Zu Geschichte und Theorie der therapeutisch-klinischen Forschung bei chronischen Krankheiten am Beispiel der Chemotherapie der chronischen Lungentuberkulose

History and Theory of Therapeutic Clinical Research in Chronic Diseases, Taking as Example the Chemotherapy of Chronic Pulmonary Tuberculosis

Autoren K. Bartmann †, R. Kropp

Institut Das Deutsche Tuberkulose-Archiv, Fulda

eingereicht 14.8.2008
akzeptiert 8.9.2008

Bibliografie
DOI 10.1055/s-2008-1038272
Online-Publikation: 6.11.2008
Pneumologie 2009; 63: 31–40

ISSN 0934-8387

Korrespondenzadresse
Dr. med. Robert Kropp
Bahnhofstr. 4
36037 Fulda
dr.robert.kropp@gmx.de

Zusammenfassung

Die historischen und gegenwärtigen Probleme therapeutisch-klinischer Prüfungen bei chronischen Krankheiten werden am Beispiel der Chemotherapie der chronischen Lungentuberkulose beschrieben und diskutiert.

Abstract

Historical and present problems of therapeutic clinical trials in chronic diseases are outlined and analysed, taking the chemotherapy of chronic pulmonary tuberculosis as an example.

Grundlagen

Ein Ziel jeder wissenschaftlichen Bemühung ist es, zu Ergebnissen zu gelangen, die keine problemspezifischen ergebnisrelevanten Zweifel mehr aufkommen lassen und daher Allgemeingut der Forschergemeinschaft werden können. In der therapeutisch-klinischen Forschung werden kausale Fragen bearbeitet: Ist die untersuchte Prüfsubstanz Ursache (*kausale Bedingung*) der nach ihrer Anwendung zu beobachtenden Erscheinungen? An die Antwort ist die wissenschaftliche Forderung zu stellen, dass sie soweit wie möglich nicht mehr angezweifelt werden kann, dass sie nach heutiger Terminologie Evidenz-basiert ist, im Interesse der Patienten und nicht nur zur Befriedigung wissenschaftlicher Desiderata.

Jede kausal bedingte Beziehung manifestiert sich als Veränderung eines bestehenden Zustandes oder Prozesses. Die Änderung wird durch einen Vergleich erkannt. Ein Vergleich ist nur verlässlich, wenn die zu vergleichenden Gegenstände gleich sind und die außer der Ursache noch mitwirkenden Einflussfaktoren ausreichend eliminiert oder neutralisiert sind. Zu vergleichende Kollektive müssen möglichst „homogen" sein. Unter „*Homogenität*" von verschiedenen Kollektiven wird ihre qualitative und quantitative Gleichheit in Bezug auf alle relevanten, bekannten, bekannten, aber nicht erfassten, und nicht bekannten Einflussgrößen verstanden. Die Gleichheit bezieht sich auf die Häufigkeitsverteilung von Eigenschaften zwischen Gruppen (ihre „Struktur"; *interkollektive Homogenität*) und/oder auf die Gleichheit der Einheiten in einer Gruppe (*intrakollektive Homogenität*). Was „Gleichheit" genau ausdrücken soll, wird später erörtert. Unter natürlichen Verhältnissen besteht keine Homogenität. Sie muss für die Zwecke des Vergleichs geschaffen werden. Das gelingt stets nur mehr oder weniger gut. Für das Kriterium „gut genug" liefert uns die Statistik bisher nur ungenügende Definitionen.

Verbesserungen der Homogenität, gleichbedeutend mit Verringerungen der *Heterogenität*, lassen sich auf verschiedene Weise erreichen. Zu unterscheiden sind prospektive Homogenisierung vor Prüfungsbeginn und retrospektive Homogenisierung nach Ende der Behandlung. Die Methoden werden weiter unten genauer erörtert und hier nur zur Übersicht skizziert. Prospektiv wird fast stets die Homogenisierung durch die Wahl geeigneter Zulassungskriterien angewandt, welche eine Aufnahme in die Prüfung vom Vorliegen bestimmter Eigenschaften wie Alter, Krankheitsform oder -Dauer abhängig machen. Weitere Methoden sind die Bildung in sich möglichst homogener Untergruppen („Schichten", „Strata"), wobei die Schichtungsvariable als Kovariable in die Analyse mit einzubeziehen ist, bzw. die Bildung möglichst homogener Paare, deren Mitglieder 2 verschiedenen Gruppen zugewiesen werden. Mit allen genannten Verfahren werden nur die ins Auge genommenen Einflussfaktoren berücksichtigt, nicht andere bekannte und die nicht bekannten Faktoren. Dies ist nur

mit zufälliger Zuteilung der Patienten zu den Gruppen („*Randomisierung*“) möglich. Neutralisieren lassen sich im Patienten über längere Zeit stabil vorliegende Einflussgrößen beim Individualvergleich, in welchem am Einzelfall der Trend des Krankheitsverlaufs vor und während der Prüfung miteinander verglichen wird. – Retrospektive Homogenisierung ist möglich mithilfe nachträglicher Paarbildung und Schichtung.

Alle Studienpläne, in denen prospektiv eine Prüfung-interne Vergleichsgruppe mitgeführt wird, und der Prüfung-interne Individualvergleich sind als kontrollierte Studien zu bezeichnen. Zu Zweifeln am Individualvergleich wird weiter unten Stellung genommen. Alle Studien mit Prüfung-externem, so genanntem historischen Vergleich werden nicht als kontrolliert klassifiziert. Die amtliche Neuzulassung von Medikamenten verlangt heute in den meisten Staaten kontrollierte Studien mit Ausnahme von schweren Krankheiten, die bisher überhaupt nicht therapierbar waren wie seinerzeit die tuberkulöse Meningitis.

Außer unzureichender Vergleichbarkeit (Strukturgleichheit) gibt es noch weitere nichtzufällige Störungen, die alle „systematische Fehler“, auch „*bias*“ oder „Verzerrungen“ genannt werden. Die wichtigsten Verzerrungen betreffen, abgesehen vom schon erwähnten Design einer Studie, die Wahl der Kontrolltherapie; die Form der Zuteilung auf die verschiedenen Gruppen beim Kollektivvergleich; die Durchführung; die Ermittlung der Ergebnisse; deren Auswertung, Interpretation und Bericht.

Außer den systematischen Verzerrungen gibt es die der natürlichen zufälligen Variation. Diese lässt sich nicht beseitigen. Man kann sie nur durch intrakollektive Homogenisierung und große Fallzahlen reduzieren. Der Einfluss der dann noch verbleibenden Variation muss durch geeignete statistische Verfahren ermittelt und beim therapeutischen Urteil in Rechnung gestellt werden. Eine befriedigende Angleichung der Verteilungshäufigkeit von Eigenschaften durch Steigerung der Fallzahlen wird nach dem so genannten Gesetz der großen Zahlen[1] erst mit einer hohen Patientenzahl je Gruppe oder Stratum erreicht. Derartige Großversuche sind jedoch nur selten realisierbar. Es sind in letzter Zeit Versuche unternommen worden, die mit kleineren Fallzahlen trotz Randomisierung auftretenden störenden Ungleichheiten durch „eingeschränkte Randomisierung“ zu reduzieren. Dazu wird im Text Stellung genommen. – Die systematischen Verzerrungen drücken sich in der Lage von Durchschnittswerten aus, die zufällige Variation in deren Streubereich.

Anwendung der Grundlagen

Die Anfänge kontrollierter Studien

Klinische chemotherapeutische Untersuchungen werden bei der Lungentuberkulose erstmals mit Einführung der Goldtherapie 1917 relevant, über die im Lauf von etwa 20 Jahren weltweit mehr als 500 überwiegend positiv eingestellte Arbeiten publiziert wurden [1]. Die Ergebnisse widersprachen einander, ein klares Bild war nicht zu gewinnen [1,2]. Die erste Studie, in der versucht wurde, die bisherigen methodischen Unzulänglichkeiten systematisch zu vermeiden, wurde 1931 von Amberson u. Mitarb. [2] publiziert. Sie wurde bereits als kontrollierte Studie mit zufälliger Zuteilung und einem detaillierten Untersuchungsprotokoll angelegt. Aus Patienten mit einem guten Allgemeinzustand und Formen der Tuberkulose, die nach Literaturangaben gut auf die Goldtherapie ansprechen sollten, wurden 2 Gruppen aus einander möglichst ähnlichen Paaren zusammengestellt, von denen ein Mitglied einer Gruppe A, das andere einer Gruppe B zugeteilt wurde. Jede Gruppe bestand aus 12 Patienten. Dann wurde durch Münzwurf entschieden, welches der beiden Kollektive die Prüf- bzw. die Kontrollgruppe bildete. Zufällig war also nur die Zuteilung der Gruppen zu den Therapieformen (= Clusterrandomisierung), nicht die Zuteilung der zugelassenen Patienten zu einer Gruppe. Die Prüfgruppe erhielt das Goldpräparat, der Kontrollgruppe wurde das Lösungsmittel des Goldpräparates injiziert. Alle übrigen Bedingungen waren für beide Gruppen gleich. Die Behandlung wurde nach einer Vorbeobachtung von 30 Tagen begonnen und 8 Wochen fortgesetzt. Die Goldtherapie erwies sich als unwirksam. In der Diskussion wurde allgemein auf die methodischen Erfordernisse therapeutisch-klinischer Prüfungen eingegangen. Hervorgehoben wurden folgende Möglichkeiten der Verzerrung: die natürliche Fluktuation im Krankheitsverlauf; Fehlen eines über die Routine hinausgehenden Untersuchungsprogramms; eine zu ungenaue und zu laxe Art der Beobachtung; keine ausreichende Vergleichbarkeit der Gruppen wegen mangelnder Berücksichtigung des Charakters der Läsionen; Nichtberücksichtigung der vorzeitigen Ausscheider; keine Beschreibung der angewandten bakteriologischen Technik; Fehlinterpretation von Röntgenbefunden und Auskultation.

