen wurden stichprobenartig von Herrn Prof. Dr. Beck-Bornholdt insgesamt drei Ausgaben des Lancet auf Publikationen überprüft, die nach den oben genannten Regeln einzuschließen seien. Es fanden sich hierbei keinerlei Abweichungen von den tatsächlich eingeschlossenen Arbeiten.

3.2. Aufbau des Fragebogens

Zur Datengewinnung wurde ein Extraktionsbogen verwendet, der vom Autor auf der Grundlage eines Extraktionsbogens von Herrn Prof. Dr. Beck-Bornholdt erstellt wurde. Der Extraktionsbogen dient im Sinne einer Checkliste der einheitlichen Erfassung der in der Einleitung als relevant herausgestellten Punkte, die Einfluss auf den Fehler erster Art nehmen können.

Der Extraktionsbogen umfasst insgesamt neun Items: sieben dichotome Ja/Nein-Fragen und zwei numerische Items. Im Folgenden werden die einzelnen Items und die Regeln nach denen sie ausgefüllt wurden vorgestellt:

3.3. Handelt es sich um den Endbericht einer Originalstudie?

Dieses Item beschreibt, ob es sich um den endgültigen Bericht einer einzigen Originalstudie handelt. Wenn es Hinweise dafür gab, dass die Publikation sich auf mehrere Studien bezog oder eine Studie in mehrere Publikationen aufgeteilt wurde, so wurde dieses Item verneint, ansonsten bejaht.

3.4. Werden primäre Endpunkte definiert?

Mit diesem Item wird überprüft, ob in der Publikation primäre Endpunkte definiert werden.

Als Angabe für den primären Endpunkt wurden Formulierungen akzeptiert, die folgende Schlüsselwörter enthielten:

- primary endpoint (primärer Endpunkt)
- primary outcome / primary outcome measure (primäres Ergebnis / primäre Ergebnismessung)
- main outcome (Hauptergebnis)

3.5. Wie viele primäre Endpunkte werden definiert?

Dieses Item gibt an, wie viele primäre Endpunkte in der Publikation angegeben wurden.

3.6. Wird ggf. für diese Form der multiplen Testung korrigiert?

Dieses Item beschreibt, ob die Autoren der Publikation angeben, für das Vorhandensein mehrerer primärer Endpunkte korrigiert zu haben. Diese Angabe kann entweder in der Erläuterung der Korrekturmethode bestehen oder darin zu erwähnen, „dass für das Vorhandensein von mehr als einem primären Endpunkt korrigiert wurde“. Dieses Item beschreibt weder, ob die Korrekturmethode für den entsprechenden Fall sinnvoll ist, noch ob sie erfolgreich angewendet wurde.

3.7. Wird irgendeine Form der multiplen Testung in Bezug auf einen primären Endpunkt durchgeführt?

Mit diesem Item wird festgehalten, ob sich in der Publikation Hinweise für Interim-, Subgruppenanalysen oder Vergleiche von mehr als zwei Studiengruppen untereinander finden. Sollte dies der Fall sein, so wurde das Item bejaht, ansonsten verneint.

3.8. Wird ggf. für diese Form multiple Testung korrigiert?

Dieses Item wurde nach den gleichen Regeln ausgefüllt, wie bei der multiplen Testung durch das Vorhandensein von mehr als einem primären Endpunkt (siehe oben), allerdings bezogen auf die multiple Testung in Bezug auf einen primären Endpunkt (siehe voriges Item).

3.9. Handelt es sich nach Angaben der Autoren um eine Intention-to-treat Analyse?

Dieses Item beschreibt, ob die Autoren der Publikation angeben, dass es sich um eine Intention-to-treat Analyse handelt. Mit diesem Item wird nicht überprüft, ob tatsächlich eine Analyse durchgeführt wurde, die den Namen Intention-to-treat Analyse verdient. Dies wird im folgenden Item überprüft.

3.10. Gehen alle Probanden in die Endauswertung ein?

Dieses Item beschreibt, ob alle Probanden, die randomisiert wurden, in die Auswertung des primären Endpunktes eingehen. In dieser Arbeit wird nicht über die Begründung für den Ausschluss der Probanden aus der Auswertung geurteilt, sondern lediglich in strengstem Sinne jede Publikation identifiziert, in der ersichtlich ist, dass nicht alle randomisierten Probanden in die Auswertung eingehen. Diese Information kann entweder durch eine explizite Aussage der Autoren gewonnen oder anhand der Flussdiagramme festgestellt werden, die widerspiegeln, wie die Patienten durch die Studie geführt wurden.

3.11. Welches Signifikanzniveau wurde gewählt?

Mit diesem Item wird festgehalten, welches Signifikanzniveau die Autoren der Publikation explizit angeben.

4. Ergebnisse

4.1. Endberichte

Von den 72 untersuchten Publikationen waren 59 Endberichte, d.h. Publikationen über die kompletten Endergebnisse des primären Endpunkts einer randomisierten Studie (siehe Abbildung 1). Die übrigen 13 Publikationen waren Publikationen über mehr als eine randomisierte Studie (90), über einen Teil der Ergebnisse (50, 35, 30), über vorläufige bzw. vorzeitige Ergebnisse (45, 23, 86, 43, 77), über „follow-up“-Ergebnisse (83, 71) oder es wurden Teile der Studie bereits anderswo veröffentlicht (80, 32).

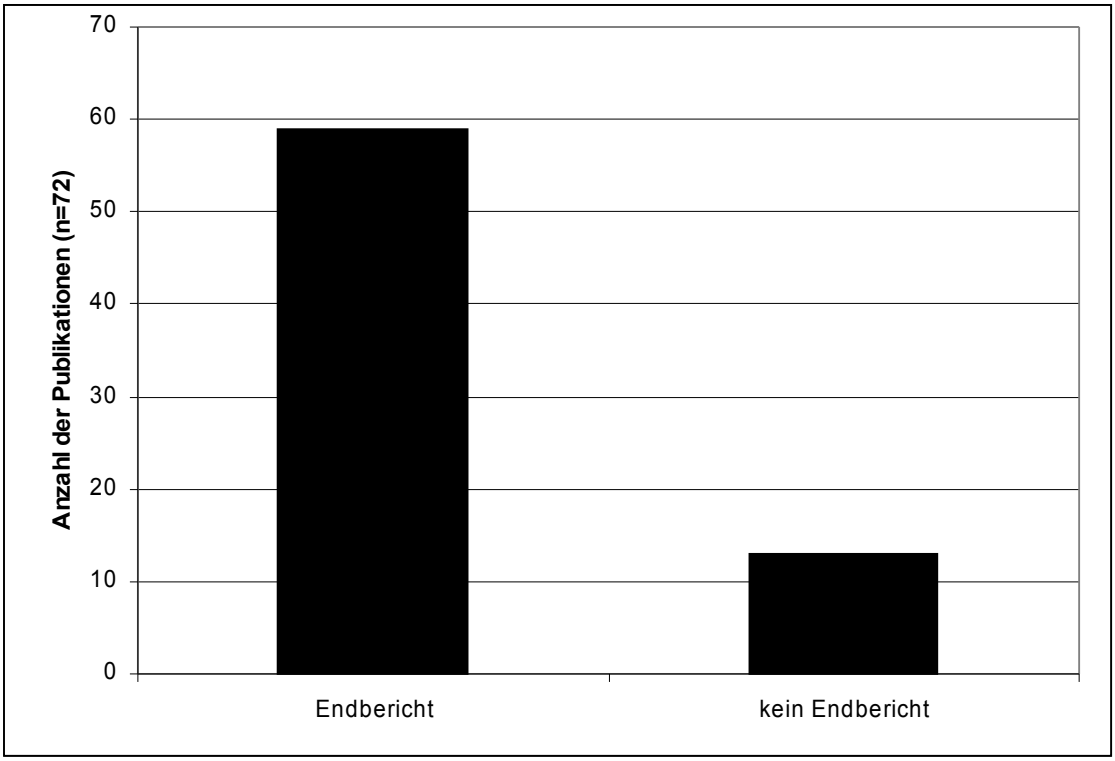

Abbildung 1 – Endberichte

4.2. Anzahl primärer Endpunkte und Korrektur bei Vorhandensein von mehr als einem primären Endpunkt

In 66 der 72 Publikationen wird mindestens ein primärer Endpunkt definiert. In 16 dieser 66 Publikationen wurde mehr als ein primärer Endpunkt getestet. In drei dieser 16 Publikationen wurde angegeben, dass für das Vorhandensein von mehr als einem primären Endpunkt korrigiert wurde (44, 77, 37).

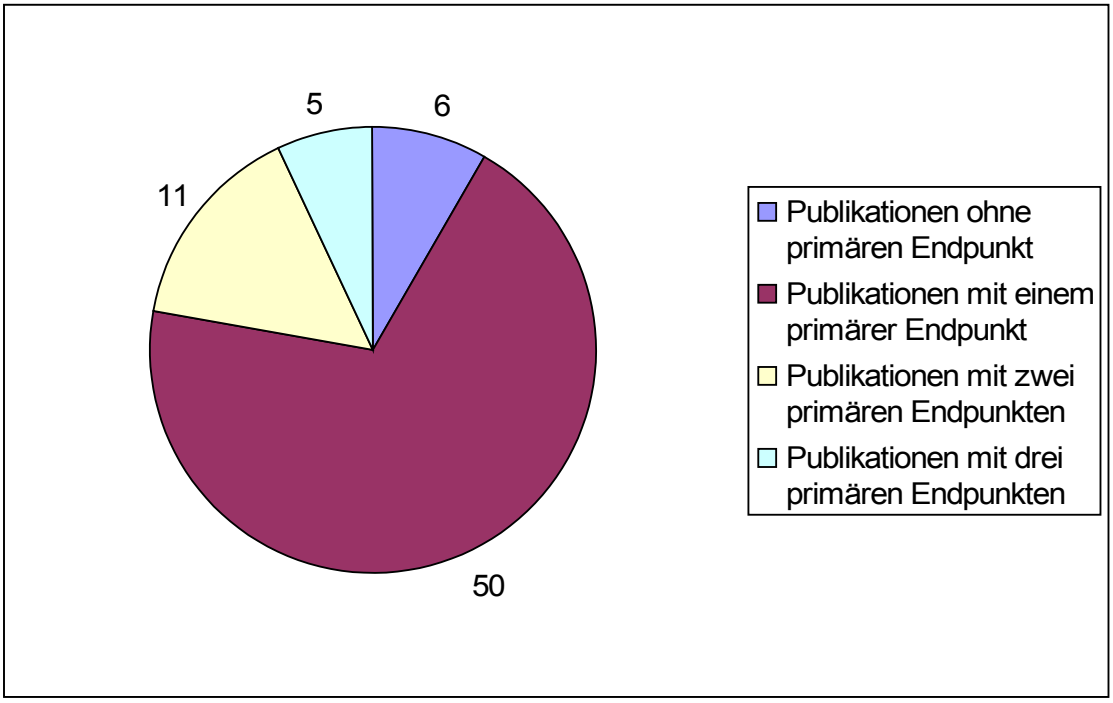

Abbildung 2 - Anzahl der primären Endpunkte

4.3. Multiple Testung bezüglich eines primären Endpunkts

In 35 der 66 Publikationen, die mindestens einen primären Endpunkt angeben, wird deutlich, dass eine Mehrfachtestung für mindestens einen primären Endpunkt durchgeführt wird. Hierbei handelte es sich um Mehrfachtestungen durch:

- Vergleiche zwischen mehr als zwei Studiengruppen
- Subgruppenanalysen
- Interimanalysen

Bei elf dieser 35 Publikationen, die eine Mehrfachtestung für mindestens einen primären Endpunkt angeben, findet sich ein Hinweis darauf, dass für diese Mehrfachtestung eine Korrektur vorgenommen wurde (siehe Abbildung 3).

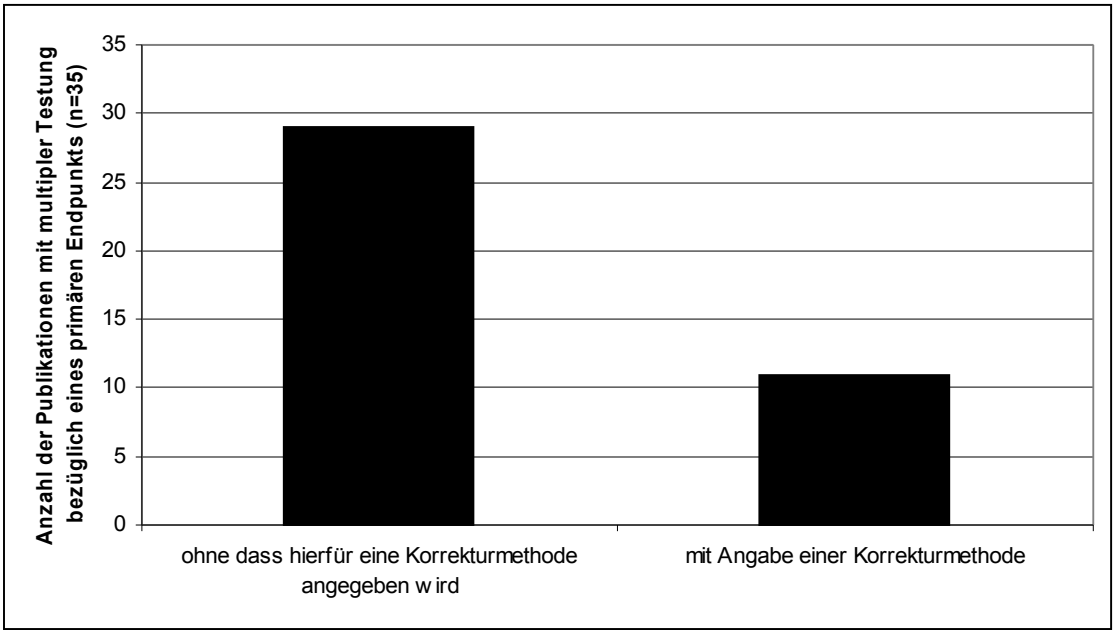

Abbildung 3 – Multiple Testung bezüglich eines primären Endpunktes

4.4. Intention-to-treat-Analyse

In 58 von 72 Publikationen geben die Autoren an, eine Intention-to-treat Analyse durchzuführen. Die Autoren von 31 der 58 Publikationen, in denen eine Intention-to-treat Analyse als Methode angegeben wird, beschreiben, dass nach erfolgter Randomisierung ein Teil der Probanden nicht in die Auswertung eingeht (siehe Abbildung 4).

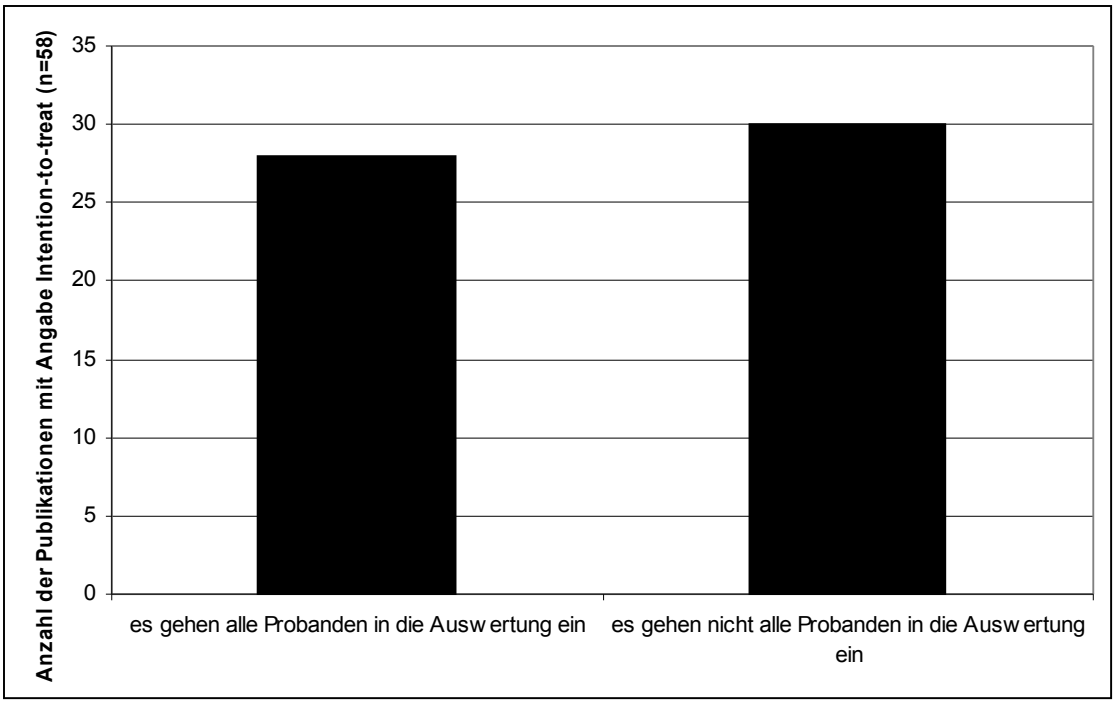

Abbildung 4 - Auswertung der Probanden bei angegebener Intention-to-treat Analyse

4.5. Angabe des Signifikanzniveaus

In 54 der 72 Publikationen wurde angegeben, auf welchem Signifikanzniveau getestet wurde. In 48 dieser 54 Publikationen wurde ein Signifikanzniveau von $p<0,05$ angegeben. In den anderen Publikationen wurde dreimal $p<0,01$ und je einmal $p<0,025$, $p<0,03$ und $p<0,049$ als Signifikanzniveau angegeben.

4.6. Zusammenfassung der einzelnen Publikationen bezüglich des Fehlers erster Art

Im Folgenden werden die in die vorliegende Arbeit eingeschlossenen Publikationen des Jahrgangs 2004 des Lancet in der Reihenfolge ihres Erscheinens einzeln zusammengefasst hinsichtlich der Mängel in Bezug auf den Fehler erster Art und dessen mögliche Inflation:

4.7. Beschreibung der untersuchten Publikationen

Tran et al. (90) untersuchten im Rahmen zweier in einer Veröffentlichung zusammengefassten Studien die Sicherheit und Effektivität von Dihydroartemisinin-piperaquin Kombinationspräparaten im Vergleich zur Standardbehandlung von unkomplizierter Plasmodium falciparum Malaria in Vietnam mit Artesunat-mefloquin. Es zeigten sich keine Unterschiede zwischen den Behandlungen. Die Autoren schließen aus ihren Ergebnissen, dass Dihydroartemisinin-piperaquin Kombinationspräparate als sichere, hoch effektive und gleichzeitig kostengünstige Präparate einen bedeutsamen Beitrag zur Bekämpfung der Malaria liefern können.

Die Autoren präsentieren die Ergebnisse von zwei Studien (einer Pilotstudie und einer Hauptstudie) in einer Publikation. In der Hauptstudie findet eine Mehrfachtestung statt, indem drei Gruppen miteinander verglichen werden, ohne dass hierfür korrigiert wird. Positiv zu vermerken ist, das dies eine der wenigen Publikationen ist, die dem Anspruch einer Intention-to-treat Analyse tatsächlich gerecht wird, da alle randomisierten Patienten in die Auswertung eingehen.

Gilliland et al. (41) überprüften die Hypothese, dass Null-Genotypen für GSTM1 und GSTT1, sowie Varianten des Codon 105 von GSTP1 (alle drei kodieren für Proteine der Glutathion-S-Transferase Superfamilie) Schlüsselrollen bei der Verstärkung von allergischen Reaktionen durch Diesel Abgaspartikel spielen. Es zeigte sich ein stärkerer Anstieg der IgE-Konzentration und Histaminausschüttung nach Exposition mit Diesel Abgaspartikeln und einem zusätzlichen Allergen bei Individuen mit Null-GSTM1 und homozygoten GSTP I105 Genotypen. Die Autoren schließen aus den Ergebnissen, dass GSTM1 und GSTP1 eine genetisch anfällige Population für die Verstärkung von allergischen Reaktionen durch Diesel Abgaspartikel identifizieren.

In dieser Publikation wurde kein primärer Endpunkt definiert. Es wurden drei Studiengruppen miteinander verglichen, ohne dass für diese Form der multiplen Testung eine Korrekturmethode angegeben wird.

Von Koningsveld et al. (93) untersuchten, ob die zusätzliche Gabe von Methylprednisolon zur Standardtherapie mit intravenös applizierten Immunglobulinen (IVIg) beim Guillain-Barré-Syndrom (GBS) zu besseren Ergebnissen führt als die alleinige Gabe von Immunglobulinen. Primärer Endpunkt der Studie war die Verbesserung um einen oder mehrere Punkte gegenüber dem Ausgangswert auf der GBS Invaliditätsskala nach vier Wochen. Es zeigten sich keine signifikanten Unterschiede.
Trotz der fehlenden statistischen Signifikanz schließen die Autoren, dass die beiden Medikamente synergistisch wirken könnten[1]. Kritisch anzumerken ist ferner, dass die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, obwohl die Ergebnisse von acht der 233 Patienten nach erfolgter Randomisierung nicht in die Auswertung eingehen.

Harrison et al. (47) untersuchten, ob die Verdopplung der Dosis inhalierbarer Kortikosteroide bei außer Kontrolle geratenem Asthma die Anzahl der Patienten senkt, die orale Kortikosteroide (Prednisolon) benötigen. Es wurde in Bezug auf die Häufigkeit der Inanspruchnahme von oralen Kortikosteroiden kein Unterschied gefunden zwischen den Studienteilnehmern, die eine doppelte Dosis inhalierbarer Kortikosteroide einnahmen im Vergleich zu denen, die eine einfache Dosis erhielten. Die Autoren schließen aus ihren Ergebnissen, dass es Patienten mit außer Kontrolle geratenem Asthma nicht zu raten sei, die Dosis der inhalierbaren Kortikosteroide zu verdoppeln.
Die Autoren geben an, eine Intention-to-treat Analyse zu machen, es gehen jedoch die Ergebnisse von 37 der 390 randomisierten Patienten nicht in die Auswertung ein. Weiterhin findet sich keine Explizite Angabe über das verwendete Sifnifikanzniveau.

Carr et al. (18) untersuchten den Effekt von Rosiglitazon auf die an Armen und Beinen gemessene Lipoatrophie bei HIV-positiven Erwachsenen, die eine antiretrovirale Therapie erhalten. Es zeigte sich kein Effekt von Rosiglitazon auf die Lipoatrophie HIV-Infizierter unter antiretroviraler Therapie. Die Autoren

[1] “Although our findings did not indicate a significant difference in treatment effect between patients given methylprednisolone and IVIg and those given IVIg alone, we believe the two drugs might work synergistically.” (93)

schließen aus ihren Ergebnissen, dass Rosiglitazon nicht zur Therapie der Lipoatrophie unter antiretroviraler Therapie bei HIV-Infizierten empfohlen werden könne.
Es wurde eine Interimanalyse ohne entsprechende Korrektur des Signifikanzniveaus durchgeführt. Dies dürfte jedoch kaum zu einer Inflation des Fehlers erster Art führen, da das Signifikanzniveau für die Iterimsanalyse mit $p<0,001$ sehr klein war. Die Ergebnisse von fünf der 108 randomisierten Patienten gehen nicht in die Auswertung ein, obwohl die Autoren angeben, dass eine Intention-to-treat Analyse durchgeführt wurde.

Muir et al. (69) untersuchten, ob Magnesiumsulfat, das innerhalb der ersten zwölf Stunden nach Apoplex intravenös appliziert wird, eine Verringerung der Mortalität oder Behinderung nach 90 Tagen zur Folge hat. Es zeigte sich keine Verringerung der Mortalität oder Behinderung nach der oben beschriebenen Intervention. Die Autoren geben zu bedenken, dass die Patientenzahl zu klein gewesen sei, um einen geringen, aber klinisch relevanten Unterschied auszuschließen.
Nach erfolgter Randomisierung gehen die Ergebnisse von 206 der 2589 Patienten nicht in die Auswertung ein. Die ausgeschlossenen Patienten werden im Rahmen einer Sicherheitsanalyse berücksichtigt, nicht jedoch in der Effizienzanalyse.

Holmberg et al. (50) untersuchten die Sicherheit einer Hormonsubstitutionstherapie zur Minderung menopausaler Symptome bei Frauen nach behandeltem Brustkrebs. Die Studie wurde vorzeitig abgebrochen, da das Risiko des erneuten Auftretens von Brustkrebs in der hormonbehandelten Studiengruppe inakzeptabel hoch war. Eine Erklärung für die harmloseren Ergebnisse einer gleichzeitig laufenden, ähnlichen Studie in Stockholm (Schweden) finden die Autoren nicht.
Die Publikation beschreibt die vorläufigen Ergebnisse der Sicherheitsanalyse. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen die Ergebnisse von 89 der 434 randomisierten Patienten nicht in die Auswertung ein. Die Autoren geben an, Interimanalysen durchzuführen, bei

denen es sich um Sicherheitsanalysen handele, welche das Signifikanzniveau nicht beeinflussen würden.

Jocham et al. (56) untersuchten, ob die adjuvante Therapie mit einem Impfstoff aus autologen Nierentumorzellen das Risiko einer Tumorprogression bei Patienten mit Zustand nach radikaler Nephrektomie senkt. Es wurde eine absolute Reduktion des Risikos einer Tumorprogression bei adjuvanter Therapie mit autologen Nierentumorzellen von 12,7% gefunden. Die Autoren schließen aus ihren Ergebnissen, dass der Einsatz eines Impfstoffes aus autologen Nierentumorzellen bei Patienten mit Zustand nach radikaler Nephrektomie als Therapieoption bedacht werden könne.
Die Autoren geben an, eine Intention-to-treat-analyse durchzuführen, es gehen jedoch die Ergebnisse von 179 der 558 randomisierten Patienten nicht in die Auswertung ein.

Christ-Crain et al. (20) untersuchten den Effekt einer Procalcitonin-Kontrolle auf die Verwendungshäufigkeit von Antibiotika bei Infektionen der unteren Atemwege. Bei gleichem klinischen Ergebnis wurden in der Therapie unter Procalcitoninkontrolle bei akuter Infektion der unteren Atemwege 47% weniger und bei akuter Exazerbation einer COPD 56% weniger Antibiotika verabreicht. Die Autoren schließen aus diesen Ergebnissen, dass eine Procalcitonin-Kontrolle bei der Therapie akuter Atemwegsinfektionen oder akuter COPD-Exazerbationen klinische und finanzielle Bedeutung habe.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen die Ergebnisse von 21 der 242 Patienten nicht in die Auswertung ein.

Klareskog et al. (59) verglichen den Effekt der Kombination von Etanercept und Methotrexat mit den Monotherapien beider Substanzen bei der Behandlung von Patienten mit rheumatoider Arthritis. Es wurden bessere Ergebnisse für die Kombinationstherapie nach den Kriterien des American College of Rheumatology erzielt als für die Monoterapien. Die Autoren schließen aus ihren Ergebnissen, dass eine Kombinationstherapie aus Etanercept und Methotrexat einer Monotherapie mit einer der beiden Substanzen vorzuziehen sei.

In dieser Publikation werden zwei primäre Endpunkte untersucht, ohne dass hierfür korrigiert wird. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen die Ergebnisse von 4 der 686 randomisierten Patienten nicht in die Auswertung ein.

Collins et al. (22) untersuchten den Effekt der Cholesterinsenkung mit Simvastatin auf Schlaganfälle oder andere vaskuläre Ereignisse bei Patienten mit cerebrovaskulärer Erkrankung oder anderen Hochrisikofaktoren. Es wurde eine Abnahme der Schlaganfallhäufigkeit unter Simvastatin gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die Therapie mit einem Statin routinemäßig bei Patienten mit hohem Risiko für einen Schlaganfall erwogen werden sollte.
In dieser Publikation fehlt die Übersicht über Patienteneinschlüsse und -ausschlüsse, so dass nicht beurteilt werden kann, ob alle randomisierten Patienten in die Auswertung eingehen.

Gale et al. (40) untersuchten die Wirkung von hoch dosiertem Nicotinamid auf die Entstehung eines Typ I Diabetes. Es wurde kein Unterschied zur Placebo-Gruppe gefunden. Die Autoren schließen aus diesem Ergebnis, dass Nicotinamid in der hier verwendeten Dosierung keinen Einfluss auf die Entstehung von Typ I Diabetes hat. Dennoch behaupten die Autoren, dass ihre Studie unter anderen Studien gezeigt habe, dass Hoffnung bestehe, Diabetes Typ I könne beizeiten zu einer vermeidbaren Erkrankung werden.[1]
In dieser Publikation geben die Autoren an, zwei Interimanalysen durchgeführt zu haben mit definierten Kriterien für den Abbruch der Studie: $p<0,0001$ für die Wirksamkeit der Behandlung und $p<0,05$ für eine Verschlechterung des Diabetes durch die Behandlung. Bei der zweiten Interimanalyse wurde die Studie abgebrochen, da die untersuchte Behandlung nutzlos zu sein schien. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen

[1] Im Ergebnisteil: „Nicotinamide treatment did not have any discernible effect on the primary outcome – i.e., progression to diabetes." und im Diskussionsteil: "Results from ENDIT [das ist die Studie der Autoren] and DPT-1 have shown that this era is coming to an end, and give us hope that type 1 diabetes will, in time, be a preventable disease."

gehen die Ergebnise von 3 der 552 randomisierten Patienten nicht in die Auswertung ein.

Emerson et al. (31) untersuchten die Hypothese, dass augensuchende Fliegen (insbesondere Musca sorbens) Vektoren für eine Conjunctivitis trachomatosa sind und prüften weiterhin, ob die Errichtung von Latrinen die Prävalenz von trachomatösen Conjunctivitiden senkt. Reduktion der Fliegen-Augenkontakte hatte eine Reduktion der Prävalenz von Conjunctivitis trachomatosa zur Folge. Die Autoren schließen hieraus, dass augensuchende Fliegen Vektoren für eine Conjunctivitis trachomatosa sind. In der „Latrinen-Gruppe“ gab es weniger Augenkontakte von Musca sorbens als in der Kontroll-Gruppe. Die Autoren schließen hieraus, dass die Versorgung mit Latrinen die Fliegen-Augenkontakte durch Musca sorbens reduziere und damit eine Reduktion der Trachomprävalenz einhergehe, obwohl diese Reduktion nicht signifikant war.[1]
Es fand eine Überprüfung zweier primärer Endpunkte statt, ohne dass hierfür eine Korrekturmethode angegeben wird. Ferner geben die Autoren nicht an, auf welchem Signifikanzniveau getestet wurde.

Turnbull et al. (91) untersuchten die Hypothese, dass bei drei ausgewählten Komplikationen während der Schwangerschaft (Hypertonus mit und ohne Proteinurie, sowie verfrühter und vorzeitiger Blasensprung = pPROM) klinische, psychosoziale und ökonomische Ergebnisse in einer Tagesbetreuung denen einer stationären Aufnahme gleich oder sogar überlegen sind. In Bezug auf klinische und ökonomische Ergebnisse wurde kein Unterschied gefunden. Die allgemeine Zufriedenheit der Patientinnen in der Tagesbetreuungsgruppe war besser als die der stationär aufgenommenen Patientinnen. Die Autoren geben zu bedenken, dass die stärkste Einschränkung der Studie die niedrige Patientenrekrutierung sei: Lediglich zwei Drittel der angestrebten Patientenzahl wurde rekrutiert. Aus ihren Ergebnissen schließen die Autoren, dass eine Studie lohnend wäre, die einen einzigen primären Endpunkt und genügend definitive empirische Evidenz hat, um die Äquivalenz der Tagesbetreuung

[1] „The reduction in trachoma was not significant, but our results show that latrine provision is effective at reducing fly-eye contact which is an additional, and important, public-health benefit of safe disposal of human faeces.”

gegenüber der stationären Aufnahme in Bezug auf Komplikationen während der Schwangerschaft zu prüfen. Obwohl die Autoren keinen Unterschied in Bezug auf die ökonomischen Ergebnisse fanden, geben sie an, dass die Tagesbetreuung eventuell sogar günstiger sein könnte als die stationäre Behandlung.