Martinis grundlegende Arbeiten

Die nächste Publikation, die sich mit der Methodik der klinischen Prüfung einer medikamentösen Therapie der Lungentuberkulose auseinandersetzt, ist unseres Wissens die von Martini aus dem Jahre 1934 [3]. Paul Martini (1889 – 1964), Ordinarius für Innere Medizin in Bonn, hat im deutschen Sprachraum das Bewusstsein der Ärzte dafür geweckt, dass die Ermittlung des Wertes therapeutischer Maßnahmen nicht einfach auf klinischen Eindrücken und Intuition basieren darf, wenn sie verbindliche Geltung beanspruchen will. Über 30 Jahre, bis zu seinem Tod, hat sich Martini für die Aufstellung und Beachtung von Regeln der Prüfung eingesetzt, die ein möglichst verzerrungsfreies und statistisch fundiertes therapeutisches Urteil erlauben sollen. Er war zuerst ein Rufer in der Wüste, wurde aber dank seiner beharrlichen, unerschrockenen, sachlich fundierten und formulierten Kritik zu einer maßgebenden Instanz. So sind auch seine Richtlinien zur therapeutischen Prüfung bei Lungentuberkulose auf Wunsch von Institutionen verfasst worden und in seiner Methodenlehre ab 1947 wiedergegeben [3 – 7]. Mit seinem Doktoranden Rosendahl veröffentlichte er eine umfassende vernichtende Kritik der Goldtherapie bei Tuberkulose [1]. Er hat nur eine einzige eigene therapeutische Prüfung bei Tuberkulose publiziert, zusammen mit seinen Mitarbeitern. Sie betrifft das Thiosemicarbazon (TSC) Conteben® [5]. Allen genannten Veröffentlichungen Martinis liegt seit 1934 eine einheitliche generelle Einstellung zur Problematik zugrunde. Die Zitate können daher als pars pro toto gelten.

Martini weist auf eine ganze Reihe von Möglichkeiten der Verzerrung hin, die sich mit denen von Amberson u. Mitarb. weitgehend decken. Im Gegensatz zu den amerikanischen Autoren hält er aber den Individualvergleich im Prinzip für eine bessere Methode als den Kollektivvergleich [3,7, S. 32]. Der Individualvergleich verlangt jedoch nach Martini bei der Tuberkulose eine Vorbeobachtungsperiode von 2 – 3 Monaten zum Ausschluss von in der Testperiode auftretenden „spontanen“ Besserungen,

[1] Das Gesetz der großen Zahlen besagt, dass sich ein Mittelwert mit wachsendem Stichprobenumfang dem Erwartungswert (= „wahrer“, unbekannter Mittelwert der Grundgesamtheit) nähert.

die fälschlich als Therapieeffekt gedeutet werden. Die frischen, noch überwiegend akuten Stadien der Lungentuberkulose laufen jedoch zu schnell ab, als dass mit der Prüfmedikation 2 – 3 Monate gewartet werden könnte. Infolgedessen kommt für diese Formen nur der Kollektivvergleich mit zufälliger Zuteilung der Patienten zu den Kollektiven infrage. Dabei müssen sämtliche Patienten untereinander in Bezug auf die zahlreichen den natürlichen Krankheitsverlauf beeinflussenden Faktoren weitgehend gleich (homogen) sein ([6], S. 163). Dazu wird meist die Bildung von Untergruppen (Strata) erforderlich sein. – Der Individualvergleich ist also nur bei chronischer Tuberkulose durchführbar, und zwar in Perioden ohne Tendenz zur Besserung. Der Individualvergleich kann auch nicht zum sukzedanen Vergleich zweier Antituberkulotika dienen, die beide eine gewisse Wirkung haben. Denn es muss entgegen der Auffassung von Martini angenommen werden, dass dann eine Testperiode von einigen Wochen doch schon die Prüfbedingungen für die zweite später gegebene Testsubstanz irreversibel verändert. Dafür sprechen die vielen Berichte über eine schon wenige Tage nach Behandlungsbeginn nachweisbar werdende Besserung selbst durch schwach wirkende Substanzen wie TSC oder PAS (Paraaminosalizylsäure). Sie lässt sich beobachten bei Haut- und Schleimhauttuberkulosen (z. B. [8 – 11]) und ist bei pulmonalen Prozessen an der Bakterienausscheidung auch messbar (z. B. [12]). Außerdem besteht bei 2 hinter einander geschalteten Monotherapien für 2 – 3 Monate das Risiko einer Resistenzentwicklung der Bakterien gegen beide Medikamente. Der individuelle Vergleich kann also nur die Frage beantworten: Ist die Behandlung mit der Prüfsubstanz P wirksamer als die gleiche Behandlung ohne die Substanz P, bei im Übrigen gleich gebliebenen Umständen? Trotz dieser Einschränkung könnte der Individualvergleich eine wichtige Rolle in der Phase der ersten Wirksamkeitsprüfung eines neuen Therapiekandidaten spielen. Es lohnt sich deshalb, diese Methodik zu analysieren, zumal man aus ethischen Gründen Patienten mit frischer und daher schneller heilender Tuberkulose die Chance einer Dauerheilung durch die bereits zur Verfügung stehenden Mittel nicht nehmen darf. Damit ist der ethische Aspekt von Arzneimittelprüfungen angedeutet. Er hat nach unser aller Grundverständnis der Arzt-Patient-Beziehung Vorrang vor allen methodologischen Forderungen und wirkt sich auf die Studienplanung begrenzend aus.

Für den Individualvergleich führt Martini vor allem 2 Argumente ins Feld. Das erste behauptet, dass Kollektivvergleiche bei chronischer Tuberkulose unmöglich seien; das zweite will begründen, dass die Ergebnisse von Individualvergleichen besonders verlässlich seien.

Martinis 1. These, mit kritischen Bemerkungen

Die erste These besagte damals, weil bei Chronikern Kollektivvergleiche und Paarbildung unmöglich seien, es gäbe keine Alternative zum Individualvergleich bei ihnen, ein Schluss ex negativo ad positivum. Bei chronischer Lungentuberkulose sei jeder Fall, bedingt durch die im Laufe der Zeit eingetretenen verschiedenen Abwandlungen der Krankheit, zu einem Individuum geworden, so dass ein Kranker nur in sich selbst und mit sich selbst verglichen werden könne [13]. Am konkretesten sind späte Ausführungen Martinis. Danach hat ein statistisch signifikanter Unterschied nur dann einen realen Beweiswert, wenn die zu vergleichenden Gruppen homogen sind ([7], S. 223). Jedoch: „Beim chronisch Kranken muss damit gerechnet werden, dass die Angriffsmöglichkeiten eines zu prüfenden Mittels bei einem mehr oder minder großen Teil der Patienten völlig andere sind als bei (zu ergänzen: den übrigen Fällen mit) der gleichen ‚Diagnose'" ([7], S. 26). Hier ist also klar Homogenität auf gleiches Einwirkungsvermögen der Prüfsubstanz bezogen und deren Vorhandensein infrage gestellt. Dieses Problem ist aber nicht theoretisch, sondern nur empirisch zu lösen. Martini sieht zu sehr auf die anatomischen und physiologischen Unterschiede, die bei den einzelnen chronisch Kranken die Pharmakokinetik und die Wirkungsbedingungen der Prüfsubstanz in den Herden beeinflussen. Er berücksichtigt zu wenig, dass diese Faktoren in Relation zu Wirkungstyp und Wirkungsintensität der Substanz gesehen werden müssen. Hohe Wirksamkeit kann viele Unterschiede in den Wirkungsbedingungen überspielen. Wie weit das zutrifft, kann nur nach Abschluss der Prüfperiode geklärt werden. Inhomogenes Reagieren auf die Prüfsubstanz wird erst bei Ende der Prüfung erkennbar. Das ist anzunehmen, wenn Versager bzw. besonders gut ansprechende Patienten bestimmte gemeinsame unterscheidende Merkmale aufweisen, welche das abweichende Verhalten erklären können und dafür prognostisch geeignet sind, z. B. [15]. – Falls Martinis These der Unmöglichkeit von Kollektivvergleichen zutrifft, dann dürften mehrere voneinander unabhängige Kollektivvergleiche mit gleicher Fragestellung und vergleichbarer Methodik nicht zu gleichen Ergebnissen führen. 1958 hat Bartmann durch eine statistisch gestützte Analyse (heute: „Metaanalyse") wahrscheinlich gemacht, dass die gleichartigen Ergebnisse mehrerer Studien beim Vergleich von Isoniazid (INH) allein mit INH in Kombination nicht zufallsbedingt sind [16]. Bis 1985 hatte die antituberkulöse Chemotherapie einen Wirksamkeitsgrad erreicht, der bei 98 – 100 % unvorbehandelter Kranker zur klinischen Heilung führte und mit Rückfällen von < 2,5 % nach 2 Jahren einherging [17]. Unterschiede zwischen den Kranken hinsichtlich der Wirkungsmöglichkeit von Chemotherapeutika spielten keine Rolle mehr. Martinis Argument ex negativo war für die Tuberkulose hinfällig geworden. Es ist auch hinfällig aus einem anderen Grund, den schon Hill angeführt hat [18]. Wenn keine Vergleichsgruppen wegen der Individualität der Krankheitsfälle gebildet werden können, dann ist auch kein historischer Vergleich möglich und beim Individualvergleich keine Einzel-Prognose der Therapiechancen, weil diese Vergleiche die Bildung von Gruppen zur Erfassung von Typen voraussetzen.