In dieser Publikation wurde kein primärer Endpunkt definiert.

Leung et al. (61) verglichen die Fünf-Jahres-Überlebensraten und die rezidivfreien Zeiten bei laparoskopischer und laparotomischer Resektion eines Rektosigmoidkarzinoms. Es wurden keine Unterschiede gefunden. Die Autoren schließen aus ihren Ergebnissen, dass der laparoskopische Eingriff bei Rektosigmoidkarzinomresektion in Bezug auf die Überlebensrate und rezidivfreie Zeit der Lapatomie nicht nachsteht. Die unmittelbar postoperativ gefundenen Ergebnisse, weniger Schmerzen, früheres Wiedererlangen der Darmtätigkeit und schnellere Mobilisation lassen die Autoren den laparoskopischen Eingriff bevorzugen.

In dieser Arbeit wurde in den Textabschnitten kein primärer Endpunkt beschrieben und explizit festgelegt. Es findet sich lediglich im Patientenflussdiagramm Bezeichnung „primary endpoints“. Es scheinen somit mindestens zwei Endpunkte getestet worden zu sein. Für diese Form der multiplen Testung wurde keine Korrekturmethode angegeben. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen die Ergebnisse von 66 der 403 randomisierten Patienten nicht in die Auswertung ein.

Van Leth et al. (94) verglichen die Häufigkeit des Therapieversagens unter Nevirapin (bei einmal und zweimal täglicher Gabe), Efavirenz sowie eine Kombination der beiden bei gleichzeitiger Therapie mit Stavudin und Lamivudin in der Therapie von HIV-Infektionen untereinander. Für die unterschiedlichen Dosen von Nevirapin wurden keine Unterschiede gefunden. Die Autoren geben an, dass keine Evidenz für eine Überlegenheit von Efavirenz gegenüber Nevirapin gefunden wurde, obwohl die Äquivalenz innerhalb einer a priori

festgelegten 10% Grenze nicht gezeigt werden konnte.[1] Die Kombination von Nevirapin und Efavirenz zeigte eine erhöhte Versagerquote gegenüber der Gabe von Efavirenz allein, jedoch nicht höher als bei der Gabe von Nevirapin einmal täglich. Die Autoren schließen aus ihren Ergebnissen, dass Nevirapin und Efavirenz einzeln gegeben gleichermaßen geeignet für eine antiretrovirale Therpie seien. Eine Kombination der beiden sei nicht angezeigt. Bei der alleinigen Gabe von Nevirapin gäbe es keine Unterschiede zwischen einmaliger und zweimaliger täglicher Gabe.

Positiv zu vermerken ist, dass dies eine der wenigen Publikationen ist, die in Bezug auf den Fehler erster Art keine Mängel aufweisen.

Sakkers et al. (79) untersuchten die Wirksamkeit von Olpadronat in der Therapie bei Kindern mit Osteogenesis imperfekta über zwei Jahre. Es wurden Unterschiede in der dualen Röntgen-Absorptiometrie und in der Frakturrate gefunden. Eine Veränderung einer Reihe funktioneller Parameter wurde nicht gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Olpadronat für die Therapie von Kindern mit Osteogenesis imperfekta geeignet ist, weisen jedoch darauf hin, dass das therapeutische Fenster für dieses Medikament noch nicht ausreichend bestimmt sei. In Bezug auf die funktionellen Ergebnisse geben die Autoren zu bedenken, dass es möglicherweise eine zu große Varianz bei den Messungen gab, dass zu wenig Kinder rekrutiert worden sein könnten und dass die Interventionsdauer zu kurz gewesen sein könnte im Hinblick auf die funktionellen Ergebnisse.

In dieser Publikation werden drei primäre Endpunkte dargestellt, ohne dass für diese Form der multiplen Testung eine Korrekturmethode angegeben wird. Es wird nicht erwähnt, welches Signifikanzniveau gewählt wurde.

[1] „Although, overall, treatment failure was numerically lower in the efavirenz group than in the nevirapine-only groups, our findings show no evidence that efavirenz is superior to nevirapine twice daily in terms of treatment failure. However, we could not show equivalence within the 10% limits of these treatment groups even though the study was adequately powered for such an analysis."

Halliday et al. (45) untersuchten die Wirkung der Karotiden-Endarterektomie auf die perioperative Morbidität und Mortalität, sowie auf die Inzidenz von nicht perioperativen Schlaganfällen bei Patienten mit Karotisstenose. Die vorliegende Publikation stellt vorläufige Ergebnisse (5-Jahres-Follow-up bei einer nicht genau angegebenen geplanten Studiendauer) dar. Es wurden Unterschiede zwischen Patienten mit und ohne Karotiden-Endarterektomie gefunden. Die Autoren schließen aus ihren Ergebnisse, dass die Karotiden-Endarterektomie, wenn erfolgreich angewandt, eine geeignete Möglichkeit ist, um Patienten mit Karotisstenose zu behandeln. Die Autoren geben zu bedenken, dass fehlschlagende Operationen großen Schaden anrichten könnten, so dass eine sorgfältige Risiko-Nutzen-Abwägung notwendig sei. Diese Abwägung sei bei den vorhandenen 5-Jahres-Daten noch nicht in ausreichendem Maße möglich.
Bei dieser Publikation handelt es sich nicht um einen Endbericht, sondern um die Veröffentlichung der 5-Jahres-Daten. Im Rahmen der Publikation wird die Untersuchung dreier primärer Endpunkte dargestellt, ohne dass hierfür korrigiert wird. Weiterhin geben die Autoren nicht an, auf welchem Signifikanzniveau getestet wurde.

Singahl et al. (83) untersuchten in einem geplanten Follow-up ihrer zwei Studien von 1982-85 die Auswirkung von Muttermilchfütterung auf das Lipoproteinprofil (insbesondere das Verhältnis von LDL zu HDL) bei ehemals frühgeborenen Jugendlichen. Es wurde ein Unterschied zur Kontrollgruppe gefunden, die Muttermilchersatznahrung erhielt. Frühgeborene, die Muttermilch erhielten, hatten im Follow-up ein niedrigeres Verhältnis von LDL zu HDL. Die Autoren schließen aus ihren Ergebnissen, dass Muttermilchfütterung bei Frühgeborenen das spätere Lipoproteinprofil positiv beeinflusst und die Säuglingsernährung eine Schlüsselrolle in der Entwicklung späterer atherosklerotisch-kardiovaskulärer Erkrankungen spielt.
Die Autoren führten zwei parallele Studien durch, deren gemeinsames Follow-up in der vorliegenden Publikation dargestellt wird. Es handelt sich folglich nicht um den Endbericht einer klinischen Studie.

Barnes et al. (13) untersuchten die Effektivität von rektal appliziertem Artesunat in der Therapie von mittelgradig schwerer Malaria (Plasmodium falciparum) im Vergleich zu parenteral appliziertem Quinin. Es stellte sich heraus, dass die Parasitämie nach zwölf Stunden bei der Artesunat-Gruppe ähnlich häufig unter 60% lag, wie in der Quinin-Gruppe. Die Autoren schließen aus ihren Ergebnissen, dass Artesunat-Suppositorien für die Therapie einer mittelgradig schweren Malaria geeignet sind.
Dies ist eine der wenigen Publikationen, in denen sich keine Hinweise auf eine Inflation des Fehlers erster Art finden.

Andrews et al. (11) verglichen die Überlebensrate von Patienten mit solitären oder multiplen Hirnmetastasen nach Hirnbestrahlung mit und ohne stereotaktischem radiochirurgischem Boost. Für die Gruppe mit solitären Hirnmetastasen wurde eine Verbesserung der Überlebensrate gefunden. Die Autoren schließen aus ihren Ergebnissen, dass ein radiochirurgischer Boost für Patienten mit solitären Hirnmetastasen und therapeutischer Hirnbestrahlung geeignet sei, die Überlebensrate zu erhöhen.
In dieser Studie wurde der primäre Endpunkt (das Überleben der Patienten) an zwei Patientengruppen getestet (Patienten mit solitären und multiplen Hirnmetastasen). Die Autoren geben explizit an, dass hierfür nicht korrigiert wurde[1]. Für darüber hinausgehende Subgruppenanalysen geben die Autoren jedoch an, mit einem P=0,0056 korrigiert zu haben. Es wurden zwei Interimanalysen durchgeführt, ohne dass hierfür eine Korrekturmethode angegeben wird.

Anand et al. (10) untersuchten die Effekte von Morphin-Analgesie bei beatmeten Frühgeborenen. Es wurden keine Unterschiede in Bezug auf intraventrikuläre Blutungen, periventrikuläre Leukomalazie oder Tod eines Neugeborenen gefunden. Die Autoren schließen aus ihren Ergebnissen, dass kontinuierliche Morphininfusion für beatmete Neugeborene keine Veränderung der Variablen im primären Endpunkt bewirken, geben jedoch zu bedenken,

[1] „We treated al outcomes as independent hypotheses, and we did not adjust for multiple comparisons." (S. 1668)

dass intravenöse Morphintherapie bei beatmeten Neugeborenen wie jede potenten Therapie vernünftig und vorsichtig angewendet werden sollte.
Dies ist eine der wenigen Publikationen in denen sich keine Hinweise auf eine Inflation des Fehlers erster Art finden.

Brooks et al. (16) untersuchten die Effekte von Zinkgabe auf die Dauer einer antimikrobiell behandelten, schweren, bakteriellen Pneumonie bei Kindern unter fünf Jahren. Es wurde eine Verkürzung der Erholungszeit unter Zinkgabe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die Gabe von Zink zusätzlich zu einer antimikrobiellen Therapie eine klinisch relevante Verkürzung der Erholungszeit von schwerer, bakterieller Pneumonie bewirkt.
Es wird kein primärer Endpunkt festgelegt. Lediglich im Patientenflussdiagramm ist die Rede von primären Ergebnissen. In der Publikation befindet sich kein expliziter Statistikteil.

Myles et al. (70) untersuchten den Effekt einer Bispektral-Index Überwachung (BIS) auf die Häufigkeit der Bewusstwerdung während einer Allgemeinanästhesie mit anschließender Erinnerung an dieses Ereignis. In der Gruppe mit Bispektral-Index Überwachung war das Risiko einer Bewusstwerdung während Anästhesie um 82% geringer als in der Gruppe ohne Bispektral-Index Überwachung. Die Autoren schließen aus ihren Ergebnissen, dass die Überwachung einer Allgemeinanästhesie mit dem Bispektral-Index die Inzidenz der Bewusstwerdung während Anästhesie reduzieren kann.
Die Autoren geben an, eine Intention-to-treat Analyse durchzuführen. Sie beschreiben jedoch nicht, wie 65 der insgesamt 2503 randomiserten Patienten mit in die Auswertung eingeschlossen werden, die keine Fragebögen zur Messung des primären Endpunktes ausfüllten. Im Patientenflussdiagramm befindet sich außerdem ein Fehler in Bezug auf die Patienten der Kontrollgruppe: zehn Patienten verschwinden im Verlauf ohne dass sich hierfür eine Erklärung findet.

Brouwer er al. (17) verglichen die Wirksamkeit verschiedener antimykotischer Therapien auf eine HIV-assoziierte Cryptokokken-Meningitis. Es wurden

Unterschiede gefunden. Die Kombination von Amphotericin B plus Flucytosin hatte eine stärkere antimykotische Wirkung als Amphotericin B plus Fluconazol und als Amphotericin B plus Flucytosin plus Fluconzol. Die Autoren schließen aus ihren Ergebnissen, dass die Kombination von Amphotericin B plus Flucytosin bei der Therapie der Cryptokokken-Meningitis zu bevorzugen sei.
In dieser Publikation werden Mehrfachvergleiche in Bezug auf den primären Endpunkt angeführt, ohne dass hierfür eine Korrekturmethode beschrieben wird. Ferner wird nicht erwähnt, welches Signifikanzniveau gewählt wurde.

Alloueche et al. (3) untersuchten das Sicherheitsprofil von Chlorproguanil-Dapson in der Therapie der unkomplizierten Plasmodium falciparum Malaria. Als Kontrolle diente die Behandlung mit Sulfadoxin-pyrimethamin. In der Chlorproguanil-Dapson-Gruppe zeigten sich häufiger körperliche Symptome, die mit der Medikation in Verbindung gebracht wurden. Es zeigt sich in dieser Publikation eine deutliche Diskrepanz zwischen Überschrift, hauptsächlichem Inhalt und dem primären Ziel der Studie. In Überschrift und großen Teilen der Publikation gehen die Autoren auf Unterschiede in der Wirksamkeit zwischen Chlorproguanil-Dapson und Sulfadoxin-pyrimethamin ein, wohingegen als primäres Ziel die Evaluation des Sicherheitsprofils von Chlorproguanil-Dapson angegeben wird.
In dieser Publikation wird kein primärer Endpunkt im engeren Sinn definiert. Die Autoren beschreiben lediglich das primäre Ziel der Studie („primary aim of the study“) als Evaluation des Sicherheitsprofils ohne genauere Angaben dazu zu machen, was genau wie überprüft werden soll.

To et al. (89) prüften, ob eine Gebärmutterhalscerclage (mit Shirodkar-Naht) die Rate der Geburten vor vollendeter 33er Schwangerschaftswoche reduzieren kann. Es wurde keine Reduktion der Frühgeburtenrate vor vollendeter 33er Schwangerschaftswoche gefunden. Die Autoren schließen aus ihren Ergebnissen, dass eine Gebärmutterhalscerclage (mit Shirodkar-Naht) das Risiko einer Frühgeburt vor vollendeter 33er Schwangerschaftswoche nicht verringere.

Es werden zwei Interimanalysen durchgeführt, ohne dass hierfür eine Korrekturmethode angeben wird.

Barwell et al. (14) verglichen die Wirksamkeit der Kombination eines oberflächlichen gefäßchirurgischen Eingriffs und anschließender Anlage eines Kompressionsverbandes mit dem alleinigen Anlegen eines Kompressionsverbandes in Bezug auf die Heilung (innerhalb von 24 Wochen) und das erneute Auftreten (innerhalb eines Jahres) chronischer venöser Ulzerationen. In der Gruppe mit zusätzlichem gefäßchirurgischem Eingriff war die Rate der erneut auftretenden chronischen venösen Ulzerationen niedriger als in der Gruppe mit Kompressionsverband allein. In Bezug auf die Heilungsrate wurden keine Unterschiede gefunden. Die Autoren schließen aus ihren Ergebnissen, dass oberflächliche gefäßchirurgische Eingriffe keinen Vorteil in Bezug auf die Heilungsrate bei chronischen venösen Ulzerationen bringen aber einen positiven Einfluss auf die Rezidivrate haben.
In dieser Publikation werden zwei primäre Endpunkte dargestellt, ohne dass hierfür eine Korrekturmethode angegeben wird. Es werden Subgruppenanalysen bezüglich der primären Endpunkte durchgeführt, ohne dass hierfür Korrekturmethoden angeben werden. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen 74 der insgesamt 500 randomisierten Patienten nicht in die Auswertung des Endpunktes bezüglich der Rezidivrate ein.

Althabe et al. (5) untersuchten die Wirkung eines obligatorisch einzuholenden Zweitgutachtens in Bezug auf die Rate von Kaiserschnitten in lateinamerikanischen Krankenhäusern. Es wurde eine Reduktion der Rate von Kaiserschnitten bei obligatorisch einzuholendem Zweitgutachten gefunden. Da die Reduktion kleiner war als in der Hypothese formuliert (7,3% statt 25% relative Reduktion) schließen die Autoren aus ihren Ergebnissen, dass die Umsetzung des obligatorisch einzuholenden Zweitgutachtens davon abhängt, ob das jeweilige Krankenhaus die Ausgaben für diese Reduktion für gerechtfertigt hält.

In Bezug auf eine Inflation des Fehlers erster Art wurden bei dieser Publikation keine Hinweise gefunden festgestellt.

McCarey et al. (64) untersuchten die Wirkung von Atorvastatin auf die DAS28 (Disease activity score) bei Patienten mit rheumatoider Arthritis. Es wurde ein kleiner, aber signifikanter (p=0,004) Unterschied zur Kontrollgruppe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Stoffwechselwege, die von Statinen beeinflussbar sind, Möglichkeiten zur Therapie entzündlicher Erkrankungen seien.
In dieser Publikation fanden sich keine Hinweise auf eine Inflation des Fehlers erster Art.

Julius et al. (57) verglichen die Wirksamkeit von Valsartan und Amlodipin auf die Verzögerung des erstmaligen Auftretens kardialer Morbidität und Mortalität bei hypertensiven Risikopatienten für kardiovaskuläre Erkrankungen. Es wurden keine Unterschiede gefunden.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen 68 der 15313 randomisierten Patienten nicht in die Endauswertung ein.

Courtney et al. (23) untersuchten die Wirkung von Donepezil auf die Häufigkeit der Notwendigkeit institutioneller Pflege und das Fortschreiten der Behinderung von Patienten mit Morbus Alzheimer. Es wurden im Vergleich zur Plazebogruppe keine Unterschiede in Bezug auf die Notwendigkeit institutioneller Pflege gefunden. Es wurde eine Verbesserung auf der „Bristol activity of daily living score“ (BADLS) nach mehr als zwölf Wochen um durchschnittlich einen Punkt unter Donezepil festgestellt. Die Autoren schließen aus ihren Ergebnissen, dass Donezepil nicht geeignet sei, die Kosten in der Behandlung von Patienten mit Morbus Alzheimer zu senken. Obwohl die Autoren zunächst angeben, die Verbesserung auf der BADLS als einen primären Endpunkt zu werten, gehen sie in der Interpretation der Ergebnisse hauptsächlich auf die Kosteneffektivität ein.
Es werden zwei primäre Endpunkte untersucht, ohne dass hierfür eine Korrekturmethode angegeben wird. Es wird eine Mehrfachtestung in Bezug auf

die primären Endpunkte durchgeführt, da es zwei Studienphasen gibt, und am Ende der ersten Phase bereits die primären Endpunkte überprüft werden. Für dieses Vorgehen wird keine Korrekturmethode angegeben.

Koblin et al. (60) untersuchten, ob Verhaltensinterventionen die Rate neu auftretender HIV-Infektionen bei Männern reduzieren, die gleichgeschlechtlichen Sexualkontakt haben. Es wurde eine Risikoreduktion für neu auftretende HIV-Infektionen von 18,2% gegenüber der Kontrollgruppe gefunden. Die Autoren schließen aus diesen Ergebnissen, dass Verhaltensinterventionen eine Risikoreduktion für neu auftretende HIV-Infektionen bei Risikopopulationen zumindest kurzfristig erzielen können. Einschränkend geben die Autoren zu bedenken, dass vor allem Langzeiteffekte erreicht werden müssten.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, geht ein Patient der 4296 randomisierten Patienten nicht in die Auswertung ein. Es fehlt weiterhin eine Angabe über das gewählte Signifikanzniveau, auf dem getestet wurde.

Wollert et al. (98) untersuchten die Wirkung eines intracoronaren Transfers autologer Knochenmarksstammzellen auf die Funktionsregeneration des linken Herzventrikels nach Myokardinfarkt anhand der LVEF (linksventrikuläre Ejektionsfraktion). Es wurde ein Unterschied von sechs Prozent in der LVEF nach sechs Monaten zwischen der Knochenmarkszell-Gruppe und der Kontrollgruppe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass autologe Knochenmarksstammzellen die Regeneration der Funktion des linken Herzventrikels nach Myokardinfarkt verbessern. Sie betonen, dass es sich um die Untersuchung eines Surrogatendpunktes handelt, und dass größere Studien, die klinische Endpunkte wie Herzversagen und Überleben untersuchen, notwendig seien.
Fünf der 65 randomisierten Patienten gingen nicht in die Auswertung ein, da bei ihnen wegen Klaustrophobie oder Adipositas keine Magnetresonanztomographie durchgeführt werden konnte, obwohl die Autoren

angeben, dass alle Patienten nach erfolgter Randomisierung eine Magnetresonanztomographie erhalten hätten[1].

Grigor et al. (44) untersuchten den Effekt von straffen Kontrollen der Krankheitsaktivität als Behandlungsstrategie für rheumatoide Arthritis auf den Verlauf dieser Erkrankung. Unter straffer Kontrolle der Krankheitsaktivität sprachen mehr Patienten auf die Behandlung an und es gab eine stärkere Reduktion auf der Krankheitsaktivitätsskala. Die Autoren schließen aus ihren Ergebnissen, dass straffere Kontrollen der Krankheitsaktivität als Behandlungsstrategie bei rheumatoider Arthritis möglich und außerdem vorteilhaft für die Patienten seien.
Es wurden zwei primäre Endpunkte getestet. Es handelt sich um eine der wenigen Publikationen, in der explizit eine Korrekturmethode (jeweils $p<0,01$ für beide primäre Endpunkte) angegeben wird.

Diener et al. (27) verglichen die Kombination von Clopidogrel und Aspirin mit der alleinigen Gabe von Clopidogrel nach kürzlich aufgetretenem ischämischem Schlaganfall oder transienter ischämischer Attacke bei Hochrisikopatienten. Es wurden keine als signifikant eingestuften Unterschiede gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die zusätzliche Gabe von Aspirin zu Clopidogrel bei Hochrisikopatienten mit kürzlich aufgetretenem ischämischem Schlaganfall oder transienter ischämischer Attacke einer Risiko-Nutzen-Abwägung nicht stand hält und somit nicht gegeben werden sollte.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen gehen nicht alle Patienten in die Auswertung ein.

Stephenson et al. (86) verglichen die Effekte von Sexualerziehung durch Lehrer mit der durch Peers auf das Auftreten von ungeschütztem Geschlechtsverkehr vor dem Vollenden des 16. Lebensjahres. Es wurde kein Unterschied in Bezug auf diesen primären Endpunkt gefunden. Die Autoren beziehen sich in der Diskussion hauptsächlich auf die sekundären Endpunkte, z.B. auf die

[1] Zunächst geben die Autoren an: „After randomisation, all patients underwent cardiac MRI" (S. 142) und dann „After randomisation, five patients were withdrawn because they could not undergo cardiac MRI, either because of claustrophobia or severe obesity" (S. 144).

Zufriedenheit der Schüler mit der Sexualerziehung und schließen aus ihren Ergebnissen, dass Sexualaufklärung durch Peers Vorteile gegenüber der durch Lehrer biete.
Es handelt sich bei dieser Publikation nicht um einen Endbericht sondern um die Veröffentlichung der Ergebnisse der ersten Phase der Studie. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen nicht alle Patienten in die Auswertung ein. Es gingen von insgesamt 29 randomisierten Schulen drei komplette Schulen und 828 zusätzliche Schüler von insgesamt ca. 9508 Schülern nicht in die Auswertung ein. Die genaue Anzahl randomisierter Schüler lässt sich nicht nachvollziehen, da von zwei Schulen, die nach erfolgter Randomisation aus der Analyse ausgeschlossen werden keine Angabe der Schülerzahl gemacht wird.

Anand et al. (9) untersuchten die Langzeiteffekte von Darusentan (selektiver Endothelin A Blocker) auf das Remodelling des linken Herzventrikels bei Patienten mit chronischem Herzversagen. In Bezug auf den primären Endpunkt, das endsystolische Volumen im linken Herzventrikel, wurden keine Unterschiede zur Kontrollgruppe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass eine chronische selektive Endothelin A Blockade bei Patienten mit chronischem Herzversagen das Remodelling nicht abmildert oder die Symptome verbessert.
In dieser Studie wird ein Endpunkt anhand von 6 Studiengruppen miteinander verglichen ohne dass für die Mehrfachvergleiche eine Korrekturmethode angegeben wird.
Es ist unklar, wieso insgesamt 94 Patienten nicht in die Analyse eingehen, obwohl sie die Behandlung abgeschlossen haben. Diese 94 Patienten und weitere 63, die wegen verschiedener Gründe (z.B. Tod, unerwünschte Arzneimittelwirkungen, Rückzug der Teilnahme) nicht in die Hauptauswertung eingingen wurden in einer zweiten Analyse berücksichtigt. In dieser zweiten Analyse wurden ebenfalls keine Unterschiede zwischen den Studiengruppen gefunden. Unklar bleibt außerdem, warum nur bei 590 Patienten die Basismagnetresonanztomographie beurteilt werden konnte, obwohl 642

randomisiert wurden. Über die hierbei aus der Auswertung herausgefallenen Patienten findet sich keine Aussage.

Remuzzi et al. (76) verglichen die Wirksamkeit von Mycophenolat mofetil mit der von Azathioprin in Bezug auf die Prävention akuter Abstoßungen nach Nierentransplantationen. Es wurden keine Unterschiede in Bezug auf den primären Endpunkt gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Mycophenolat mofetil nicht besser als Azathioprin bei der Verhinderung akuter Abstoßungen nach Nierentransplantationen wirke. Die Autoren geben jedoch zu bedenken, dass in ihrer Studie die Wirksamkeit von Mycophenolat mofetil auf spätere Abstoßungen im Gegensatz zu akuten Abstoßungen nicht untersucht wurde.
Die Autoren geben an eine Intention-to-treat-Analyse zu machen, obwohl nicht alle Patienten in die Auswertung eingehen. Zwei der insgesamt 336 randomisierten Patienten gehen nicht in die Auswertung ein, da sie keine Transplantation erhielten. Außerdem vollenden insgesamt 119 weitere der 336 randomisierten Patienten nicht die zweite Phase der Studie (Phase B). Gründe für den Ausschluss waren: Nichterfüllung der Einschlusskriterien für Phase B (56%) und die übrigen 44% Bedenken bezüglich der Sicherheit für die Patienten, Rückzug der Teilnahmebereitschaft, Verstöße gegen das Studienprotokoll, Bedarf von Medikamenten, die vom Studienprotokoll ausgeschlossen waren, Nebenwirkungen oder andere Gründe („practical reasons“).

Thornton et al. (88) verglichen die Häufigkeit von Todesfällen oder Behinderungen bei Feten/Kindern im Alter von zwei Jahren oder älter (korrigiert für das Gestationsalter bei der Geburt), die entweder sofort oder verzögert entbunden wurden, wenn der behandelnde geburtshelferliche Arzt unsicher über den Zeitpunkt der Geburtseinleitung war. In der Diskussion schließen die Autoren aus ihren Ergebnissen, dass der Trend in Richtung häufigere Behinderungen bei Sofortentbindungen die Empfehlung rechtfertige, die Entbindung so lange wie möglich hinauszuzögern.

Es wurden Interimanalysen durchgeführt mit dem Hinweis, dass diese bereits veröffentlicht worden seien, ohne dass hierfür eine Korrekturmethode angegeben wird. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen 13 der insgesamt 588 randomisierten Patienten nicht in die Auswertung ein. In Abbildung 1 (S. 514) wird in der rechten Spalte deutlich, dass ein Patient für die „primary analysis“ verloren geht, ohne dass sich eine Angabe über diesen Ausschluss findet.

Vain et al. (92) untersuchten die Wirksamkeit von oropharyngealer und nasopharyngealer Absaugung vor der Geburt für die Vermeidung des „Mekonium Aspirationssyndroms“ bei Neugeborenen. Es wurde kein Unterschied in der Häufigkeit des Mekonium Aspirationssyndroms zwischen den abgesaugten und den nicht abgesaugten Neugeborenen gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die Empfehlung zur Absaugung aufgehoben werden sollte.

Es ist kritisch anzumerken ist, dass im Verlauf der Studie zwei geplante Interimsanalysen ohne entsprechende Korrektur des Signifikanzniveaus durchgeführt wurden.

Schnitzer et al. (80) verglichen Lumiracoxib mit Naproxan und Ibuprofen in Bezug auf die Entstehung gastrointestinaler Ulzera bei der Therapie von Osteoarthritis und rheumatoider Arthritis. Unter Gabe von Lumiracoxib war die Rate von gastrointestinalen Ulzera niedriger als unter Naproxan und Ibuprofen. Die Autoren schließen aus ihren Ergebnissen, dass die Inzidenz von Ulcuskomplikationen durch eine Behandlung mit Lumiracoxib anstelle von Naproxan und Ibuprofen um bis zu 80% gesenkt werden könne.

Dieses ist die erste von zwei Publikationen im Lancet zu dieser Studie. Es handelt sich somit nicht um einen Endbericht. Die zweite Publikation stammt von Farkouh et al. (siehe unten).

Die Autoren geben an, eine modifizierte Intention-to-treat Analyse durchzuführen, bei der Ereignisse während der ersten 48 Stunden der Behandlung ausgeschlossen werden. Es gehen 39% der Patienten nicht in die Auswertung ein. Begründung für knapp 21% der Patienten, die nicht in die

Auswertung eingehen, ist ein unbefriedigender therapeutischer Effekt. Andere Gründe sind z.B. unerwünschte Arzneimittelwirkungen, Bereitschaft der Teilnahme an der Studie zurückgezogen und Verstoß gegen das Studienprotkoll. Die Ergebnisse dieser ausgeschlossenen Patienten werden in der Publikation nicht veröffentlicht. Positiv anzumerken ist in Bezug auf den Fehler erster Art, dass auf einem Signifikanzniveau von p=0.025 getestet wurde, weil zwei Subgruppen mit einer Kontrollgruppe verglichen wurden.

Farkouh et al. (32) verglichen Lumiracoxib mit Naproxan und Ibuprofen in Bezug auf kardiovaskuläre Morbidität und Mortalität bei der Therapie von Osteoarthritis und rheumatoider Arthritis. Es wurden keine Unterschiede zwischen den Gruppen gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die Gabe von Lumiracoxib kein anderes Risiko für kardiovaskuläre Morbidität und Mortalität darstellt als Naproxan und Ibuprofen.
Bei dieser Arbeit handelt es sich um die zweite Publikation einer Studie, die bereits in derselben Ausgabe des Lancet (Vol 364, August 2004) erschienen ist (siehe oben). Es handelt sich folglich ebenfalls nicht um den Endbericht einer Studie, sondern um eine Teilpublikation.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen 39% der Patienten nicht in die Auswertung ein. Begründung hierfür ist für 21% der Patienten, die nicht in die Auswertung eingehen, ein unbefriedigender therapeutischer Effekt. Die Ergebnisse dieser ausgeschlossenen Patienten werden in der Publikation nicht veröffentlicht. Es werden Mehrfachvergleiche zwischen drei Studiengruppen durchgeführt, ohne dass hierfür eine Korrekturmethode angegeben wird.

Colhoun et al. (21) untersuchten die Wirksamkeit von Atorvastatin bei der Prävention kardiovaskulärer Erkrankungen bei Patienten mit Diabetes mellitus Typ II. Es wurde eine Wirkung festgestellt, aufgrund derer die Studie vorzeitig abgebrochen wurde. Die Autoren schließen aus ihren Ergebnissen, dass Atorvastatin ein sicheres und effizientes Medikament sei, das Risiko für das erste Auftreten einer kardiovaskulären Erkrankung bei Patienten mit Diabetes mellitus Typ II zu reduzieren

Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen 22 der 2841 randomisierten Patienten nicht in die Auswertung ein.

Greenhalgh et al. (43) verglichen die Technik der Endovaskulären Aneurysma Reparatur (EVAR) mit der Technik der offenen Operation bei Patienten mit Aneurysma der Aorta abdominalis in Bezug auf die Mortalitätsrate. Es wurden Unterschiede gefunden. Es starben prozentual weniger Patienten in der EVAR-Gruppe, als in der Gruppe mit offener Operation. Die Autoren betonen, dass es sich erst um Kurzzeitergebnisse handelt und dass diese Ergebnisse lediglich als Lizenz zu verstehen seien, klinische Untersuchungen zum Vergleich der beiden Techniken durchzuführen und nicht etwa, diese Ergebnisse als Rechtfertigung für eine Umsetzung in der klinischen Praxis anzusehen.
Bei dieser Publikation handelt es sich nicht um einen Endbericht, sondern um die Darstellung der der Ergebnisse der ersten 30 postoperativen Tage.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen gehen die Ergebnisse von 35 der 1082 randomisierten Patienten nicht in die Auswertung ein.