Martinis 2. These

Martinis 2. These besagt, dass bei der von ihm vorgeschlagenen Methode des Individualvergleichs schon jeder Einzelfall einen erheblichen Beweisgrad in sich trägt. Er nähere sich der exakten Induktion (damit wird meist eine Induktion auf der Grundlage von Experimenten gemeint), im Gegensatz zur generalisierenden Induktion einer allgemeinen Materialsammlung ([6], S. 99). In welchen Merkmalen drückt sich die Beweiskraft aus? „Je rascher, unerwarteter, günstiger die Abweichung des Verlaufs nach Einsatz der zu prüfenden Therapie in einem Einzelfall zum Ausdruck kommt, umso beweiskräftiger ist dieser Einzelfall für sich allein" ([7], S. 236). Was unter „rascher", „unerwarteter", „günstiger" in diesem Zusammenhang genau verstanden werden soll, ist nicht definiert oder an Beispielen gezeigt. Die „Beweiskraft" ist unter diesen Umständen kein Begriff, der intersubjektiv zuverlässig gebraucht werden kann. Solange das unmöglich ist, kann er nur als Ausdruck einer subjektiven Überzeugung für die Nichtzufälligkeit der betreffenden klinischen Beobachtung angesehen werden. – In der Diskussion dieser Fragen ist auch behauptet worden, dass schon bei Vorliegen eines einzelnen Falles von chronischer Lungentuberkulose, bei welchem die

Kriterien für einen Erfolg der Prüftherapie besonders stark ausgeprägt sind, ein positives Urteil über die Prüfsubstanz abgegeben werden könne [19,20], worauf Berg prompt ein Gegenbeispiel lieferte [21]. Martinis These von der Beweiskraft des Einzelfalls ist von ihm selbst relativiert worden ([6], S. 99): „… es haftet dem einzelnen Fall doch immer noch soviel Zufälliges und Unkontrollierbares an, dass wir auch hier zu einem ausreichenden Beweis immer eine Reihe von klinischen Beobachtungen brauchen". Also ist der Einzelfall kein zuverlässiger Beweis. So subjektiv eindrucksvoll ein Einzelfall bei der Individualanalyse auch sein mag, er ist, wenn es sich nicht um eine Krankheitsform handelt, die bis dahin unbeeinflussbar war, nur *ein* Element für das therapeutische Urteil wie jeder andere Proband der Studie bzw. der betreffenden Prüfgruppe beim kollektiven Vergleich.–
In den letzten Jahren ist von übernational wirkenden Behörden, die für die Aufstellung von Leitlinien zur Wirksamkeitsprüfung von Arzneimitteln zuständig sind (ICH, EMEA), die Klassifizierung des Individualvergleichs als kontrollierte Prüfung infrage gestellt worden. In EMEA [22] heißt es: „In so-called base-line controlled studies the patients' state over time is compared with their baseline state. Although these studies are sometimes thought to use „the patient as his own control" they do not have in fact an internal control. Rather, changes from baseline are compared with an estimate what would have happened to the patient in the absence of treatment with the test drug … Such estimates are generally made on the basis of general knowledge without reference to a specific control population … Designers and analysts of such trials need to be aware to justify its use". Dass bei der Individualanalyse der Effekt beim behandelten Patienten mit seiner Prognose verglichen wird, ist von den Anhängern der Methode selbst klar ausgesprochen worden [5,23]. Auf welch schwachen Füßen sie bezüglich der Intersubjektivität steht, ergibt sich aus Bemerkungen Martinis. Die Verlässlichkeit der Prognose hängt von einer ausreichenden Vorbeobachtung ab. Diese muss solange durchgeführt werden, dass „man sich mit Wahrscheinlichkeit vor dem Auftreten spontaner Veränderungen gesichert fühlen kann" ([6], S. 160; [7], S. 224/5). Und es gilt, „dass eine komplexe Prognosestellung überhaupt und erst recht nur sehr erfahrenen Fachleuten erlaubt ist" ([6], S. 163). Die Prognose ist also das Produkt eines sich sicher fühlenden sehr erfahrenen Spezialisten. Das ist natürlich ein recht unsicherer Punkt für eine Methodik, die zu intersubjektiv verbindlichen Ergebnissen führen soll. Hinzu kommt, dass keine noch so lange Vorbeobachtung vor dem Auftreten spontaner Schwankungen schützen kann. Denn diese Schwankungen sind nach Martini unvorhersehbar und unerklärbar ([3], S. 88), erfüllen also die Kriterien der Zufälligkeit. Das bedeutet, dass man im Einzelfall nicht vorhersehen kann, wann ein solches Ereignis eintritt, und, da es zudem selten sein soll ([5], S. 555), ob es überhaupt eintritt. Und selbst wenn das der Fall sein sollte, ist es bei einem genügenden Umfang des Kollektivs für die Beurteilung des Ergebnisses meist praktisch irrelevant.
Trotz alledem, der Individualvergleich ist damit nicht verloren. Denn die Prognose ist gar nicht notwendig. Man muss nur die prospektive Sichtweise aufgeben, die der Vergleich mit der Prognose erfordert, und stattdessen nach Abschluss der Therapie fragen: Gab es nach Beginn der Prüftherapie eine Trendänderung im Verlauf bei im Übrigen gleich gebliebenen Umständen? Diese Frage ist empirischer und nicht hypothetischer Natur und kann bei geeigneter Studienplanung mit den Methoden der Trendstatistik (Regressionsanalysen) – und im Erfolgsfall und bei genügend häufig durchgeführten Untersuchungen mit statistischer Signifikanz sogar für den einzelnen Patienten – beantwortet werden. Zur Klärung der Verlässlichkeit und Repräsentativität solcher Einzelergebnisse müssen diese natürlich zusammengefasst werden, um für die Gesamtheit oder ihre Untergruppen den Vertrauensbereich (Konfidenzintervall) zu ermitteln und daraus die entsprechenden Schlüsse zu ziehen. Der Individualvergleich neutralisiert die individuell permanent vorhandenen Einflussfaktoren wie Alter, Geschlecht, bleibende Schäden, aber nicht die erst im Laufe der Prüfung auftretenden individuellen und kollektiven Störungen. Letztere lassen sich nur durch Bildung paralleler Vergleichsgruppen auffangen. Der Individualvergleich ist aber noch aus einem anderen, viel allgemeineren Grund nicht verloren: Er ist unentbehrlich. Jeder Kollektivvergleich ist zwangsläufig zunächst ein Individualvergleich bei jedem Mitglied des Kollektivs. Die Bewertung der Kollektive ergibt sich nur aus der Zusammenfassung der Ergebnisse der Individualvergleiche. Wir sehen: Es gibt zwischen Individualvergleich und Kollektivvergleich keinen Unterschied in der Sache. Beide benötigen den anderen. Der Unterschied liegt lediglich in der Gewichtung der beiden Vorgehensweisen[2].

Kontrollierte Studien in Großbritannien und USA

Eine weitere richtungweisende Veröffentlichung aus der Anfangszeit der antituberkulösen Chemotherapie ist eine Empfehlung von Hinshaw und Feldman aus dem Jahre 1945 [24]. Sie diente der Vermeidung methodischer Unzulänglichkeiten, welche die amerikanischen Veröffentlichungen über die klinische Wirksamkeit verschiedener Sulfone unverwertbar machten. Einige Monate später begann die klinische Prüfung von Streptomycin (SM), an der Hinshaw und Feldman maßgeblich beteiligt waren [25]. Ihre Empfehlungen sind darin nicht beachtet. Es gibt in dieser Studie keine Vergleichsgruppe, es liegt auch kein ausreichender Individualvergleich vor, die Bewertung der röntgenologischen Veränderungen ist nicht definiert, die bakteriologische Technik ist nicht angegeben, auch nicht die Zeitabstände, in denen die Daten erhoben wurden. – Zu historischen Details der klinischen Prüfung von SM siehe [26 – 29].
Bei den Untersuchungen in den folgenden Jahren begegnen wir einem eigentümlichen Phänomen. Realisiert werden kontrollierte randomisierte Studien zunächst nur von 3 Institutionen in der Welt: von der US Veterans Administration in Kooperation mit Krankenhäusern der Armed Forces (VAAF), von dem US Public Health Service (USPHS) und vom British Medical Research Council (BMRC). Die 3 Einrichtungen haben folgende Merkmale gemeinsam: Sie sind Regierungsbehörden, und zwar mit dem ausdrücklichen Auftrag, Forschungen zu wichtigen Fragen der Volksgesundheit (und damit der Tuberkulose) zu initiieren, zu finanzieren und sich an der Planung und Durchführung zu beteiligen. Als Behörden kommt ihnen eine gewisse Autorität zu. In den zuständigen Abteilungen arbeiteten angesehene Kliniker zusammen mit Laborexperten und Statistikern oder Epidemiologen, alle mit wissenschaftlicher Erfahrung und organisatorischen Fähigkeiten, J.B. Barnwell und A.M. Walker bei den VAAF [30], C.E. Palmer und S.H. Ferebee beim USPHS [31], sowie P. D'A. Hart, M. Daniels und A.B. Hill beim BMRC [28,29]. Die Tuberkulose war von den Politikern als eine drohende Nachkriegsgefahr erkannt, so dass den Institutionen auch genügend Geld bewilligt wurde. Die VAAF verfügten über eigene Tuberkulose-

[2] Alternativ zum Zwei-Gruppen-Vergleich wäre auch eine Cross-Over-Versuchsanlage in Betracht zu ziehen, bei der jeder Patient als seine eigene Kontrolle dient.

krankenhäuser, der USPHS und der BMRC mussten geeignete Kliniken für die Studien gewinnen, wobei sich der USPHS auf die Trudeau Society, die wissenschaftliche US-Tuberkulosegesellschaft, stützte [32] und der BMRC auf die engen persönlichen Kontakte seiner Forscher mit großen Kliniken, vor allem in London. Es war eine wirklich einmalige Konstellation von Umständen, welche die Durchführung großer multizentrischer randomisierter kontrollierter Studien ermöglichte. Im deutschen Sprachraum gab es weder in Deutschland noch in der Schweiz oder Österreich entsprechende Institutionen, auch nicht in Frankreich. Nicht nur in all diesen Ländern wurden klinisch-therapeutische Prüfungen, zumindest im ersten Nachkriegsjahrzehnt, so gut wie gar nicht in der Weise durchgeführt und/oder publiziert, dass sie den Ansprüchen zur Vermeidung von Verzerrungen genügt hätten. Das trifft auch für die Prüfungen in USA und England zu, die nicht in Zusammenarbeit mit den genannten Regierungsinstitutionen durchgeführt wurden. 22 therapeutische Prüfungen, die in der Zeitschrift Chest von 1946 – 1954 veröffentlicht wurden [33 – 53], haben wir nach lediglich 4 Gesichtspunkten überprüft:

- kontrolliert als Kollektiv- oder Individualvergleich bzw. durch Paarbildung?;
- wenn Kollektivvergleich: randomisiert?;
- interkollektive Homogenität überprüft?;
- Ergebniskriterien adäquat und durch andere Wissenschaftler überprüf- und nachmachbar?

In keiner Studie wurde ein Individualvergleich durchgeführt, in einer eine Paarbildung, in 4 ein Kollektivvergleich, davon nur in einer mit Randomisierung. 17 von 22 Prüfungen waren also unkontrolliert. Bei den zur gleichen Zeit erschienenen Arbeiten in der American Review of Tuberculosis sah es etwas besser aus: in keiner von 17 Prüfungen [54 – 71] ein Individualvergleich, keine Paarbildung, in 7 ein Kollektivvergleich, 4 davon mit Randomisierung durch Alternation, nicht kontrolliert also 10 Untersuchungen; Prüfung auf interkollektive Homogenität in 7 Arbeiten, davon 4-mal mit unbefriedigendem Ergebnis. Keine der Arbeiten erfüllt aber alle 4 genannten Kriterien. In England sind Publikationen von chemotherapeutischen Prüfungen relativ spärlich. In 4 Zeitschriften sind wir auf 13 Veröffentlichungen gestoßen [72 – 84]. Von diesen waren nur 3 kontrolliert; 1 durch Individualvergleich, 1 durch Paarbildung, 1 durch Kollektivvergleich mit Alternierung. Keine der 13 Arbeiten erfüllte alle Kriterien. – Auch bei der VAAF und dem USPHS war es nicht möglich, von Anfang an eine Randomisierung durchzuführen. Es bedurfte einer „Erziehung“ ([30], S. 27) und des Lernens aus Fehlern und der zunehmenden Erfahrung mit kooperativen Prüfungen, die bei den Prüfern das Bewusstsein für die Relativität der eigenen Ergebnisse schärfte. Die VAAF begann mit Individualvergleichen [30]. Die konsequente Randomisierung innerhalb jeder beteiligten Klinik wurde erst im Oktober 1948 eingeführt [30], aber noch 1952 wurde eine kooperative nicht kontrollierte Studie mit Viomycin begonnen [54]. Über die Gründe findet sich nur eine sybillinische Antwort von Walker, zitiert bei [30]: „… for reasons that can be visualized we did not adopt the method“ (der Alternierung oder Randomisierung von Anfang an). Bei den vom USPHS und der Trudeau Society unterstützten SM-Studien gab es keine parallele Kontrollgruppe [32], jedoch in der ersten vom USPHS selbst organisierten und im November 1947 begonnenen multizentrischen SM-Großstudie. In ihr sind die Patienten durch Alternierung nach der Endzahl ihrer Patientennummer randomisiert [85]. Die ersten Isoniazid-Prüfungen in den USA sind ebenfalls nicht als kontrollierte Prüfungen angelegt [44, 45, 92]. Selbst Martini genügt in seiner Conteben-Studie nicht den eigenen Maßstäben: es gibt keine Trenddarstellung, nur eine tabellarische, nicht quantifizierte Beschreibung der Fälle, keine Teststatistik, keine Definition der Krankheitsschweregrade, keine Angaben zur bakteriologischen Technik, keine Erwähnung von Resistenzbestimmungen, über die Ausfälle wegen Verschlechterung nur die Angabe ihrer Zahl.

Was sind die Ursachen dieser desillusionierenden Differenzen zwischen Soll und Ist, zwischen Ideal und Realität? Sie sind auch in anderen Wissensgebieten anzutreffen, z. B. in der Ursachenforschung von Infektionskrankheiten ([86], S. 322 f.). Der Forderung, dem Problem-spezifischen Zweifel so weit wie möglich die Basis zu entziehen, wird nicht nachgekommen. Urteile sind dann Entscheidungen unter Unsicherheit, gefällt unter äußerem oder innerem Druck oder wegen praktischer Beschränkungen.

Zwei terminologische Zwischenbemerkungen

- Unter Randomisierung verstehen wir jedes Zuteilungsverfahren der Patienten auf die Gruppen nach dem Zufallsprinzip. In der Statistik wird jetzt oft nur dann von Randomisierung gesprochen, wenn jede mögliche Manipulierung bei der Gruppenbildung ausgeschlossen ist. Daher wird eine Alternierung nicht als Randomisierung klassifiziert. Man kann aber von einer alternierenden Zuteilung, die korrekt durchgeführt ist, nicht sagen, dass sie nicht randomisiert ist. Wir benutzen daher Randomisierung als Oberbegriff und unterscheiden bei Bedarf manipulierbare und nichtmanipulierbare Randomisierung. Eine „nichtmanipulierbare“ Randomisierung gibt es natürlich nur, wenn in ausreichendem Maß Kontrollinstanzen vorhanden sind. Man muss hier unterscheiden zwischen dem Zuteilungsverfahren (der Randomisierung) an sich und der Durchführung der Randomisierung.
- Zum Begriff der „*observational study*“ („Beobachtungsstudie“). Im statistisch-epidemiologischen Schrifttum ist die Definition von Cochran, 1965, allgemein gebräuchlich. Sie lautet nach Rosenbaum [116], in Übereinstimmung mit anderen Zitaten, so: Die observational study ist „an empirical comparison of treated and controlled groups in which the objective is to elucidate cause – and effect relationships [in which it] is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effect it is desired to discover, or to assign subjects at random to different procedures“. Also, ein Kollektivvergleich ist eine therapeutische „observational study“, wenn sie nicht *als willkürlicher, gezielter und auf seine Wirksamkeit kontrollierter Eingriff* angelegt werden kann oder nicht randomisiert werden kann (I). Der Einfachheit halber wollen wir die Wortfolge in (I) von „als“ bis „kontrollierter Eingriff“ auf „willkürlicher Eingriff*“ abkürzen. Wie der Gebrauch von (I) in der Literatur zeigt, wird (I) verstanden als: ” wenn er nicht als willkürlicher Eingriff* angelegt oder nicht randomisiert ist. (II). Diese verneinende Formulierung des „wenn“-Satzteils in (II) kann logisch korrekt ins Positive umgewandelt werden zu: wenn er ein willkürlicher Eingriff* ist *und* randomisiert ist (III). Dieser positive “wenn“-Satz (III) kann nun den „wenn“-Teil von (I) ersetzen: Ein Kollektivvergleich ist eine observational study, wenn es nicht der Fall ist, dass er als willkürlicher Eingriff* angelegt ist *und* nicht randomisiert ist (IV). Aber wie ist dann der häufig vorkommende Fall zu klassifizieren, der die Kriterien des willkürlichen Eingriffs* erfüllt, aber nicht randomisiert ist? Er wird von Cochrans Definition

nicht erfasst. Trotzdem wird in der statistisch-epidemiologischen Literatur auf „controlled, but not randomized“ „observational“ angewendet, siehe auch [116], obwohl ja überhaupt nicht rein beobachtet, sondern „controlled experimentation“ getrieben wird. Damit wird eine Bedeutungsverschiebung vorgenommen. Sie besteht sprachlogisch darin, dass Cochran zu „observational study“ einen anderen Begriff als Kontradiktion setzt als er in vielen Sprachen wie Englisch, Französisch, Deutsch üblich ist. In allen diesen Sprachen gilt als Kontradiktion, i. e. als eine bedeutungsgleiche Negation von „observational study“ „controlled study“ im Sinne von: „nicht: willkürlicher Eingriff*“. Cochran dagegen kreiert die neue Kontradiktion „nicht: willkürlicher Eingriff* und nicht: „randomisiert“. Dieser scheinbar minimale Unterschied zerstört unsere international verbreitete Gebrauchsweise von „observational“. Man sollte Cochrans Kontradiktion nicht übernehmen. Denn sie hat weitreichende Folgen für unseren Sprachgebrauch von „controlled“ und „observational“, Folgen, von denen wir heute noch nicht eindeutig sagen können, ob sie *generell* Nutzen bringen, solange das Leistungsverhältnis zwischen nicht-randomisierten und randomisierten kontrollierten Studien sowie der Erfolg der Verfahren zur Verbesserung unzulänglich randomisierter Studien nicht voll geklärt sind. Wir werden daher unter „observational study“ wie bisher eine geplante Untersuchung verstehen, in der eine Gruppe von Personen mit einem nicht in der Studie beigefügten (natürlichen) Merkmal verglichen wird mit einer gleichartigen Gruppe, die lediglich dieses Merkmal nicht aufweist. Cochrans Definition liefert die Suchregel für die weltweiten Literaturrecherchen nach kontrollierten Studien durch die Cochrane-Zentren. Dieser Regel fällt die größte Menge der Studien zum Opfer. Damit verwirft man ein gewaltiges Material, aus dem sich durch *gute* Metaanalysen vermutlich viele nicht weiter bezweifelbare Informationen herausholen ließen. – Psychologisch könnte man die skizzierte Bedeutungsverschiebung als wissenschaftspolitisch wohlbekannten Versuch interpretieren, mithilfe von Umdefinitionen ein Arbeitsgebiet zum Teilgebiet eines anderen zu machen, im vorliegenden Fall die klinische therapeutische Prüfung zu einem Teilgebiet der Epidemiologie, statt sie wie bisher als klinische Pharmakologie anzusehen.

Kontrollierte Studien in Deutschland, die Bedeutung der W.A.T.L.