Poole-Wilson et al. (75) untersuchten den Langzeiteffekt von Nifedipin auf die Sterblichkeit und kardiovaskuläre Morbidität bei Patienten mit stabiler Angina pectoris, die bereits wegen dieser Erkrankung transdermal oder oral behandelt wurden. Es wurden keine Unterschiede in Bezug auf den primären Endpunkt (ereignisfreies Überlebensintervall bezüglich kardiovaskulärer Ereignisse) gefunden[1]. Die Autoren schließen dennoch aus ihren Ergebnissen, dass Nifedipin für die Langzeitbehandlung von Patienten mit Angina pectoris eingesetzt werden kann, da es sicher sei und kardiovaskuläre Ereignisse hinauszögere[2]. Es wird nicht klar, wie die Autoren zu diesem Schluss kommen.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen 132 der 7797 randomisierten Patienten nicht in die Auswertung ein. Die Angabe, wann der primäre Endpunkt gemessen wird, ist vage. Die Autoren geben an, dass bis zur Studienendvisite gemessen wird, geben jedoch den

[1] „No significant difference was noted with respect to the primary efficacy endpoint." (S. 855)
[2] Nifedipine GITS can be use safely for the long-term treatment of patients with coronary disease and angina pectoris because, in addition to relieving symptoms of angina, it prolongs cardiovascular event and procedure-free survival." (S. 856)

Zeitpunkt dieser Visite in der Publikation nicht an. Es fehlt eine Angabe über die angestrebte Patientenzahl.

Fisher et al. (35) untersuchten das rezidivfreie Überleben und die Gesamtüberlebensrate bei Frauen bei der Behandlung eines lymphknotennegativen, östrogenrezeptorpositiven Brustkrebses. Präsentiert werden die Langzeitergebnisse zweier Studien, „B-14“ und „B-20“, bei denen einmal Tamoxifen mit Placebo („B-14“) und einmal Tamoxifen allein mit zwei Kombination von Tamoxifen plus sequentieller Behandlung mit verschiedenen Chemotherapeutika verglichen werden („B-20“). Es wurden bessere Ergebnisse bei Tamoxifenbehandlung gegenüber Placebo erzielt und bessere Ergebnisse bei Tamoxifen plus sequentieller Chemotherapie verglichen mit Tamoxifenbehandlung allein. Die Autoren geben zu bedenken, dass die Ergebnisse der beiden dargestellten Studien erzielt wurden durch Testung zweitrangiger Effekte (Subgruppenanalysen) und dass somit die Power nicht ausreiche, um Ergebnisse auf dem üblichen Signifikanzniveau zu erzielen. Hieraus ziehen die Autoren die Konsequenz, auch Ergebnisse für unterstützend zu halten, welche sich dem Signifikanzniveau lediglich annäherten.
In der Zusammenfassung der Publikation werden zwei primäre Endpunkte genannt. Im Statistik-Teil der Arbeit werden sie dann nicht mehr als primäre Endpunkte bezeichnet, sondern lediglich als Endpunkte. Für das Vorhandensein zweier primärer Endpunkte geben die Autoren keine Korrekturmethode an. Ferner geben die Autoren nicht bekannt, auf welchem Signifikanznivau getestet wurde.

Manandhar et al. (63) untersuchten den Effekt einer Intervention zum Thema Geburt in Nepal. Primärer Endpunkt war die Neugeborenensterblichkeit. Es wurde eine Reduktion um 30% in der Interventionsgruppe gegenüber der Kontrollgruppe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass diese Form der Intervention geeignet sei, die Neugeborenensterblichkeit in armen und entlegenen Gemeinden zu senken.
Es wurde eine Interimanalyse durchgeführt, ohne dass hierfür korrigiert wurde.

Fernandez-Aviles et al. (33) verglichen eine invasive Behandlungsstrategie innerhalb von 24 Stunden nach Thrombolyse mit einem ischämieabhängigen, konservativen Ansatz bei der Therapie des akuten Myokardinfarktes mit ST-Strecken-Hebung. Es wurden bei der invasiven Behandlungsstrategie weniger Tode, nicht-letale Myokardinfarkte und ischämieinduzierte Revaskulisationen nach einem Jahr gefunden. Die Autoren schließen aus ihren Ergebnissen, dass frühe invasive Intervention gegenüber konservativer Therapie bei Patienten mit akutem Myokardinfarkt überlegen sei.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, gehen 10 der 500 randomisierten Patienten nicht in die Auswertung ein.

Kelly et al. (58) verglichen die Wirksamkeit der Kombination von Tacrolimus und Steroiden mit der Kombination aus Ciclosporin, Steroiden und Azathioprin in der Behandlung von Kindern mit transplantierter Leber. In der Tacrolimus-Gruppe wurde im Vergleich zur Ciclosporin-Gruppe eine um 15,3% niedrigere Rate von akuten Abstoßungen gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Tacrolimus in der Kombination mit Steroiden gut geeignet sei für die Behandlung von lebertransplantierten Kindern.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen gehen vier der insgesamt 185 randomisierten Patienten nicht in die Auswertung ein.

Verweij et al. (96) untersuchten, ob eine Dosisverdopplung von Imatinib bei der Therapie von gastrointestinalen Stromatumoren eine Verlängerung des progressionsfreien Überlebens bewirkt. Es wurde eine Verlängerung des progressionsfreien Überlebens bei Patienten mit doppelter Dosis Imatinib täglich gefunden. Die Autoren schließen aus ihren Ergebnissen, dass eine Erhöhung der Dosis von Imatinib bei der Therapie gastrointestinaler Stromatumoren zu erwägen sei.
Für eine Inflation des Fehlers erster Art fanden sich in dieser Arbeit keine Hinweise.

Addo-Yobo et al. (1) überprüften die Äquivalenz von oral appliziertem Amoxicillin mit intravenös appliziertem Penicillin bei der Therapie der schweren Pneumonie bei Kindern im Alter von drei bis 59 Monaten. Es wurden gleiche Ergebnisse für beide Therapieformen gefunden. Die Autoren schließen aus ihren Ergebnissen, dass oral appliziertes Amoxicillin ebenso wirksam wie intravenös appliziertes Penicillin sei.
In dieser Publikation fanden sich keine Hinweise für eine Inflation des Fehlers erster Art.

Hommes et al. (51) untersuchten die Wirksamkeit intravenöser Immunoglobuline auf die sekundäre Progression der Multiplen Sklerose. Es wurden keine Unterschiede zur Kontrollgruppe mit Placebo-Gabe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die Gabe von intravenösen Immunoglobulinen zur Verhinderung bzw. Hinauszögerung der sekundären Progression bei Patienten mit Multipler Sklerose nicht zu empfehlen sei.
Die Autoren dieser Arbeit geben an, dass sie zwei primäre Endpunkte untersuchen, ohne dass hierfür eine Korrekturmethode angegeben wird.

De Kraker et al. (26) prüften, ob sich die Dauer der postoperativen Chemotherapie bei Kindern mit anaplastischem Wilms-Tumor (Stage 1) ohne Verschlechterung des therapeutischen Ergebnisses verkürzen lässt. Es wurden zwischen der Gruppe mit achtwöchiger postoperativer Chemotherapie und der mit vierwöchiger postoperativer Chemotherapie keine Unterschiede im zweijährigen ereignisfreien Überleben oder im fünfjährigen Gesamtüberleben gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die Reduktion der postoperativen Chemotherapie bei Kindern mit anaplastischem Wilms-Tumor (Stage 1) sinnvoll sei.
Bei dieser Publikation konnten keine Anzeichen für eine Inflation des Fehlers erster Art gefunden werden.

Jindani et al. (55) verglichen drei verschiedene Chemotherapieregime für die Behandlung einer neu diagnostizierten Pulmonaltuberkulose. Es wurden bessere Ergebnisse in Bezug auf den Anteil an Patienten mit negativer

Sputumkultur bei dem sechsmonatigen Regime (zwei Monate Ethambutol, Isoniazid, Rifampizin und Pyrazinamid täglich, gefolgt von vier Monaten täglicher Rifampizin- und Isoniazidgabe) gegenüber den beiden achtmonatigen Regimen gefunden. Die Autoren schließen aus ihren Ergebnissen, dass ihre Studie wichtige Ergebnisse für die Behandlung der Pulmonaltuberkulose ergeben hat.
In dieser Studie wurde in folgender Hinsich mehrfach getestet: Der primäre Endpunkt wurde sowohl zwei, als auch zwölf Monate nach Ende der Behandlung getestet. Außerdem wurden die drei Regime jeweils miteinander verglichen. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen gehen 160 der insgesamt 1355 randomisierten Patienten nicht in die Auswertung mit ein.

Roberts et al. (77) untersuchten den Effekt von intravenös applizierten Kortikosteroiden bei Patienten mit klinisch signifikanter Kopfverletzung auf die Todesrate innerhalb der ersten 14 Tage. Es wurde eine höhere Todesrate unter Kortikosteroidtherapie gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Kortikosteroidgabe bei Patienten mit klinisch signifikanter Kopfverletzung kontraindiziert sei.
Es handelt sich bei der Publikation nicht um einen Endbericht, sondern um die vorzeitige Veröffentlichung der 14-Tage-Ergebnisse der CRASH-Studie. Es wurden zwei primäre Endpunkte untersucht, ohne dass hierfür eine Korrekturmethode angegeben wird.

Crawford et al. (24) untersuchten den Effekt einer Kurzintervention bei Patienten mit Alkoholmissbrauch in einer Notaufnahme. Sechs Monate nach Intervention wurde eine Reduktion der zugeführten Alkoholmenge festgestellt, zwölf Monate nach Intervention war die Reduktion gegenüber der Kontrollgruppe nicht mehr signifikant ($p>0,05$). Die Autoren schließen aus ihren Ergebnissen, dass die Kurzintervention bezüglich des Alkoholmissbrauchs in einer Notaufnahme den Alkoholkonsum vorläufig senkt.
In dieser Studie wird der primäre Endpunkt lediglich im Patientenflussdiagramm erwähnt, nicht jedoch im Methodenteil definiert.

Alonson et al. (4) untersuchten die Effektivität, Immunogenität und Sicherheit von RT,S/AS02A (präerythrozytärer Impfstoff gegen Plasmodium Falciparum Malaria) bei afrikanischen Kindern. Im Vergleich zur Kontrollgruppe wurden weniger Nebenwirkungen und weniger Infektionsepisoden unter RT,S/AS02A gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die Möglichkeit bestünde, vor diesem Hintergrund neue Malaria-Impfstoffe zu entwickeln.
Es wird eine Interimanalyse durchgeführt, ohne dass hierfür korrigiert wird. Es findet sich keine Angabe über das gewählte Signifikanzniveau.

Filippi et al. (34) untersuchten, ob Interferon beta-1a bei der Behandlung der Multiplen Sklerose die Umwandlung zu einer klinisch gesicherter Multiplen Sklerose verringert. Es wurde eine geringere Umwandlungsrate unter Interferon beta-1a gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Interferon beta-1a die Umwandlung zu gesicherter Mulipler Sklerose reduziert.
Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen gehen 46 der 309 randomisierten Patienten nicht in die Auswertung ein. Es findet sich außerdem keine Angabe auf welchem Signifikanzniveau getestet wurde.

Homs et al. (52) verglichen eine einmalige (single dose) Brachytherapie mit der Implantation eines Stents bei der palliativen Behandlung von Dysphagie bei Patienten mit Ösophaguskarzinom. Die Dysphagie wurde durch die Stentimplantation initial schneller gebessert, nach 30 Tagen jedoch brachte die Brachytherapie eine größere Linderung der Beschwerden. Die Autoren schließen aus ihren Ergebnissen, dass die einmalige (single dose) Brachytherapie der Stentimplantation vorzuziehen sei bei der palliativen Behandlung von Dysphagie bei Patienten mit Ösophaguskarzinom. Eine Ausnahme seien Patienten mit kurzer Lebenserwartung oder solche, bei denen der Tumor nach Brachytherapie weiter wächst.
In dieser Arbeit fanden sich keine Hinweise auf eine Inflation des Fehlers erster Art.

Israel et al. (54) verglichen die Nebenwirkungen von Albuterol (beta-2-adrenerger Agonist) bei Asthmapatienten mit Arg/Arg- und Gly/Gly-Genotyp. Bei Patienten mit Arg/Arg-Genotyp verbesserte sich die PEFR (maximale Expirationsflussrate), wenn das Albuterol abgesetzt wurde im Gegensatz zu Patienten mit Gly/Gly-Genotyp. Die Autoren schließen aus ihren Ergebnissen, dass Patienten mit Arg/Arg-Genotyp vom Absetzen des Albuterols profitieren würden.
In dieser Studie wurde mehrfach getestet, indem zwei Subgruppen gegen Placebo getestet wurden, ohne dass hierfür korrigiert wurde. Außerdem geben die Autoren an, dass der primäre Endpunkt mehrfach gemessen wurde ohne Korrektur hierfür. Es gingen zwölf der 180 randomisierten Patienten nicht in die Auswertung ein, obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen.

El Arifeen et al. (30) untersuchten das integrierte Management bei Krankheiten im Kindesalter (Integrated Management of Childhood Illness, IMCI) in Bangladesh. Es wurde eine Verbesserung der medizinischen Versorgung der Kinder mit integriertem Management gefunden. Die Autoren schließen aus ihren Ergebnissen, dass eine volle Implementierung des integrierten Managements bei Krankheiten im Kindesalter zu einer verbesserten Nutzung des öffentlichen Gesundheitswesens führen könne.
Diese Publikation ist kein Endbericht, sondern stellt die Ergebnisse einer Interimanalyse dar. Es wird keine Korrekturmethode für diese Interimsanalyse angegeben. Es findet sich bezüglich der mit dem Extraktionsbogen zu überprüfenden Items lediglich eine Angabe zum gewählten Signifikanzniveau, andere Angaben fehlen.

Fletcher et al. (37) untersuchten, ob die Anwendung eines multidimensionalen, geriatrischen Assessments auf alle Patienten einer Allgemeinarztpraxis (General Practitioner) in Großbritannien Vorteile gegenüber der gezielten Anwendung auf schwerkranke Patienten hat. Es wurden keine Unterschiede gefunden. Die Autoren schließen aus ihren Ergebnissen, dass eine universelle Anwendung des multidimensionalen, geriatrischen Assessments auf alle

Patienten einer Allgemeinarztpraxis keinen Vorteil gegenüber der Anwendung auf schwerkranke Patienten habe.
In dieser Publikation geben die Autoren an, drei primäre Endpunkte zu untersuchen. Allerdings wird auf einem Signifikanzniveau von p=0,01 getestet, was als Korrektur für das Vorhandensein von mehr als einem primären Endpunkt gewertet werden kann. Die Autoren geben an, eine Intention-to-treat Analyse durchzuführen. Aus der Publikation ist jedoch nicht eindeutig ersichtlich, ob alle Patienten in die Auswertung eingehen.

Harper et al. (46) untersuchten die Effizienz eines bivalenten Impfstoffes mit L1-Virus ähnlichen Partikeln in der Prävention von Infektionen mit Humanen Papillomaviren (HPV) vom Typ 16 und 18. Es wurden Unterschiede zur Placebo-Kontrollgruppe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass der L1-Impfstoff hochwirksam in der Prävention von HPV Typ 16 und 18 Infektionen sei.
Es fanden sich in dieser Arbeit keine Anzeichen für eine Inflation des Fehlers erster Art.

Smeets et al. (84) verglichen die chirurgische Exzision und Mohs mikrografische Chirurgie für Basalzellkarzinome des Gesichts in Bezug auf das Auftreten von Rezidiven. Es wurden keine signifikanten Unterschiede auf einem 5%-Signifikanzniveau gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Mohs mikrografische Chirurgie die Rezidivrate für Basalzellkarzinome ein wenig aber nicht signifikant erniedrigt gegenüber der chirurgischen Exzision.
In dieser Studie wurde der primäre Endpunkt an zwei Subgruppen untersucht, ohne dass hierfür eine Korrekturmethode angegeben wird.