Im deutschsprachigen Raum wurden zunächst keine Kollektivvergleiche durchgeführt. Die erste randomisierte Studie wurde unseres Wissens von Tanner und Merian in der Schweiz gemacht und 1958 veröffentlicht [87], die zweite wurde 1965 von Schütz und Bartmann in Deutschland publiziert [88]. Die ersten multizentrischen kontrollierten und randomisierten Prüfungen wurden in Deutschland 1964 von der Wissenschaftlichen Arbeitsgemeinschaft für die Therapie von Lungenkrankheiten (W.A.T.L.) begonnen [12]. Beide Autoren waren daran aktiv beteiligt. Die Abneigung oder Indifferenz gegen Kollektivvergleiche hatte eine Reihe von Gründen. Zunächst den Umstand, dass in diesem Raum die in USA und England gegebenen Voraussetzungen nicht bestanden. Dazu gehört auch, dass die Verteilung der Prüfsubstanzen und damit die Auswahl der Prüfer und die Formulierung der Prüfpläne nicht wie bei SM in USA und England maßgeblich in der Hand öffentlicher Institutionen lag, sondern von den forschenden Pharmafirmen bestimmt bzw. beeinflusst wurden. Im Falle des INH z. B. wurde in Deutschland die Testsubstanz von der Bayer AG wie bei TSC an einzelne Chefärzte vergeben, die in der Planung und Durchführung ihrer Prüfung offensichtlich weitgehend frei waren. In der Schweiz wurden die Prüfungen multizentrisch unter intensiver Mitwirkung der Firma Hoffmann La Roche zentral geplant, nach einem gemeinsamen Programm ohne Vergleichsgruppen durchgeführt und zentral ausgewertet. Verfasser der Publikation waren Mitarbeiter der Firma [89]. Prinzipiell gleich wurde bei der Prüfung von Cycloserin vorgegangen [90, 91]. Natürlich lag den Entdeckern und Herstellern von INH daran, möglichst schnell und ökonomisch Klarheit darüber zu gewinnen, ob sie mit einer Zulassung rechnen und die Großproduktion vorbereiten konnten. Auch die Prüfer waren daran interessiert, sich so schnell wie möglich ein Bild zu machen, primär der Sache wegen, sekundär oder tertiär aber auch, um durch frühe und daher häufig zitierte Veröffentlichungen den eigenen Bekanntheitsgrad zu steigern. Nicht anders war es in USA und England, wenn die ersten klinischen Untersuchungen durch die Pharmaindustrie initiiert wurden, siehe [92, 93], ebenso in Schweden. Dort wurden nach den ersten durch Firmen veranlassten Prüfungen vom Therapeutic Trials Committee of the Swedish National Association against Tuberculosis methodisch ausgezeichnete kontrollierte randomisierte multizentrische Prüfungen durchgeführt [94, 95]. – Viele Kliniker waren der Ansicht, dass ein Vergleich mit früheren Erfahrungen bei ähnlich gelagerten Fällen als Kontrolle ausreichend sei. Das kann aber nur für Krankheitsformen mit stets ungünstigem Ausgang gelten, nicht, wenn auch ohne spezifische Therapie Heilungen eintreten. Was tatsächlich zufällig passieren kann, wird oft unterschätzt, ebenso das Risiko einer Verallgemeinerung der eigenen Ergebnisse aus einer beschränkten Zahl von Beobachtungen. Der Mensch glaubt oft mehr zu wissen, als er jeweils wissen kann ([86], S. 333). Von dieser Schwäche sind auch Wissenschaftler nicht frei.

In den amerikanischen und englischen Studien, die nicht von VAAF, USPHS oder BMRC durchgeführt sind, werden Fragen der Versuchsplanung kaum angesprochen, nur in 5/52 der von uns zitierten Arbeiten. In den Veröffentlichungen aus dem deutschen Sprachraum ist das jedoch häufig der Fall. Meist werden unter Berufung auf Martini Kollektivvergleiche abgelehnt, weil eine ausreichende Homogenisierung der Gruppen unmöglich sei [20, 23, 96 – 99, 100]. Weitere Argumente sind: zu großer praktischer Aufwand [98, 100]; zwangsläufig eingeschränkte Zuverlässigkeit in Großversuchen wie denen von VAAF und USPHS [98, 102, 103], ethische Unverantwortbarkeit, wobei manchmal irrtümlich unterstellt wird, dass zum Vergleich nur eine Plazebo-Gruppe dienen könne. Zu diesen Einwänden ist zu sagen, dass, wie bereits ausgeführt, die sich auf Martini stützenden Gegenargumente nicht stichhaltig sind. – Bei den Großversuchen können die Ergebnisse nicht die sein, die mit der besten röntgenologischen und bakteriologischen Technik herauskämen, sondern Werte, die vermutlich zu optimistisch sind, weil nicht-optimale Untersuchungstechniken wegen unzureichender Sensitivität die Raten von Kavernenschluss und Sputumkonversion scheinbar erhöhen. Das muss aber nicht den Vergleich der Gruppen verzerren. Eine Illustration zu den Problemen der Großversuche liefert die Diskussionsbemerkung eines Teilnehmers der VAAF-Studien [104]. Die Auseinandersetzung mit methodischen Fragen hatte aber nicht zur Folge, dass die eigenen Untersuchungen der meisten Kritiker des Kollektivvergleichs methodisch gut geplant waren. Das trifft nach einer Analyse von 22 Studien durch Trendelenburg auch noch für die Jahre 1960 – 1963 zu [105].

Ethische Aspekte kontrollierter Studien

Die ethischen Aspekte sind immer wieder diskutiert worden, bis in die Gegenwart, besonders nachdem die ersten wirksamen Antituberkulotika zur Verfügung standen. Es sei nicht verantwortbar, der Vergleichsgruppe eine wirksame Therapie vorzuenthalten [98 – 100]. Dieser Einwand ist berechtigt, war aber in praxi nicht aktuell, da es immer wieder Patienten gab, deren Bakterien gegen die verfügbaren wirksamen Antituberkulotika resistent geworden waren, so dass man ihnen überhaupt kein ausreichend wirksames Mittel vorenthalten konnte. Bei den Vergleichen verschiedener Behandlungsschemata aus bekannten wirksamen Substanzen stellt sich das Problem gar nicht, man sucht ja nur nach einer gegenüber dem Standard besseren Lösung. Ein weiterer Argumentationsstrang betrifft die Frage einer wesentlichen Störung des Arzt-Patientenverhältnisses. Der Patient werde zu einer Nummer, zum Versuchskaninchen degradiert; kontrollierte, randomisierte Prüfungen würden gegen den hippokratischen Eid verstoßen usw. ([106 – 109]; als Antwort: [110 – 112]). Das Problem adäquat zu behandeln würde den Rahmen dieses Aufsatzes sprengen. Unter der Durchführung einer kontrollierten Prüfung muss das Verhältnis zwischen behandelndem Arzt und Patient nicht leiden, vorausgesetzt, der Patient wurde fair über das Vorhaben aufgeklärt und nicht zur Teilnahme gedrängt. Der Patient wird ja nicht bei der Durchführung zur Nummer, sondern erst bei der Auswertung, welche das Arzt-Patientenverhältnis überhaupt nicht berührt. Die Einstellung für oder gegen kontrollierte Studien hängt letztlich davon ab, welche ethische Grundposition man vertritt. Die Gegner sind der Auffassung, die Lilford [107] klar formuliert hat, dass „the obligation to respect individual autonomy outweighs the common good in all but most extreme cases“. Die Befürworter dagegen geben der Gemeinschaft gegenüber dem Individuum Priorität, weil kein Mensch im Laufe seines Lebens ohne Mitmenschen weiter existieren kann, und daher Handlungen, welche dem Wohl der Gemeinschaft dienen, letztlich in der Regel auch dem Einzelnen dienen. Über die Frage der Rangordnung dieser beiden Werte kann man nicht mehr diskutieren, weil wir mit ihr die höchste Ebene hinter uns gelassen haben, auf der noch sachliche Argumente zur Verfügung stehen. Offensichtlich hält die Mehrzahl der Ärzte kontrollierte randomisierte Studien, die unter gebührender Beachtung der humanen Aspekte durchgeführt werden, nicht für unethisch, und sie werden von Patienten akzeptiert.

Neuere methodische Aspekte des Kollektiv-Vergleichs

Wir wollen uns nun wieder den methodischen Aspekten der Kollektivvergleiche zuwenden. Dass solche Vergleiche möglich und brauchbar sind, steht außer Frage. Aber wie verlässlich sind die Ergebnisse? Die Randomisierung soll gegen Heterogenität der Gruppen schützen. Das kann sie, wenn nicht stratifiziert wird, zuverlässig nur, wenn der Gruppenumfang groß genug ist, wie zuvor ausgeführt; dies war nur in wenigen Großprüfungen der Fall. In den letzten Jahrzehnten sind verschiedene Methoden entwickelt worden, um Ungleichheiten bei der zufälligen Zuteilung vorzubeugen, was übrigens Martini schon 1940 praktiziert hat und „ausgleichende Alternierung“ nannte ([6], S. 17; [113]). Sie wird heute als „eingeschränkte Randomisierung“ bezeichnet. Damit lässt sich zwar ein guter Ausgleich für bekannte Einflussfaktoren wie Alter oder Krankheitstyp usw. erreichen. Wie sich dabei die Verteilung von nicht erfassten bekannten und unbekannten Einflussfaktoren verschieben kann, ist bisher nicht bekannt. Das CPMP-Komitee der European Agency for the Evaluation of Medicinal Products (EMEA) schreibt 2003: „Dynamic allocation (Verfahren, welche die zufällige Zuteilung der Patienten bei interkollektiver Ungleichheit der Einflussgrößen durch eine kompensierende Zuteilung unterbrechen) is strongly discouraged“ ([114], S. 4). – Mit weiteren Argumenten für die Randomisierung haben sich Abel und Koch auseinander gesetzt ([115]; s. jedoch [129]), u. a. dem, Randomisierung sei die Basis für statistische Signifikanzteste. Die Randomisierung ist jedoch weder hinreichende noch notwendige Bedingung für verlässliche derartige Schlüsse. Die Randomisierung kann auch nicht als Basis für Schlüsse auf die kausale Rolle einer Behandlung bei beobachteter Besserung dienen. Statistische Analysen können niemals eine Verursachung nachweisen. Das geht nur mit dem Kausalexperiment [86]. Die Randomisierung ist auch keine hinreichende oder notwendige Bedingung für eine maskierte („verblindete“) Zuteilung der Patienten. Der einzige methodische Grund für eine Randomisierung ist der Ausgleich unbekannter bzw. bekannter, aber nicht quantifizierbarer oder nicht komparativ ordnungsfähiger Einflussfaktoren (z. B. Dispositionen wie Infektionsabwehr). Deswegen muss, wer spezifische Zweifel soweit wie möglich ausräumen will, randomisieren. Wegen des oft nicht voll befriedigenden Ausgleichs muss statistisch zu Beginn und – wegen der Ausfälle – auch am Ende der Studie geprüft werden, ob die Gruppen als Stichproben aus der gleichen Grundgesamtheit anzusehen sind – siehe dazu [128]. Metaanalysen von weiteren gleichartigen Prüfungen müssen zeigen, dass das Therapieergebnis ausreichend reproduzierbar ist. Die Metaanalysen müssen später unbedingt durch langfristige Nachbeobachtungen, die ihre eigenen methodischen Probleme haben, ergänzt werden.

Die Aussagekraft therapeutischer nichtrandomisierter Kollektivstudien wurde mithilfe verschiedener statistischer Verfahren verbessert (Übersichten in [116,117]). Sie zielen auf eine Schätzung der möglichen Einflüsse unbekannter Faktoren und auf die nachträgliche Bereinigung bei bekannten Faktoren ab. Unterschiede in der Verlässlichkeit zwischen randomisierten und nichtrandomisierten Studien sind wegen methodischer Mängel der durchgeführten Metaanalysen [118 – 126] nicht abschätzbar. Dazu mit Recht Eysenck: „A good review is based on intimate personal knowledge of the field, the participants, the problems that arise, the reputation of different laboratories, the likely trustworthiness of individual scientists, and other partly subjective but extremely relevant considerations. Meta-analysis rules out any such subjective factors. It can be done by simply feeding the published results to a computer (so geschehen in den betreffenden Metaanalysen) and coming up with an effect size. The computer avoids the bias of the subjective approach but simply adds together the biases of the authors of the original reports – which may or may not balance out“ [127].

Schlussfolgerung

Alle Arten des Vergleichs von Kollektiven haben einen Schwachpunkt gemeinsam: die Feststellung der Vergleichbarkeit der Gruppen, des Grades der interkollektiven Homogenität, ihrer Ähnlichkeit bezüglich der Wirkungsmöglichkeiten eines Therapeutikums. Die zur Verfügung stehenden statistischen Verfahren erlauben nur eine Aussage darüber, ob 2 Stichproben aus derselben Grundgesamtheit stammen. Diese Aussage bezieht sich lediglich auf statistische Signifikanteste, die Überlegenheit nachweisen sollen. Gleichheit kann mit diesem Ansatz nicht nachgewiesen werden. Für derartige Fragestellungen könnte man aber Äquivalenz- oder Nicht-Unterlegenheitsstudien heranziehen, wobei für die diskutierte Vergleichbarkeit von Gruppen die direkte Gegenüberstellung mittels eines Tests insgesamt

kritisch zu bewerten ist [128]. Gefragt ist aber nach der Ähnlichkeit der Wirkungsmöglichkeiten der Therapeutika *zwischen* den Mitgliedern einer Grundgesamtheit. Diese Kenntnis ist erforderlich, weil auch zwischen den Mitgliedern derselben Grundgesamtheit, die sich ja nicht durch gleiches Einwirkungsvermögen eines Therapeutikums definieren lässt, mit erheblichen Differenzen in den Merkmalen gerechnet werden muss, welche die Bedingungen des Erfolges sind, die ein Therapeutikum haben kann. Die meisten dieser Bedingungen wie Infektionsabwehr oder Regenerationsfähigkeit, die zudem nur als Komplex positiver und negativer Einflüsse wirksam werden, sind qualitativ oft, quantitativ meist derzeit nicht erfassbar. Und selbst wenn sie es wären: Umfang und Grad der notwendigen Übereinstimmung, deren Kenntnis für ein Urteil über die Vergleichbarkeit erforderlich ist, sind keine für alle Vergleiche feste Größe, was ja auch die Entwicklung der antituberkulösen Therapie zeigt. Der notwendige Grad der Vergleichbarkeit hängt, wie schon erwähnt, wesentlich auch von Eigenschaften der Testsubstanz, von ihrer Wirksamkeit und z.T. auch von ihrem Wirkungstyp ab. Ist die Wirksamkeit groß, werden viele Störfaktoren irrelevant. Da die klinische Wirkung einer neuen Substanz nicht sicher vorauszusehen ist, sollte die Vergleichbarkeit der Gruppen bei ganz neuen Mitteln so gut wie möglich gesichert werden. Die Basis therapeutischer Prüfungen ist der Vergleich, die Basis des Vergleichs ist die Vergleichbarkeit, die zumindest gegenwärtig nicht durch Maß und Zahl bestimmt werden kann, sondern subjektiv gefärbt mittels Erfahrung und Intuition geschätzt werden muss. Hierfür sollte nach Abhilfe gesucht werden. Vielleicht könnten Verfahren der taxonomischen Statistik nützlich sein.

Die Methodologie therapeutischer Prüfungen kann sich also über einen Mangel an Aufgaben nicht beklagen. Und: Problemlösungen evozieren neue Probleme, solange der Mensch forscht.

Abkürzungen

Medikamente

CS	=	Cycloserin
DATC	=	Thiocarlid
ETH	=	Ethionamid
INH	=	Isoniazid
PAS	=	Para-Aminosalizylsäure
SM	=	Streptomyzin
TSC	=	alle Thiosemicarbazone

Andere Abkürzungen

BMRC	=	British Medical Research Council
CPM-Komitee	=	Committee for Proprietary Medicinal Products
EMEA	=	European Agency for the Evaluation of Medicinal Products
ICH	=	International Conference on Harmonisation

Danksagung

Wir danken Herrn Professor Dr. Dierk Brockmeier, Gießen, Frau Dr. rer. medic. Nicole Heussen, Aachen, und Herrn Franz C. von Lichtenberg M. D, Boston, MA, U.S.A., herzlich für die kritische Durchsicht des Manuskriptes.

Literatur

1 *Martini P, Rosendahl A.* Bilanz der Goldtherapie der Lungentuberkulose. Z Tuberk 1938; 80: 20 – 26 und 1940; 84: 330 – 340
2 *Amberson JrJB, McMahon BT, Pinner M.* A clinical trial of Sanocrysin pulmonary tuberculosis. Am Rev Tuberc 1931; 24: 401 – 435
3 *Martini P.* Die Gesetze der Prüfung von Heilmitteln bei Lungentuberkulose. Beitr Klin Tuberk 1934; 84: 86 – 98
4 *Martini P.* Richtlinien zur Prüfung von Heilmitteln bei Tuberkulose. Z Tuberk 1950; 94: 117 – 128
5 *Martini P, Moers H, Gansen H.* Conteben in der Behandlung der Lungentuberkulose. Beitr Klin Tuberk 1951; 104: 515 – 578
6 *Martini P.* Methodenlehre der therapeutisch-klinischen Forschung. Berlin, Göttingen: Springer-Verlag, 1947
7 *Martini P, Oberhoffer G, Welte E.* Methodenlehre der therapeutisch-klinischen Forschung, 4. Aufl. Berlin, Heidelberg, New York: Springer-Verlag, 1968
8 *Kalkoff KW.* Zur Behandlung der Hauttuberkulose mit Tb I 698/E. In: Domagk G (Hrsg). Chemotherapie mit den Thiosemikarbazonen. Stuttgart: Georg Thieme Verlag, 1950: 142 – 172
9 *Arold C.* Die Erfolgsmöglichkeiten bei der Tuberkulose der oberen Luftwege mit älteren und neueren Behandlungsmethoden. In: Domagk G (Hrsg). Chemotherapie der Tuberkulose mit den Thiosemikarbazonen. Stuttgart: Georg Thieme Verlag, 1950: 173 – 187
10 *Schürmann F, Radenbach KL.* Tuberkulose und das Thiosemicarbazon TbI/698. Behandlungsergebnisse bei Lungentuberkulose mit sekundärer Kehlkopftuberkulose. Schweiz Z Tuberk 1950; 7: 99 – 114
11 *Carstensen B.* Para-aminosalicylic acid (PAS) in pulmonary and extrapulmonary tuberculosis. Am Rev Tuberc 1950; 61: 613 – 620
12 *Wissenschaftliche Arbeitsgemeinschaft für die Therapie von Lungenkrankheiten (W.A.T.L.).* Kooperative, kontrollierte Prüfung von Thiocarlid (DATC), PAS und Bettruhe in kurzfristiger Monotherapie bei kavernöser vorbehandelter Lungentuberkulose. Beitr Klin Tuberk 1969; 139: 115 – 139
13 *Martini P.* Grundsätzliches und Methodisches zur therapeutisch-klinischen Forschung. Dtsch Med Wochenschr 1949; 74: 1349 – 1353
14 *Martini P.* Grundsätzliches zur therapeutisch-klinischen Versuchsplanung. Method Inf Med 1962; 1: 1 – 5
15 *Auersbach K, Bartmann K, Kauffmann G-W et al.* Die frühe Erkennung des ungenügenden Effekts der chemisch-konservativen Behandlung bei kavernöser Lungentuberkulose. Fortschr Tuberk Forsch 1961; 11: 122 – 192
16 *Bartmann K.* Statistische Überlegungen. In: Walter AM (Hrsg). Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien. Stuttgart: Georg Thieme Verlag, 1958: 47 – 49
17 *Bartmann K, Radenbach KL, Zierski M.* Wandlungen in den Auffassungen und der Durchführung der antituberkulösen Chemotherapie. Prax Klin Pneumol 1985; 39: 397 – 420
18 *Hill AB.* The clinical trial. N Engl J Med 1952; 247: 113 – 119
19 *Heilmeyer L.* Diskussionsbemerkung. In: Walter AM (Hrsg). Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien. Stuttgart: Georg Thieme Verlag, 1958: 46 – 47
20 *Trendelenburg F.* Kollektivstatistische Befundanalysen – Individuelle Verlaufsbeobachtung. In: Walter AM (Hrsg). Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien. Stuttgart: Georg Thieme Verlag, 1958: 50 – 55
21 *Berg G.* Schlusswort. In: Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien. Stuttgart: Georg Thieme Verlag, 1958: 55 – 56
22 *European Medicines Agencies (EMEA).* Note for guidance on choice of control group in clinical trials (CPMP/ICH/364/96). Im Internet unter http://www.emea.eu.int, 2001
23 *Düggeli O, Trendelenburg F.* Frühergebnisse der Rimifon – Behandlung bei Lungentuberkulose. Schweiz Z Tuberk 1952; 9: 267 – 280
24 *Hinshaw HC, Feldman WH.* Evaluation of chemotherapeutic agents in clinical tuberculosis. Am Rev Tuberc 1945; 50: 202 – 213
25 *Pfuetze KH, Glover RP, White JrEF et al.* Clinical use of streptomycin in the treatment of tuberculosis. Chest 1946; 12: 515 – 519
26 *Hinshaw HC.* Historical notes on earliest use of streptomycin in clinical tuberculosis. Am Rev Tuberc 1954; 70: 9 – 14
27 *Feldman WH.* Streptomycin: some historical aspects of its development as a chemotherapeutic agent in tuberculosis. Am Rev Tuberc 1954; 69: 859 – 868
28 *Yoshioka A.* Use of randomisation in the Medical Research Council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940's. BMJ 1998; 317: 1220 – 1223