Lux et al. (62) verglichen die Wirksamkeit einer hormonellen Therapie mit der Gabe von Vigabatrin in Bezug auf infantile Spasmen nach 14 Tagen. Es wurden Unterschiede gefunden. Die Autoren schließen aus ihren Ergebnissen, dass eine hormonelle Therapie die Häufigkeit von infantilen Spasmen stärker herabsetzt als die Gabe von Vigabatrin.

Es wurden mehr als zwei Studiengruppen miteinander verglichen, ohne dass hierfür eine Korrekturmethode angegeben wird. Es fehlt eine Angabe über das gewählte Signifikanzniveau.

Chintu et al. (19) untersuchten die Wirksamkeit von Co-trimoxazol als Prophylaxe gegen opportunistische Infektionen bei HIV-infizierten Kindern in Zambia. Es wurden deutliche Unterschiede zur Plazebo-Kontrollgruppe gefunden, die Studie wurde vorzeitig abgebrochen. Die Autoren schließen aus ihren Ergebnissen, dass alle HIV-positiven Kinder in Afrika eine Co-trimoxazol Prophylaxe erhalten sollten.
Es gibt in dieser Studie zwei primäre Endpunkte, ohne dass hierfür korrigiert wird. Es fanden eine Interimanalyse und Subgruppenanalysen statt, ohne dass hierfür korrigiert wurde. Es wurden sieben Patienten nach erfolgter Randomisierung ausgeschlossen, obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen.

Gournay et al. (42) untersuchten die Wirksamkeit einer prophylaktischen Gabe von Ibuprofen auf den Verschluss des Ductus arteriosus bei Frühgeborenen. Es wurden Unterschiede zur Placebo-Kontrollgruppe gefunden. Obwohl sich die prophylaktische Gabe von Ibuprofen positiv auf den Verschluss des Ductus arteriosus auswirkte, schließen die Autoren aus ihren Ergebnissen, dass die kurative Gabe von Ibuprofen der prophylaktischen Gabe aufgrund der in der Studie aufgetretenen Nebenwirkungen vorzuziehen sei. Die Studie wurde aufgrund der aufgetretenen Nebenwirkungen frühzeitig abgebrochen.
In dieser Arbeit konnten keine Hinweise für eine Inflation des Fehlers erster Art gefunden werden.

Van Overmeire et al. (95) untersuchten die Wirksamkeit einer prophylaktischen Gabe von Ibuprofen an Frühgeborene auf die Entwicklung schwerer intraventrikulärer Hämorrhagien. Es wurden keine Unterschiede zur Placebo-Kontrollgruppe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass die prophylaktische Gabe von Ibuprofen an Frühgeborene die Entwicklung schwerer intraventrikulärer Hämorrhagien nicht beeinflusse.

Es wurde eine Interimanalyse durchgeführt, ohne dass hierfür korrigiert wurde. Es findet sich keine Angabe zum gewählten Signifikanzniveau.

Staedke et al. (85) Untersuchten die Wirksamkeit verschiedener Kombinationstherapien (Chloroquin+Sulfadocin+Pyrimethamin vs. Amodiaquin+Sulfadoxin+Pyrimethamin vs. Amodiaquin+Artesunat) bei unkomplizierter Malaria (Plasmodium falciparum). Es wurden Unterschiede gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Amodiaquin+Sulfadoxin+Pyrimethamin die am besten geeignete Kombination in Regionen Afrikas sei, in denen die Resistenz gegenüber dieser Kombination niedrig bleibt.
Es werden drei primäre Endpunkte untersucht, ohne dass hierfür korrigiert wird. Es werden eine Interimanalyse durchgeführt und mehr als zwei Studiengruppen miteinander verglichen, ohne dass hierfür korrigiert wird. Es gingen 34 von 418 randomisierten Patienten nicht mit in die Endauswertung ein, obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen.

Morris et al. (68) untersuchten, ob Geldzuwendungen an Schwangere und Mütter mit jungen Kindern die Rate der Nutzung präventiver Gesundheitsmaßnahmen erhöhen. Es wurden Unterschiede zur Kontrollgruppe gefunden. Die Autoren schließen aus ihren Ergebnissen, dass Geldzuwendungen an Schwangere und Mütter mit jungen Kindern die Rate der Nutzung präventiver Gesundheitsmaßnahmen erhöht.
Es werden drei primäre Endpunkte untersucht, ohne dass hierfür korrigiert wird. Es werden mehr als zwei Studiengruppen miteinander verglichen, ebenfalls ohne dass hierfür eine Korrekturmethode angegeben wird.

Newnham et al. (71) veröffentlichen ein Follow-up einer Studie, welche die Effekte intensivierter pränataler Ultraschalluntersuchungen auf die Ergebnisparameter in der Kindheit untersuchen. Es wurden Unterschiede gefunden. Die Kinder in der intensivierten Ultraschalluntersuchungsgruppe hatten bei der Geburt eine signifikant ($p=0,011$) geringere Körperlänge. Dieser Unterschied war jedoch bei Untersuchungen zwischen dem 1. - 8. Lebensjahr

nicht mehr vorhanden. Als weiterer Unterschied wurde gefunden, dass Kinder aus der intensivierten Ultraschalluntersuchungsgruppe seltener Auffälligkeiten bei der Sprachentwicklung hatten, als Kinder aus der Kontrollgruppe. Dieses Ergebnis führen die Autoren auf die Möglichkeit zurück, dass die hohe Anzahl an Endpunkten in dieser Studie es wahrscheinlich werden lassen, dass ein Unterschied zufällig signifikant erscheint. Es sei möglich, dass die Sprachentwicklung aufgrund einer veränderten elterlichen Aufmerksamkeit in der intensivierten Ultraschalluntersuchungsgruppe anders verlaufe. Die Autoren schließen insgesamt aus ihren Ergebnissen, dass es keine Evidenz für eine schädliche Wirkung von intensivierten Ultraschalluntersuchungen auf die kindliche Entwicklung gäbe. Trotzdem geben die Autoren zu bedenken, dass es notwendig sei, weitere Untersuchungen zur Sicherheit pränataler Ultraschalluntersuchungen durchzuführen, zumal die Energiestärken, mit denen gemessen werde, zunähmen.

Es handelt sich bei dieser Publikation nicht um einen Endbericht, sondern um ein Follow-up. Obwohl die Autoren angeben, eine Intention-to-treat Analyse durchzuführen, wird aus der Publikation nicht eindeutig klar, ob alle Patienten in die Auswertung eingehen. Es findet sich keine Angabe über das ausgewählte Signifikanzniveau.

5. Diskussion der Methoden

5.1. Der Extraktionsbogen

Der verwendete Extraktionsbogen ist in mancher Hinsicht kritikwürdig. Es handelt sich um einen für diese Arbeit auf der Grundlage eines Fragebogens von Herrn Prof. Dr. Beck-Bornholdt eigens vom erstellten Extraktionsbogen, der somit keine breite Anwendung findet oder in seiner Validität und Reliabilität bereits geprüft wurde. Die Interraterreliabilität wurde stichprobenartig überprüft, indem Extraktionsbögen für mehrere Publikationen vom Autor und von Herrn Prof. Dr. Beck-Bornholdt ausgefüllt wurden und die Ergebnisse miteinander verglichen wurden. Es fand sich eine Übereinstimmung in 90% der Fälle. Es kann daher davon ausgegangen werden, dass der Extraktionsbogen und die Beschreibung der Extraktionsbogen-Items in einer Form vorliegen, in der verschiedene Anwender auf annähernd übereinstimmende Ergebnisse kommen. Dennoch gibt durchaus ernst zu nehmende Abweichungen, die auf eine prinzipielle Einschränkung der in dieser Arbeit verwendeten Methoden deuten. Es handelt sich bei den zu extrahierenden Daten um sprachliche Äußerungen. Es ist davon auszugehen, dass die Informationen, die der Sender einer sprachlichen Nachricht übermitteln will, nicht immer unmissverständlich beim Empfänger ankommen. Vielmehr ist es so, dass man erst über sprachliche Rückkoppelung, d.h. über Nachfragen sich immer näher an einen Konsens herantasten kann. Dieses Nachfragen kann bei schriftlichen Äußerungen nicht unmittelbar stattfinden. Es gab immer wieder mehrdeutige sprachliche Äußerungen in den Publikationen, bei denen es in der jetzigen Gestaltung des Extraktionsbogens nicht immer möglich war, eine eindeutige Entscheidung zu fällen.

Für solche Fälle ergibt sich als Verbesserungsvorschlag für eine überarbeitete Version des Extraktionsbogens das Einführen einer weiteren Möglichkeit neben den Kästchen „ja“, bzw. „vorhanden“ und „nein“, bzw. „nicht vorhanden“, nämlich „nicht beurteilbar“. Der Vorteil dieser dritten Wahlmöglichkeit für die Items wäre, dass abgebildet werden kann, wie häufig eine eindeutige

Entscheidung dem Ausfüllenden unmöglich ist. Nachteilig hingegen wäre die Verlockung, sich nicht festlegen zu müssen, was vielleicht in einigen Fällen bei intensiverem Nachdenken möglich wäre.

Im Laufe der Arbeit mit dem Extraktionsbogen hat sich ein Lerneffekt beim Autor eingestellt. Durch das wiederholte Anwenden des Bogens auf verschiedene Arbeiten wuchs der Erfahrungsschatz, was eindeutige und mehrdeutige Angaben angeht, so dass das Ausfüllen der Items bei Publikationen, die später bearbeitet wurden, differenzierter möglich war. Aus diesem Grund wurde ein zweiter Durchlauf durchgeführt, nachdem einmal alle Publikationen bearbeitet waren. Im zweiten Durchlauf wurden etliche Items verändert, was darauf hindeutet, dass ein Lerneffekt das Ausfüllverhalten verändert. Für eine zukünftige Anwendung des Bogens würde sich ein Probelauf an einer Auswahl anderer Studien empfehlen, bevor die eigentlich zu untersuchenden Publikationen bearbeitet werden.

5.2. Das untersuchte Datenmaterial

Das untersuchte Datenmaterial in dieser Arbeit waren die Publikationen randomisierter kontrollierter Studien, nicht die Studien selbst. Das bedeutet, dass mit der hier vorliegenden Arbeit prinzipiell keine eindeutige Aussage darüber möglich ist, was in den jeweiligen Studien gemacht oder nicht gemacht wurde, sondern es wird lediglich untersucht, ob die Autoren in der Publikation ihre Arbeit so beschreiben, dass es Hinweise auf eine Inflation des Fehlers erster Art gibt. Es besteht grundsätzlich die Möglichkeit, dass die sprachlichen Äußerungen in den Publikationen missverstanden werden und somit der Leser eine andere Information aufnimmt, als die Studie selbst produziert hat. Dies ist ein sprachphilosophischer Aspekt der vorliegenden Arbeit, der hier nicht weiter verfolgt werden soll, dem aber in anderem Rahmen nachgegangen werden könnte.

5.3. Uneindeutigkeit der Informationen

Um beurteilen zu können, ob eine potentielle Inflation des Fehlers erster Art vorliegt, sind unter anderem die in dem Extraktionsbogen vorkommenden Informationen notwendig, der in dieser Arbeit verwendet wurde. In den untersuchten Arbeiten fanden sich jedoch oftmals Informationen, die nicht eindeutig sind. Beispielsweise findet sich in der Arbeit von Leung et al. (61) die Festlegung des primären Endpunkts lediglich implizit im Patientenflussdiagramm, nicht aber im Methodenteil. Wenn sich im Methodenteil keine Angabe dazu findet und dort kein eindeutiger primärer Endpunkt definiert wird, so bleibt die Frage offen, warum dies nicht geschehen ist. Es ist für den Leser nicht möglich, hieraus valide Schlüsse zu ziehen. Aus dieser Verunsicherung heraus erscheint es wünschenswert, dass die im CONSORT-Statement vorgeschlagenen Leitlinien für die Publikation von randomisierten kontrollierten Studien umgesetzt werden und es einen klaren Ort für bestimmte Informationen gibt, so dass man schnell einen sicheren Überblick darüber erhält, was die Autoren schreiben und wozu sie sich vielleicht auch nicht äußern. In der in Anlehnung an das CONSORT-Statement erstellt Checkliste für Autoren des Lancet gibt es für die einzelnen zu erwähnenden Informationen je eine Spalte, in der eine Seitenzahl eingefügt werden kann. In diesem Zusammenhang entstand die Idee, dass es hilfreich für den Leser einer Publikation sein könnte, wenn diese bereits vorhandenen Informationen derart umgesetzt würden, dass beispielsweise bei den einzelnen Items aus der Checkliste (wie z.B. primärer Endpunkt) eine Randnotiz in der Publikation erfolgen könnte. Dies könnte die Eindeutigkeit mancher Aussagen erhöhen, da man nicht überlegen müsste, ob die Äußerung sich auf eine bestimmte Kategorie aus dieser Checkliste bezieht. In der Arbeit von Alloueche et al. (3) ist es vonnöten, zu entscheiden, ob die Autoren einen primären Endpunkt meinen, wenn sie von „primary aim“[1] schreiben. Auch wenn dies zunächst nahe liegt, ist keine Sicherheit dafür gegeben. Es könnte auch sein, dass die Autoren hier nur beschreiben, was das Ziel der Arbeit ist und nicht, was konkret der primäre Endpunkt ist.

[1] Engl. Primäres Ziel, primäres Anliegen (Übers. des Autors)

6. Diskussion der Ergebnisse

6.1. Endberichte

Im Jahrgang 2004 fanden sich anhand der für diese Arbeit definierten Suchkriterien (s. Material und Methoden) 13 Arbeiten, die nicht als Endberichte einer randomisierten Studie bezeichnet werden können, sondern die Ergebnisse mehrerer Studien zusammenfassen oder lediglich einen Teil einer Studie darstellen.

Die Wahrscheinlichkeit für zumindest ein falsch-positives Ergebnis in einer Publikation erhöht sich, wenn die Ergebnisse zweier Studien publiziert werden, da die Möglichkeit für das Auftreten eines Fehlers erster Art für beide Studien getrennt besteht. Auch wenn eine Studie in zwei Publikationen aufgeteilt wird, erhöht sich die Wahrscheinlichkeit für ein falsch-positives Ergebnis. Dies ist beispielsweise wenn verschiedene Daten publiziert werden. z.B., wenn die Daten im Sinne einer Interim-Analyse vorab untersucht werden. Darüber hinaus besteht in diesem Fall für jede Analyse eine getrennte Wahrscheinlichkeit für das Auftreten eines Fehlers erster Art.

Vor dem Hintergrund der Erhöhung der Wahrscheinlichkeit falsch-positiver Ergebnisse ist für eine angemessene Einschätzung der Forschungsergebnisse wichtig, dass der Leser einer Publikation möglichst eindeutig darüber informiert wird, ob es sich um einen Endbericht, eine geteilte oder zusammenfassende Publikation handelt.

Wie am Beispiel von Tran et al. (90) gesehen, ist aus dem Titel der Arbeit („Dihydroartemisinin-piperaquine against multidrug-resistant Plasmodium falciparum malaria in Vietnam: randomised clinical trial“) jedoch nicht ersichtlich, dass es sich um die Publikation zweier Studien handelt.

6.2. Anzahl primärer Endpunkte

In 16 Publikationen der 72 untersuchten Publikationen fanden sich mehr als ein primärer Endpunkt, während nur in drei dieser Arbeiten eine Korrektur für diese multiple Testung angegeben wird. Gerade das Vorhandensein von mehr als einem primären Endpunkt erhöht eindeutig die Wahrscheinlichkeit für eine Inflation des Fehlers erster Art. Das bedeutet, dass in 13 Arbeiten des Jahrgangs 2004 des Lancet das effektive Signifikanzniveau schon aufgrund der Anzahl der primären Endpunkte erhöht ist und die Ergebnisse dieser Arbeiten unsicherer sind als der Leser der Publikation aufgrund der Angabe der Autoren zunächst annimmt. Die Wahrscheinlichkeit für die Richtigkeit der dargestellten Ergebnisse wird somit tendenziell überschätzt.