29 *Hart PDA.* A change in scientific approach: from alternation to randomized allocation in clinical trials in the 1940's. BMJ 1999; 319: 572 – 573
30 *Tucker WB.* The evolution of the cooperative studies in the chemotherapy of tuberculosis of the Veterans Administration and Armed Forces of the USA. Fortschr Tuberk Forsch 1960; 10: 1 – 68
31 *Comstock GW.* In memoriam Carroll Edwards Palmer. Am J Epidemiol 1972; 95: 305 – 307
32 *Riggins HM, Hinshaw HC.* Streptomycin-tuberculosis research project of the American Trudeau Society. Am Rev Tuberc 1949; 59: 142 – 167
33 *Robitzek EH, Ornstein GG, Slater P et al.* Diasone in the treatment of pulmonary tuberculosis. Chest 1946; 12: 185 – 204
34 *Kettelkamp GD, Friedman B.* Diasone therapy in pulmonary tuberculosis. Chest 1947; 13: 23 – 32
35 *Eastlake JrC, Barach AL.* Use of para-aminosalicylic acid in chronic pulmonary tuberculosis. Chest 1949; 16: 1 – 14
36 *Rubin EH, Steinbach MM, Leiner GC et al.* Streptomycin in tuberculosis. Chest 1949; 16: 304 – 328
37 *Sweany HC, Turner GC, Lichtenstein M et al.* A preliminary report on the use of para-aminosalicylic acid in the treatment of pulmonary tuberculosis. Chest 1949; 16: 633 – 656
38 *Burns HA, Feldman WH, Hinshaw HC et al.* Treatment of tuberculosis with promizole: a clinical investigation with matched controls. Chest 1949; 16: 867 – 869
39 *Potter BP.* A clinical appraisal of the value of para-aminosalicylic acid with and without streptomycin in the treatment of tuberculosis. Chest 1950; 17: 509 – 523
40 *DeJanney NH, Cox J, Grindell-Balchum E.* Preliminary report of clinical experience with p-aminosalicylic acid. Chest 1950; 18: 413 – 429
41 *Davis JD, Netzer S, Schwartz JA et al.* Tibione: laboratory and clinical studies. Chest 1950; 18: 521 – 527
42 *Dunner E, Brown WB.* Streptomycin para-aminosalicylate in the treatment of pulmonary tuberculosis. Chest 1951; 19: 438 – 443
43 *Hughes FJ, Mardis RE, Dye WE et al.* Combined intermittent regimens in the treatment of non-miliary pulmonary tuberculosis. Chest 1952; 21: 1 – 16
44 *Selikoff IJ, Robitzek EH.* Tuberculosis chemotherapy with hydrazine derivatives of isonicotinic acid. Chest 1952; 21: 385 – 438
45 *Witkind E, Willner I.* Clinical experiences with isonicotinic acid hydrazide in tuberculosis. Chest 1953; 23: 16 – 27
46 *Krieser AE, Sanderson AG, Vik M et al.* Effects of isonicotinic acid hydrazide in mentally ill patients. Chest 1953; 23: 28 – 35
47 *Pitts FW, O'Dell ET, Fitzpatrick MJ et al.* Intermittent viomycin therapy in pulmonary tuberculosis: employed singly and in combination with intermittent streptomycin or daily para-aminosalicylic acid. Chest 1953; 23: 241 – 254
48 *Cohen SS, Frost RH, Yue W-Y.* Tibione in the treatment of pulmonary tuberculosis. Chest 1953; 23: 507 – 517
49 *Ziskind MM, Calovich E, Joffko J et al.* Use of the iso-nicotinic acid hydrazides in the treatment of tuberculosis. Chest 1953; 24: 535 – 544
50 *Sweany HC, Perez JA.* The present status of isoniazid in the treatment of pulmonary tuberculosis. Chest 1954; 25: 374 – 389
51 *Cheifetz I, Paulin C, Tuatay H et al.* Iproniazid in pulmonary tuberculosis. Chest 1954; 23: 390 – 396
52 *Cohen S, Ang E.* Treatment of pulmonary tuberculosis with isoniazide and iproniazide. Chest 1954; 25: 622 – 639
53 *Ertug C, Easom HF.* A study on therapeutic evaluation of isoniazid in treatment of pulmonary tuberculosis. Chest 1954; 26: 138 – 145
54 *Tucker WB.* Re-treatment of advanced pulmonary tuberculosis with viomycin. Am Rev Tuberc 1954; 70: 812 – 840
55 *Hinshaw HC, Feldman WH, Pfuetze KH.* Streptomycin in treatment of clinical tuberculosis. Am Rev Tuberc 1946; 54: 191 – 201
56 *Jenkins DE, Peck WM, Reid JR et al.* Effect of streptomycin on early tuberculous pulmonary lesions. Am Rev Tuberc 1947; 56: 387 – 395
57 *Tempel CW, Hughes JrFJ, Mardis RE et al.* Combined intermittent regimens employing streptomycin and para-aminosalicylic acid in the treatment of pulmonary tuberculosis. Am Rev Tuberc 1951; 63: 295 – 311
58 *Robitzek EH, Selikoff IJ.* Hydrazine derivatives of isonicotinic acid (Rimifon, Marsilid) in the treatment of active progressive caseous-pneumonic tuberculosis. Am Rev Tuberc 1952; 65: 402 – 428
59 *Miller FL, Sands JH, Walker R et al.* Combined daily terramycin and intermittent streptomycin in the treatment of pulmonary tuberculosis. Am Rev Tuberc 1952; 66: 534 – 541
60 *Childress WG, Norman JW, Ott JrRH et al.* Observations on the effect of amithiozone (Tibione) in selected tuberculous pulmonary lesions. Am Rev Tuberc 1952; 65: 692 – 708
61 *Cohen SS, Johnsen L, Lichtenstein MR et al.* A comparative study of streptomycin and dihydrostreptomycin in pulmonary tuberculosis. Am Rev Tuberc 1953; 68: 229 – 237
62 *Mahady SCF, Armstrong FL, Beck F et al.* A comparative study of streptomycin and dihydrostreptomycin in pulmonary tuberculosis. Am Rev Tuberc 1953; 68: 238 – 248
63 *Miller FL, Sands JH, Gregory LJ et al.* Daily tetracycline (Terramycin) and intermittent streptomycin in the treatment of pulmonary tuberculosis. Am Rev Tuberc 1954; 69: 58 – 64
64 *Campagna M, Calix AA, Hauser G.* Observations on the combined use of pyrazinamide (Aldinamide) and isoniazid in the treatment of pulmonary tuberculosis. Am Rev Tuberc 1954; 69: 334 – 350
65 *Adcock JD, Whintrop ND, Haley RR et al.* The use of viomycin in patients with pulmonary tuberculosis. Am Rev Tuberc 1954; 69: 543 – 553
66 *Rothstein E, Johnson MP.* Streptomycin once weekly in the treatment of pulmonary tuberculosis. Am Rev Tuberc 1954; 69: 980 – 990
67 *Schwartz WS, Moyer RE.* The chemotherapy of pulmonary tuberculosis with pyrazinamide alone and in combination with streptomycin, para-aminosalicylic acid, or isoniazid. Am Rev Tuberc 1954; 70: 413 – 422
68 *Payne HM, Quarles C, McKnight HV et al.* Intermittent use of streptomycyilidene isonicotinoyl hydrazine sulphate in the therapy of pulmonary tuberculosis. Am Rev Tuberc 1954; 70: 701 – 713
69 *Reisner D, Peizer LR, Widelock D.* Isoniazid in single and multiple drug regimens in the treatment of pulmonary tuberculosis of recent origin. Am Rev Tuberc 1955; 71: 841 – 859
70 *Armstrong FL, Beck F, Horton R et al.* Streptomycin and para-aminosalicylic acid in pulmonary tuberculosis. I. Results during treatment: 528 patients. Am Rev Tuberc 1955; 72: 242 – 244
71 *Philips S, Larkin JrJC.* Viomycin in re-treatment of pulmonary tuberculosis. Am Rev Tuberc 1955; 72: 843 – 845
72 *Weitzman D, de Wend Caley FE, Wingfield AL.* Streptomycin in the treatment of pulmonary tuberculosis. Brit J Tuberc 1950; 44: 98 – 104
73 *Livingstone R, Street EW.* Thiosemicarbazones in the treatment of pulmonary tuberculosis. Tubercle (Lond) 1951; 32: 8 – 14
74 *Fielding J, Maloney JJ.* Calciferol, streptomycin, and paraaminosalicylic acid in pulmonary tuberculosis. Lancet 1951; 258: 614 – 617
75 *Greenberg MJ.* Use of thiocarbazone with streptomycin in pulmonary tuberculosis. Tubercle (Lond) 1952; 33: 53 – 56
76 *Sita Lumsden EG, Swoboda JAF.* Isoniazid in the treatment of pulmonary tuberculosis. Tubercle (Lond) 1952; 33: 322 – 329
77 *Edwards PW, Penman AC, Cutbill LJ.* Sulphone and streptomycin in pulmonary tuberculosis. BMJ 1952; 1: 1224 – 1226
78 *Joiner CL, MacLean KS, Pritchard EK et al.* Isoniazid in pulmonary tuberculosis. Lancet 1952; 260: 843 – 849
79 *McLaren Todd R.* Treatment of primary tuberculosis with P.A.S. BMJ 1953; 1: 1247 – 1249
80 *Stewart SM, Turnbull FWA, Crofton JW.* The use of oxytetracycline in preventing or delaying isoniazid resistance in pulmonary tuberculosis. BMJ 1954; 2: 1508 – 1511
81 *Houghton LE.* Combined corticotrophin therapy and chemotherapy in pulmonary tuberculosis. Lancet 1954; 263: 595 – 598
82 *Marmion T.* A clinical appraisal of cyanacetic acid hydrazide in chronic pulmonary tuberculosis. Brit J Tuberc 1955; 49: 9 – 19
83 *Charles F.* Treatment of pulmonary tuberculosis with intravenous PAS-infusions. Tubercle (Lond) 1955; 36: 40 – 42
84 *Clegg JW.* Clinical trial of P.A.S. salt of isoniazid in treatment of pulmonary tuberculosis. BMJ 1955; 2: 1004 – 1005
85 *Long ER, Ferebee SH.* A controlled investigation of streptomycin treatment in pulmonary tuberculosis. Pub Health Rep 1950; 65: 1421 – 1451
86 *Bartmann K.* Kritik der Ursachenforschung bei Infektionskrankheiten. Stuttgart: Wissenschaftliche Verlagsgesellschaft, 2001
87 *Tanner E, Merian P.* Klinische Erfahrungen mit Viomycin (Vionactan-Pantothenat). In: Walter AM (Hrsg). Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien. Stuttgart: Georg Thieme Verlag, 1958: 64 – 70
88 *Schütz I, Bartmann K.* A controlled clinical trial of ethionamide (ETH) + cycloserine (CS) + PAS six times weekly versus 4 times weekly in the first phase of clinical retreatment. Excerpta Medica. International