6.3. Multiple Testung bezüglich eines primären Endpunktes

Auch wenn nur ein primärer Endpunkt untersucht wird, besteht die Möglichkeit, dass eine Mehrfachtestung durchgeführt wird, die zu einer Inflation des Fehlers erster Art führt. In den in dieser Arbeit untersuchten Publikationen war diese Weise nur bei 31 Arbeiten nicht der Fall. In den übrigen 35 Arbeiten mit einem primären Endpunkt findet sich das Problem, dass eine Mehrfachtestung für einen primären Endpunkt durchgeführt wird. Um die Einschränkungen der dargestellten Ergebnisse besser beurteilen zu können, sollten die Autoren auf die Erhöhung des effektiven Signifikanzniveaus hinweisen oder anhand einer durchgeführten Korrekturmethode belegen, dass keine Erhöhung des effektiven Signifikanzniveaus stattfindet. Wenn die Autoren dies offen lassen, so herrscht für den Leser Unklarheit über die Sicherheit der Ergebnisse. Diese Unklarheit findet sich in 24 Arbeiten des Jahrgangs 2004 des Lancet.

6.4. Intention-to-treat Analysen

Hollis und Campbell (49) untersuchten den Umgang mit Intention-to-treat Analysen in den Publikationen des Jahres 1997 in vier der führenden medizinischen Fachzeitschriften (The Lancet, British Medical Journal, Journal of the American Medical Association, New England Journal of Medicine). Die

Autorinnen kommen zu dem Schluss, dass die Intention-to-treat Analysen oft inadäquat beschrieben und durchgeführt würden.

In der vorliegenden Arbeit fanden sich viele Fälle, in denen die Autoren angeben, eine Intention-to-treat Analyse durchzuführen (58 von 72 Publikationen), während nur in 27 Arbeiten beschrieben wird, dass alle Patienten gemäß ihrer Zuordnung zur Studiengruppe ausgewertet wurden. In den übrigen Arbeiten wurden aus verschiedenen Gründen Patienten nach erfolgter Randomisierung ausgeschlossen, was dem Prinzip der Intention-to-treat Analyse zuwider läuft.

Hollis und Campbell (49) beschreiben, dass bezüglich ihrer Untersuchung des Jahrgangs 1997 der Anteil der Publikationen mit expliziter Intention-to-treat Analyse im Vergleich zu 1990 gestiegen sei. Die Autorinnen vermuten, dass dies darauf zurückzuführen sei, dass im Rahmen der evidenzbasierten Medizin die Beurteilung der Forschungsmethoden unterstützt würde und Leitlinien zur kritischen Beurteilung der Forschungsmethoden immer Fragen nach der Vollständigkeit der Patientendaten ebenso enthalten würde, wie die Frage, ob die Patienten in der Gruppe ausgewertet wurden, in die sie randomisiert worden seien. In der Zeitschrift The Lancet ist der Anteil an Publikationen, die laut Angabe der Autoren eine Intention-to-treat Analyse durchführen auch im Vergleich zu 1997 weiter angestiegen. Im Jahr 1997 gaben 54% der Autoren an, eine Intention-to-treat Analyse durchzuführen, in der vorliegenden Untersuchung des Jahrgangs 2004 fand sich eine solche Angabe in 80% der Publikationen. Nach Hollis und Campbell könnte diese generelle Aufmerksamkeit bezüglich Intention-to-treat Analysen dazu geführt haben, dass diese Analysen häufiger inkomplett durchgeführt werden, da in der Studienplanung nicht berücksichtigt wurde, wie mit Abweichungen vom geplanten Studienverlauf umgegangen werden solle. Es gibt keine einheitlichen Regeln, wie mit Verstößen gegen das Protokoll im Rahmen einer Intention-to-treat Analyse umgegangen werden sollte. Zunächst erscheint es evident, dass Patienten, die gegen das Protokoll verstoßen, in ihrer jeweiligen Studiengruppe ausgewertet werden müssen, das verbirgt sich schon in dem Namen Intention-

to-treat, d.h. es genügt die Intention, den Patienten in bestimmter Weise zu behandeln, um ihn in der entsprechenden Gruppe in die Auswertung eingehen zu lassen, er muss nicht tatsächlich gemäß dieser Gruppe behandelt worden sein. Bei näherem Hinsehen können jedoch sinnvollerweise auch Ausnahmen beschrieben werden. Es ist deshalb schwierig eine allgemeingültige Definition des Intention-to-treat Prinzips zu beschreiben.

Hollis et al. (49) beschreiben folgendes Beispiel: wenn ein Patient in einer Studie zum Vergleich einer aktiven und einer Plazebo-Impfung fälschlicherweise die Plazebo-Impfung erhält, so stellt das einen Sonderfall dar, weil diese Verwechslung außerhalb der Studie nicht zustande kommen könnte. Ein solcher Fall ist in den Publikationen, die in der vorliegenden Arbeit untersucht wurden, nicht vorgekommen.

Die Ergebnisse der vorliegenden Arbeit zeigen, dass auch acht Jahre nach Einführung des Consort-Statements immer noch unterschiedliche Vorgehen im Umgang mit Patienten beschrieben werden, welche aus verschiedenen Gründen die Studie nicht so durchlaufen, wie es geplant war. In rund 50% der Arbeiten fand sich ein Patientenausschluss nach erfolgter Randomisierung, obwohl die Autoren angaben, eine Intention-to-treat Analyse durchzuführen. Ein Beispiel ist die Arbeit von Carr et al. (18): in dieser Publikation „verschwinden“ in der Verum-Gruppe im Patientenflussdiagramm fünf Patienten, ohne dass es für den Leser nachvollziehbar ist, aus welchen Gründen diese Patienten ausgeschlossen wurden. Ein solcher Patientenausschluss kann eine Erhöhung des effektiven Signifikanzniveaus bedeuten. Die Autoren selbst diskutieren die möglichen Folgen, namentlich eine höhere Wahrscheinlichkeit für ein falsch-positives Ergebnis, nicht.

Als weiteres Beispiel für einen besonders unverständlichen Grund für einen Patientenausschluss diene die Arbeit von Klareskog et al. (59). Die Autoren geben an, aus den drei Studiengruppen 21, 16 und 6 Patienten aufgrund mangelnder Wirksamkeit („lack of efficacy“) ausgeschlossen zu haben. Ein solcher Patientenausschluss erhöht erheblich die Wahrscheinlichkeit, dass es

zu einer Inflation des Fehlers erster Art kommt. Zwar würde außerhalb einer Studie auch ein Abbruch der Therapie in Frage kommen, wenn die Wirksamkeit ausbleibt, in dieser Studie ist es jedoch ausgesprochenes Ziel, gerade die Wirksamkeit des Medikaments zu untersuchen. Werden dann Patienten ausgeschlossden, bei denen eine Wirksamkeit ausbleibt, so ist eine Verfälschung des Ergebnisses höchst wahrscheinlich. Folge ist auch hier, dass die Wahrscheinlichkeit für ein falsch-positives Ergebnis erhöht ist und es somit zu einer Versorgung von Patienten mit Medikamenten kommt, die nicht optimal auf die entsprechende Situation abgestimmt sind.

In der vorliegenden Arbeit wurde für die untersuchten Publikationen nicht überprüft, ob ein Patientenausschluss sinnvoll war oder nicht. Hollis et al. beschreiben in ihrer Arbeit (49) dass es nur wenige Fälle gibt, in denen ein Patientenausschluss nach erfolgter Randomisierung gerechtfertigt ist (s.o.), und sie beschreiben, dass es verschiedene Meinungen in der Wissenschaftsgemeinde darüber gäbe, wann ein solcher Patientenausschluss gerechtfertigt ist und wann nicht.

In der vorliegenden Arbeit wurde der Patientenausschluss streng beurteilt: wenn die Autoren angeben, eine Intention-to-treat Analyse durchzuführen und ein Patientenausschluss stattfand, dann wurde die Arbeit zu denen gezählt, bei denen es eine potentielle Inflation des Fehlers erster Art gibt aufgrund eines Patientenausschlusses nach erfolgter Randomisierung. Nachteil dieses Vorgehens ist, dass die Inflation des Fehlers erster Art tendenziell überschätzt wird, da in manchen Arbeiten der Patientenausschluss möglicherweise als gerechtfertigt eingestuft werden könnte und keinerlei Einfluss auf den Fehler erster Art haben würde. Um ein genaueres Ergebnis bezüglich des Einflusses von Patientenausschlüssen auf den Fehler erster Art zu erhalten, wäre eine detailliertere Untersuchung notwendig, bei der genau unterschieden wird, ob ein Patientenausschluss gerechtfertigt. Dies hätte jedoch den Rahmen der vorliegenden Arbeit gesprengt.

6.5. Schlussfolgerungen für die Qualität in anderen biomedizinischen Zeitschriften

In der vorliegenden Arbeit wurde ein Jahrgang einer der ältesten biomedizinischen Fachzeitschrift untersucht, die sich mittlerweile unter den fünf Zeitschriften mit dem höchsten Impact Factor befindet. Aus der hier durchgeführten Untersuchung allein lässt sich weder ein Trend zur Qualität innerhalb der Zeitschrift, noch eine Aussage zum Fehler erster Art in biomedizinischen Publikationen allgemein ableiten, da es sich um eine punktuelle Stichprobe handelt. In Zusammenhang mit bereits durchgeführten Untersuchungen lassen sich jedoch einige mögliche Schlussfolgerungen für den Fehler erster Art und dessen Inflation diskutieren.

In einer Übersichtsarbeit von Moher et al. (66) wird untersucht, ob es einen Effekt des Consort-Statement auf die Qualität der Publikation von randomisierten kontrollierten Studien gibt. In dieser Arbeit werden die Jahrgänge zwei Jahre vor und zwei Jahre nach Publikation des Consort-Statement aus vier Zeitschriften (BMJ, JAMA, The Lancet und The New England Journal of Medicine) untersucht. Die Autoren kommen nach ihrer Untersuchung zu dem Schluss, dass die Qualität der Publikationen randomisierter kontrollierter Studien aller untersuchten Zeitschriften sich über die Zeit verbesserte, besonders jedoch, wenn das Consort-Statement von den Zeitschriften übernommen wurde. Die Autoren geben jedoch auch zu bedenken, dass die Publikationen randomisierter kontrollierter Studien weiterhin der Verbesserung bedürfen. Die Ergebnisse der vorliegenden Arbeit weisen acht Jahre nach Einführung des Consort-Statements ebenfalls in diese Richtung. Lediglich 15 der 72 untersuchten Publikationen waren bezüglich der untersuchten Qualitätsmerkmale im Hinblick auf den Fehler erster Art ohne jedweden Mangel. In den übrigen 57 Arbeiten fanden sich mindestens ein, in den meisten Arbeiten sogar mehrere Hinweise auf eine Inflation des Fehlers erster Art. Die vorliegende Arbeit bezieht sich auf Publikationen in einer der angesehensten medizinischen Fachzeitschriften, die aufgrund der Bekanntheit

aus einer Vielzahl von zugesandten Publikationen auswählen kann und besonders auf die Qualität der publizierten Arbeiten achtet. Es ist damit zu rechnen, dass die in dieser Arbeit gefundene Häufigkeit einer Inflation des Fehlers erster Art bezogen auf die Gesamtheit der Publikationen in verschiedensten Fachzeitschriften eher unterschätzt wird. Dies könnte anhand einer Bestandsaufnahme der Publikationen verschiedener Zeitschriften mittels des in dieser Arbeit verwendeten Extraktionsbogens überprüft werden. Eine Ausweitung der Untersuchung auf andere Fachzeitschriften hätte jedoch den Rahmen der vorliegenden Arbeit gesprengt.

7. Zusammenfassung

Ziel der vorliegenden Arbeit ist die quantitative Erfassung der Inflation des Fehlers erster Art in den randomisierten, kontrollierten Studien des Jahrgangs 2004 der Zeitschrift *The Lancet*. Insgesamt wurden 72 Publikationen identifiziert, die den Ein- und Ausschlusskriterien genügten.

Für die Auswertung wurde ein spezieller Extraktionsbogen entwickelt. Dieser wurde für jede Publikation von einer Person ausgefüllt. Von einer zweiten Person wurden stichprobenartig Kontrollen durchgeführt. Für eventuelle Abweichungen wurde ein Konsens gefunden.

Bei sechs Publikationen wurde gar kein primärer Endpunkt definiert. Diese Arbeiten können lediglich Hypothesen generieren und sind nicht dazu geeignet Behandlungsempfehlungen zu geben.

Lediglich 15 der 72 untersuchten Publikationen waren frei von Hinweisen auf eine Inflation des Fehlers erster Art. In den übrigen 57 Publikationen gibt es Hinweise darauf, dass durch die aufgefundenen methodischen Mängel eine Inflation des Fehlers erster Art und somit ein erhöhtes effektives Signifikanzniveau vorhanden ist. Dadurch steigt die Wahrscheinlichkeit für ein falsch-positives Studienergebnis zum Teil erheblich.

Nur in drei von 16 Publikationen, die mehr als einen primären Endpunkt aufweisen, wird angegeben, dass für diese Form der Mehrfachtestung korrigiert wurde. Die Autoren von 35 Publikationen geben an, eine Form der Mehrfachtestung in Bezug auf mindestens einen primären Endpunkt anzuwenden. Von diesen geben lediglich zwölf an, dass sie hierfür korrigieren.

Eine Intention-to-treat Analyse wird laut Angabe der Autoren in 58 Publikationen durchgeführt. In 31 dieser 58 Publikationen gibt es Hinweise darauf, dass dennoch nicht alle randomisierten Patienten in die Auswertung eingehen,

obwohl die Autoren angeben, eine Intention-to-treat Analyse durchgeführt zu haben.

In 18 Publikationen findet sich keine Angabe über das gewählte Signifikanzniveau.

Die vorliegende Arbeit zeigt, dass es auch acht Jahre nach Erscheinen des CONSORT-Statements (Consolidated Standards of Reporting Trials) zur Verbesserung der Berichterstattung über randomisierte Studien in den Publikationen noch immer Hinweise auf eine Inflation des Fehlers erster Art zu finden sind.

Um es dem Leser die Einschätzung der methodischen Qualität von Studien zu erleichtern, wäre es hilfreich, wenn in der Publikation explizit kenntlich gemacht würde, an welcher Stelle die im CONSORT-Statement geforderten Informationen zu finden sind.