Congress Series No. 96. XVIIIe Conference internationale de la tuberculose, 1965: 115
89 *Fust B, Wernsdorfer G, Wernsdorfer W.* Rimifon. Schweiz Z Tuberk 1955; Supplementum ad vol. 12: 1 – 344
90 *Heilmeyer L.* D-Cycloserin bei Lungentuberkulose. In: Walter AM (Hrsg). Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien Stuttgart: Georg Thieme Verlag, 1958: 113 – 122
91 *Isebarth R, Wiedemann O.* D-Cycloserin bei Lungentuberkulose. Tuberkulosearzt 1960; 14: 144 – 162
92 *Ferebee SH, Long ER.* Isoniazid in the treatment of tuberculosis with a review of recent experience in the United States. Bull Int Union Tuberc 1954; 23: 50 – 87
93 *Mitchell RS, Bower GC.* Review of the English language literature on the clinical uses of chemotherapy in the treatment of pulmonary tuberculosis, 1954 – 1955 – 1956. Bull Int Union tuberc 1958; 28: 3 – 45
94 *The Therapeutic Trials Committee of the Swedish National Association against Tuberculosis.* Para-aminosalicylic acid treatment in pulmonary tuberculosis. Am Rev Tuberc 1950; 61: 597 – 612
95 *The Therapeutic Trials Committee of the Swedish National Association against Tuberculosis.* Isonicotinic acid hydrazide (isoniazid, INH) treatment in pulmonary tuberculosis. Bull Int Union Tuberc 1953; 23: 140 – 151
96 *Fanconi G, Löffler W.* Einleitung. In: Fanconi G, Löffler W (Hrsg). Streptomycin und Tuberkulose. Basel: Benno Schwabe Verlag, 1958: 7 – 10
97 *Mordasini E.* Streptomycinbehandlung bei Lungentuberkulose. In: Fanconi G, Löffler W (Hrsg). Streptomycin und Tuberkulose. Basel: Benno Schwabe Verlag, 1958: 187 – 207
98 *Schaich W, Stadler L, Keiderling W.* Ergebnisse einer zweijährigen Conteben (TbI/698)-Behandlung der Tuberkulose in der Medizinischen Klinik Freiburg i. Br. und Heilstätte St. Blasien. Beitr Klin Tuberk 1951; 104: 465 – 514
99 *Berg G.* Möglichkeiten und Grenzen der Chemotherapie der Lungentuberkulose. In: Freerksen E (Hrsg) Jahresbericht Borstel 1956/57. Berlin-Göttingen-Heidelberg: Springer Verlag, 1957: 1 – 42
100 *Brecke F, Wentz D.* Klinische Erfahrungen mit der Kombination Pyrazinamid-Isoniazid. In: Walter AM (Hrsg). Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien. Stuttgart: Georg Thieme Verlag, 1958: 78 – 93
101 *Klee P.* Die Chemotherapie der Lungentuberkulose mit Thiosemikarbazonen. In: Domagk G (Hrsg). Chemotherapie der Tuberkulose mit den Thiosemikarbazonen Stuttgart: Georg Thieme Verlag, 1950: 267 – 301
102 *Berg G.* Grundprinzipien für die Prüfung neuer Tuberkulostatika. In: Walter AM (Hrsg). Neue Tuberkulostatika und Tuberkulostatika-Resistenz von Tuberkelbakterien. Stuttgart: Georg Thieme Verlag, 1958: 43 – 46
103 *Berg G.* Rationelle tuberkulostatische Therapie. Z Tuberk 1961; 116: 165 – 171
104 *Dr. Salkin.* Diskussionsbemerkung. Trans 17th conference on the chemotherapy of tuberculosis. Washington, Veterans Administration, 1958: 45 – 46
105 *Trendelenburg F.* Über die Qualität chemotherapeutischer Veröffentlichungen. Beitr Klin Tuberk 1963; 127: 235 – 236
106 *Waugh WG.* A blast of the trumpet against the monstrous regiment of mathematics. BMJ 1951; 2: 1088
107 *Lilford RJ.* Ethics of clinical trials from a bayesian and decision analytic perspective: whose equipoise is it anyway? BMJ 2003; 326: 980 – 981
108 *Retsas S.* Treatment at random: the ultimate science or the betrayal of Hippocrates? J Clin Oncol 2004; 22: 5005 – 5008
109 *Retsas S.* In reply. J Clin Oncol 2005; 23: 4469
110 *Editor.* Out, damned spot. BMJ 1951; 2: 074 – 1076
111 *Hill AB.* Medical ethics and controlled trials. BMJ 1963; 1: 1043 – 1049
112 *Wieand S, Murphy K.* A commentary on treatment at random: the ultimate science or betrayal of Hippocrates? J Clin Oncol 2004; 22: 5009 – 5011
113 *Blittersdorf F.* Die Versuchsplanung bei akuten Infektionskrankheiten, speziell die Bedeutung der „ausgleichenden Alternierung", dargestellt am Beispiel des Scharlachs. Method Inf Med 1963; 2: 134 – 137
114 *European Medicines Agency (EMEA).* Note for guidance on statistical principles for clinical trials (CPMP/ICH/363/96). Im Internet unter http://www.emea.int,
115 *Abel U, Koch A.* The role of randomisation in clinical studies: myths and beliefs. J Clin Epidemiol 1999; 52: 487 – 497
116 *Rosenbaum PR.* Observational Study. In: Everitt BS, Howell DC (eds). Encyclopedia of Statistics in Behavioral Science. Chichester: Wiley & Sons, 2005; Vol 3: 1451 – 1462
117 *Schneider B.* Beobachtungsstudien als Mittel der Erkenntnisgewinnung über die Wirksamkeit von Arzneimitteln. Im Internet unter: www.mh-hannover.de/fileadmin/institute/biometrie/skripte/speziel/beob_studien/pdf, 2001
118 *Chalmers TC, Matta RJ, Smith JrH et al.* Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med 1977; 297: 1091 – 1096
119 *Sacks H, Chalmers TC, Smith JrH.* Randomized versus historical controls for clinical trials. Am J Med 1982; 72: 233 – 240
120 *Colditz GA, Miller JN, Mosteller F.* How study design affects outcomes in comparisons of therapy. I. Medical. Stat Med 1989; 8: 441 – 454
121 *Miller JN, Colditz GA, Mosteller F.* How study design affects outcomes in comparisons of therapy. II. Surgical. Stat Med 1989; 8: 455 – 466
122 *Joannidis JPA, Haidisch A-B, Pappa M et al.* Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001; 286: 821 – 830
123 *Kunz R, Oxman AD.* The unpredictability paradox: review of empirical comparisons of randomized and non-randomized clinical trials. BMJ 1998; 317: 1185 – 1190
124 *Benson K, Hartz AJ.* A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000; 342: 1878 – 2886
125 *Concato J, Shah N, Horwitz RI.* Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000; 342: 1887 – 1892
126 *Pocock SJ, Elbourne DR.* Randomized trials or observational tribulations? N Engl J Med 2000; 342: 1907 – 1909
127 *Eysenck HJ.* Systematic reviews: meta-analysis and its problems. BMJ 1994; 309: 789 – 792
128 *Senn S.* Testing for baseline balance in clinical trials. Statist Med 1994; 13: 1715 – 1726
129 *Windeler J, Antes G, Behrens J et al.* Kritische Evaluation ist ein Wesensmerkmal ärztlichen Handelns. Deutsches Ärzteblatt 2008; 105: A 565 – 570