8. Literaturverzeichnis

1. Addo-Yobo,E et al. (2004) Oral amoxicillin versus injectable penicillin for severe pneumonia in children aged 3 to 59 months: a randomised multicentre equivalency study. Lancet 9440:1141-1148

2. AHCPR (1992) Acute pain management: operative or medical procedures and trauma, Part 1. Agency for Health Care Policy and Research. Clin Pharm 4:309-331

3. Alloueche,A et al. (2004) Comparison of chlorproguanil-dapsone with sulfadoxine-pyrimethamine for the treatment of uncomplicated falciparum malaria in young African children: double-blind randomised controlled trial. Lancet 9424:1843-1848

4. Alonso,PL et al. (2004) Efficacy of the RTS,S/AS02A vaccine against Plasmodium falciparum infection and disease in young African children: randomised controlled trial. Lancet 9443:1411-1420

5. Althabe,F et al. (2004) Mandatory second opinion to reduce rates of unnecessary caesarean sections in Latin America: a cluster randomised controlled trial. Lancet 9425:1934-1940

6. Altman,DG (1996) Better reporting of randomised controlled trials: the CONSORT statement. BMJ 7057:570-571

7. Altman,DG und Bland,JM (1999) Statistics notes. Treatment allocation in controlled trials: why randomise? BMJ 7192:1209-

8. Altman,DG et al. (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 8:663-694

9. Anand,I et al. (2004) Long-term effects of darusentan on left-ventricular remodelling and clinical outcomes in the EndothelinA Receptor Antagonist Trial in Heart Failure (EARTH): randomised, double-blind, placebo-controlled trial. Lancet 9431:347-354

10. Anand,KJ et al. (2004) Effects of morphine analgesia in ventilated preterm neonates: primary outcomes from the NEOPAIN randomised trial. Lancet 9422:1673-1682

11. Andrews,DW et al. (2004) Whole brain radiation therapy with or without stereotactic radiosurgery boost for patients with one to three brain metastases: phase III results of the RTOG 9508 randomised trial. Lancet 9422:1665-1672

12. Armitage P et al. (1969) Repeated significance tests on accumulating data. Journal of the Royal Statistical Society :235-244

13. Barnes,KI et al. (2004) Efficacy of rectal artesunate compared with parenteral quinine in initial treatment of moderately severe malaria in African children and adults: a randomised study. Lancet 9421:1598-1605

14. Barwell,JR et al. (2004) Comparison of surgery and compression with compression alone in chronic venous ulceration (ESCHAR study): randomised controlled trial. Lancet 9424:1854-1859

15. Begg,C et al. (1996) Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 8:637-639

16. Brooks,WA et al. (2004) Zinc for severe pneumonia in very young children: double-blind placebo-controlled trial. Lancet 9422:1683-1688

17. Brouwer,AE et al. (2004) Combination antifungal therapies for HIV-associated cryptococcal meningitis: a randomised trial. Lancet 9423:1764-1767

18. Carr,A et al. (2004) No effect of rosiglitazone for treatment of HIV-1 lipoatrophy: randomised, double-blind, placebo-controlled trial. Lancet 9407:429-438

19. Chintu,C et al. (2004) Co-trimoxazole as prophylaxis against opportunistic infections in HIV-infected Zambian children (CHAP): a double-blind randomised placebo-controlled trial. Lancet 9448:1865-1871

20. Christ-Crain,M et al. (2004) Effect of procalcitonin-guided treatment on antibiotic use and outcome in lower respiratory tract infections: cluster-randomised, single-blinded intervention trial. Lancet 9409:600-607

21. Colhoun,HM et al. (2004) Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. Lancet 9435:685-696

22. Collins,R et al. (2004) Effects of cholesterol-lowering with simvastatin on stroke and other major vascular events in 20536 people with cerebrovascular disease or other high-risk conditions. Lancet 9411:757-767

23. Courtney,C et al. (2004) Long-term donepezil treatment in 565 patients with Alzheimer's disease (AD2000): randomised double-blind trial. Lancet 9427:2105-2115

24. Crawford,MJ et al. (2004) Screening and referral for brief intervention of alcohol-misusing patients in an emergency department: a pragmatic randomised controlled trial. Lancet 9442:1334-1339

25. Davidoff,F (2000) News from the International Committee of Medical Journal Editors. Ann Intern Med 3:229-231

26. de Kraker,J et al. (2004) Reduction of postoperative chemotherapy in children with stage I intermediate-risk and anaplastic Wilms' tumour (SIOP 93-01 trial): a randomised controlled trial. Lancet 9441:1229-1235

27. Diener,HC et al. (2004) Aspirin and clopidogrel compared with clopidogrel alone after recent ischaemic stroke or transient ischaemic attack in high-risk patients (MATCH): randomised, double-blind, placebo-controlled trial. Lancet 9431:331-337

28. Ederer,F (1998) History of clinical trials.

29. Egger,M et al. (2001) Value of flow diagrams in reports of randomized controlled trials. JAMA 15:1996-1999

30. El Arifeen,S et al. (2004) Integrated Management of Childhood Illness (IMCI) in Bangladesh: early findings from a cluster-randomised study. Lancet 9445:1595-1602

31. Emerson,PM et al. (2004) Role of flies and provision of latrines in trachoma control: cluster-randomised controlled trial. Lancet 9415:1093-1098

32. Farkouh,ME et al. (2004) Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), cardiovascular outcomes: randomised controlled trial. Lancet 9435:675-684

33. Fernandez-Aviles,F et al. (2004) Routine invasive strategy within 24 hours of thrombolysis versus ischaemia-guided conservative approach for acute myocardial infarction with ST-segment elevation (GRACIA-1): a randomised controlled trial. Lancet 9439:1045-1053

34. Filippi,M et al. (2004) Interferon beta-1a for brain tissue loss in patients at presentation with syndromes suggestive of multiple sclerosis: a randomised, double-blind, placebo-controlled trial. Lancet 9444:1489-1496

35. Fisher,B et al. (2004) Treatment of lymph-node-negative, oestrogen-receptor-positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project randomised clinical trials. Lancet 9437:858-868

36. Fisher, R A 1925 Statistical methods for research workers. London: Oliver and Boyd

37. Fletcher,AE et al. (2004) Population-based multidimensional assessment of older people in UK general practice: a cluster-randomised factorial trial. Lancet 9446:1667-1677

38. Freemantle,N et al. (1997) CONSORT: an important step toward evidence-based health care. Consolidated Standards of Reporting Trials. Ann Intern Med 1:81-83

39. Gail,MH (1996) Statistics in action. Journal of the American Statistical Association :1-13

40. Gale,EA et al. (2004) European Nicotinamide Diabetes Intervention Trial (ENDIT): a randomised controlled trial of intervention before the onset of type 1 diabetes. Lancet 9413:925-931

41. Gilliland,FD et al. (2004) Effect of glutathione-S-transferase M1 and P1 genotypes on xenobiotic enhancement of allergic responses: randomised, placebo-controlled crossover study. Lancet 9403:119-125

42. Gournay,V et al. (2004) Prophylactic ibuprofen versus placebo in very premature infants: a randomised, double-blind, placebo-controlled trial. Lancet 9449:1939-1944

43. Greenhalgh,RM et al. (2004) Comparison of endovascular aneurysm repair with open repair in patients with abdominal aortic aneurysm (EVAR trial 1), 30-day operative mortality results: randomised controlled trial. Lancet 9437:843-848

44. Grigor,C et al. (2004) Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single-blind randomised controlled trial. Lancet 9430:263-269

45. Halliday,A et al. (2004) Prevention of disabling and fatal strokes by successful carotid endarterectomy in patients without recent neurological symptoms: randomised controlled trial. Lancet 9420:1491-1502

46. Harper,DM et al. (2004) Efficacy of a bivalent L1 virus-like particle vaccine in prevention of infection with human papillomavirus types 16 and 18 in young women: a randomised controlled trial. Lancet 9447:1757-1765

47. Harrison,TW et al. (2004) Doubling the dose of inhaled corticosteroid to prevent asthma exacerbations: randomised controlled trial. Lancet 9405:271-275

48. Hill ,AB (1951) The clinical trial. Br Med Bull 4:278-282

49. Hollis,S und Campbell,F (1999) What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 7211:670-674

50. Holmberg,L und Anderson,H (2004) HABITS (hormonal replacement therapy after breast cancer--is it safe?), a randomised comparison: trial stopped. Lancet 9407:453-455

51. Hommes,OR et al. (2004) Intravenous immunoglobulin in secondary progressive multiple sclerosis: randomised placebo-controlled trial. Lancet 9440:1149-1156

52. Homs,MY et al. (2004) Single-dose brachytherapy versus metal stent placement for the palliation of dysphagia from oesophageal cancer: multicentre randomised trial. Lancet 9444:1497-1504

53. Huston,P und Hoey,J (1996) CMAJ endorses the CONSORT statement. Consolidation of Standards for Reporting Trials. CMAJ 9:1277-1282

54. Israel,E et al. (2004) Use of regularly scheduled albuterol treatment in asthma: genotype-stratified, randomised, placebo-controlled cross-over trial. Lancet 9444:1505-1512

55. Jindani,A et al. (2004) Two 8-month regimens of chemotherapy for treatment of newly diagnosed pulmonary tuberculosis: international multicentre randomised trial. Lancet 9441:1244-1251

56. Jocham,D et al. (2004) Adjuvant autologous renal tumour cell vaccine and risk of tumour progression in patients with renal-cell carcinoma after radical nephrectomy: phase III, randomised controlled trial. Lancet 9409:594-599

57. Julius,S et al. (2004) Outcomes in hypertensive patients at high cardiovascular risk treated with regimens based on valsartan or amlodipine: the VALUE randomised trial. Lancet 9426:2022-2031

58. Kelly,D et al. (2004) Tacrolimus and steroids versus ciclosporin microemulsion, steroids, and azathioprine in children undergoing liver transplantation: randomised European multicentre trial. Lancet 9439:1054-1061

59. Klareskog,L et al. (2004) Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. Lancet 9410:675-681

60. Koblin,B et al. (2004) Effects of a behavioural intervention to reduce acquisition of HIV infection among men who have sex with men: the EXPLORE randomised controlled study. Lancet 9428:41-50

61. Leung,KL et al. (2004) Laparoscopic resection of rectosigmoid carcinoma: prospective randomised trial. Lancet 9416:1187-1192

62. Lux,AL et al. (2004) The United Kingdom Infantile Spasms Study comparing vigabatrin with prednisolone or tetracosactide at 14 days: a multicentre, randomised controlled trial. Lancet 9447:1773-1778

63. Manandhar,DS et al. (2004) Effect of a participatory intervention with women's groups on birth outcomes in Nepal: cluster-randomised controlled trial. Lancet 9438:970-979

64. McCarey,DW et al. (2004) Trial of Atorvastatin in Rheumatoid Arthritis (TARA): double-blind, randomised placebo-controlled trial. Lancet 9426:2015-2021

65. Meinert,CL (1998) Beyond CONSORT: need for improved reporting standards for clinical trials. Consolidated Standards of Reporting Trials. JAMA 18:1487-1489

66. Moher,D et al. (2001) Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 15:1992-1995

67. Moher,D et al. (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 9263:1191-1194

68. Morris,SS et al. (2004) Monetary incentives in primary health care and effects on use and coverage of preventive health care interventions in rural Honduras: cluster randomised trial. Lancet 9450:2030-2037

69. Muir,KW et al. (2004) Magnesium for acute stroke (Intravenous Magnesium Efficacy in Stroke trial): randomised controlled trial. Lancet 9407:439-445

70. Myles,PS et al. (2004) Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. Lancet 9423:1757-1763

71. Newnham,JP et al. (2004) Effects of repeated prenatal ultrasound examinations on childhood outcome up to 8 years of age: follow-up of a randomised controlled trial. Lancet 9450:2038-2044

72. O'Brien,PC und Fleming,TR (1979) A multiple testing procedure for clinical trials. Biometrics 3:549-556

73. Pocock,SJ (1982) Interim analyses for randomized clinical trials: the group sequential approach. Biometrics 1:153-162

74. Pocock,SJ et al. (1987) Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 7:426-432

75. Poole-Wilson,PA et al. (2004) Effect of long-acting nifedipine on mortality and cardiovascular morbidity in patients with stable angina requiring treatment (ACTION trial): randomised controlled trial. Lancet 9437:849-857

76. Remuzzi,G et al. (2004) Mycophenolate mofetil versus azathioprine for prevention of acute rejection in renal transplantation (MYSS): a randomised trial. Lancet 9433:503-512

77. Roberts,I et al. (2004) Effect of intravenous corticosteroids on death within 14 days in 10008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. Lancet 9442:1321-1328

78. Sackett,DL et al. (1996) Evidence based medicine: what it is and what it isn't. BMJ 7023:71-72

79. Sakkers,R et al. (2004) Skeletal effects and functional outcome with olpadronate in children with osteogenesis imperfecta: a 2-year randomised placebo-controlled study. Lancet 9419:1427-1431

80. Schnitzer,TJ et al. (2004) Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), reduction in ulcer complications: randomised controlled trial. Lancet 9435:665-674

81. Schulz,KF (1997) The quest for unbiased research: randomized clinical trials and the CONSORT reporting guidelines. Ann Neurol 5:569-573

82. Schumacher, M. and Schulgen, G. 2007 Methodik klinischer Studien : methodische Grundlagen der Planung, Durchführung und Auswertung. Berlin ; Heidelberg ; New York, NY: Springer

83. Singhal,A et al. (2004) Breastmilk feeding and lipoprotein profile in adolescents born preterm: follow-up of a prospective randomised study. Lancet 9421:1571-1578

84. Smeets,NW et al. (2004) Surgical excision vs Mohs' micrographic surgery for basal-cell carcinoma of the face: randomised controlled trial. Lancet 9447:1766-1772

85. Staedke,SG et al. (2004) Combination treatments for uncomplicated falciparum malaria in Kampala, Uganda: randomised clinical trial. Lancet 9449:1950-1957

86. Stephenson,JM et al. (2004) Pupil-led sex education in England (RIPPLE study): cluster-randomised intervention trial. Lancet 9431:338-346

87. Sutherland,I (1998) Medical Research Council Streptomycin trial.

88. Thornton,JG et al. (2004) Infant wellbeing at 2 years of age in the Growth Restriction Intervention Trial (GRIT): multicentred randomised controlled trial. Lancet 9433:513-520

89. To,MS et al. (2004) Cervical cerclage for prevention of preterm delivery in women with short cervix: randomised controlled trial. Lancet 9424:1849-1853

90. Tran,TH et al. (2004) Dihydroartemisinin-piperaquine against multidrug-resistant Plasmodium falciparum malaria in Vietnam: randomised clinical trial. Lancet 9402:18-22

91. Turnbull,DA et al. (2004) Clinical, psychosocial, and economic effects of antenatal day care for three medical complications of pregnancy: a randomised controlled trial of 395 women. Lancet 9415:1104-1109

92. Vain,NE et al. (2004) Oropharyngeal and nasopharyngeal suctioning of meconium-stained neonates before delivery of their shoulders: multicentre, randomised controlled trial. Lancet 9434:597-602

93. van Koningsveld,R et al. (2004) Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for Guillain-Barre syndrome: randomised trial. Lancet 9404:192-196

94. van Leth,F et al. (2004) Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN Study. Lancet 9417:1253-1263

95. Van Overmeire,B et al. (2004) Prophylactic ibuprofen in premature infants: a multicentre, randomised, double-blind, placebo-controlled trial. Lancet 9449:1945-1949

96. Verweij,J et al. (2004) Progression-free survival in gastrointestinal stromal tumours with high-dose imatinib: randomised trial. Lancet 9440:1127-1134

97. Weatherall,DJ (1994) The inhumanity of medicine. BMJ 6970:1671-1672

98. Wollert,KC et al. (2004) Intracoronary autologous bone-marrow cell transfer after myocardial infarction: the BOOST randomised controlled clinical trial. Lancet 9429:141-148

9. Danksagung

In den Zeiten der Entstehung dieser Arbeit haben mich viele Menschen Unterstützt…

Ihnen allen meinen Dank von Herzen!

Insbesondere bedanke ich mich bei meinem Doktorvater und Betreuer Herrn Prof. Dr. H.-P. Beck-Bornholdt. Ich bin froh darüber, wie sich alles in allem zwischendurch ver- und dann wieder ent-wickelt hat! Das war eine wichtige Erfahrung für mich auf dem Weg zur Arztwerdung.

Ich danke dem Leben, dass es mich gibt!

10. Lebenslauf

Name:	Siemann, Benjamin
Geburtsdatum:	12.05.1981
Geburtsort:	Henstedt-Ulzburg
Familienstand:	ledig
Staatsangehörigkeit:	deutsch
Schulbildung:	
1987 – 1991	Grundschule Glückstadt
1991 – 2000	Detlefsengymnasium Glückstadt
Schulabschluss:	2000 Allgemeine Hochschulreife
Zivildienst:	
2000 – 2001	Helfer im Pflegedienst / Patienten-transport im Klinikum Lippe/Lemgo
Hochschulstudium:	
2001-2007	Studium der Humanmedizin an der Universitätsklinik Hamburg-Eppendorf der Universität Hamburg
2003	Erster Abschnitt der ärztlichen Prüfung
2007	Zweiter Abschnitt der ärztlichen Prüfung
Approbation:	November 2007 in Hamburg
Berufliche Tätigkeit:	seit Dezember 2007 Assistenzarzt in der Medizinisch-Psychosomatischen Klinik Bad Bramstedt
Sprachen:	Deutsch, Englisch

11. Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als von mir angegebene Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich und inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe.
Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig zur Promotion beworben habe.

Benjamin Siemann