1 Kontrollierte klinische Studien - eine Einführung

Martin Schumacher und Gabi Schulgen

Die Erforschung und Entwicklung neuer Therapien in der Medizin findet in der Regel nur in kleinen Schritten statt. Bahnbrechende Erfolge in der Entwicklung innovativer Behandlungen zur Heilung bislang inkurabler Erkrankungen waren in der Vergangenheit selten und sind auch in Zukunft kaum zu erwarten. Doch auch vergleichsweise kleine Effekte neuer Therapien können klinisch relevant sein und beträchtliche Auswirkungen auf das Wohlbefinden des individuellen Patienten haben. Um die Wirksamkeit und Verträglichkeit neuer Therapien zu belegen, ist ihre systematische Erprobung und Überprüfung in klinischen Studien erforderlich. Der erste Einsatz einer erfolgversprechenden medizinischen Behandlung am Menschen sollte daher als klinisches Experiment verstanden werden, mit dem Ziel, die Wirksamkeit der Therapie und ihre Verträglichkeit nachzuweisen.

Der traditionelle Wirksamkeitsnachweis bestand darin, den Behandlungserfolg neuer Substanzen mit den Ergebnissen zu vergleichen, die in einem vorausgegangenen Zeitraum mit herkömmlichen Verfahren erzielt wurden. Dieser sogenannte historische Vergleich hat jedoch vielfach nicht zu überzeugenden Erkenntnissen geführt. Bei der Behandlung der pulmonalen Tuberkulose wurden bis Mitte dieses Jahrhunderts viele - wie wir heute wissen - unwirksame Therapien über eine lange Zeit hinweg verabreicht. Die in Abbildung 1 dargestellte positive Entwicklung der Tuberkulosesterblichkeit, die mit einer Verbesserung der allgemeinen Lebensbedingungen einherging, wurde vielfach als Nachweis der Wirksamkeit neuer Behandlungen herangezogen (Silverman, 1985; McKeown, 1976). Obwohl der Erreger der Tuberkulose bereits 1882 identifiziert wurde, konnte erst im Jahre 1944 das erste Antibiotikum Streptomycin zur wirksamen Behandlung der pulmonalen Tuberkulose in den USA entwickelt werden. In dieser Zeit war die Tuberkulose die häufigste medizinische Todesursache bei jungen Erwachsenen in Europa und den USA.

Die begrenzte Verfügbarkeit dieses neuen Medikaments sowie der variable Verlauf der Erkrankung erhöhten die damaligen Anforderungen an einen Wirksamkeitsnachweis. Die Streptomycin-Studie des British Medical Research Council (MRC) zur Behandlung der pulmonalen Tuberkulose, die im Jahre 1947 durchgeführt wurde, war daher die erste randomisierte kontrollierte Studie, die weltweit durchgeführt wurde (Ederer, 1998; Sutherland, 1998). Aufgrund des variablen Verlaufs der Erkrankung wurde es als notwendig erachtet, zeitgleich eine Kontrollgruppe mitzuführen, die die Standardbehandlung (Bettruhe) erhielt. Nicht nur das Mitführen einer parallelen Kontrollgruppe war für die damalige Zeit revolu-

tionär; darüber hinaus wurde mit dieser Studie erstmalig die randomisierte, d.h. die zufällige Zuordnung der Patienten zur Therapie- und Kontrollgruppe eingesetzt. Die Randomisation war zuvor von Ronald A. Fisher in landwirtschaftlichen Versuchen Mitte der zwanziger Jahre eingeführt worden. Es war das besondere Verdienst von Sir Austin Bradford Hill, Mitglied des MRC, dieses Prinzip auch in die klinischen Versuche einzuführen und dort zu etablieren (Hill, 1951; Gail, 1996).

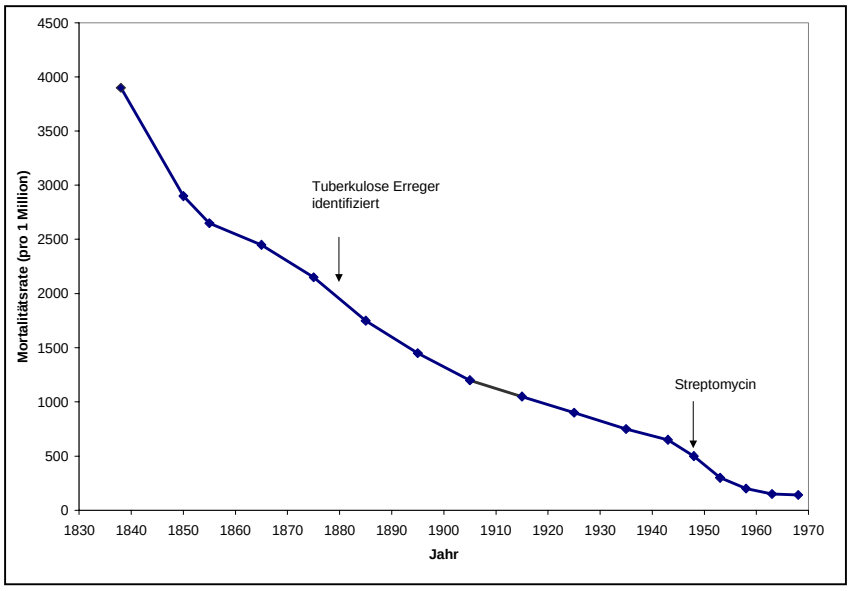

***Abbildung 1**: Entwicklung der Sterblichkeit verursacht durch die pulmonale Tuberkulose in England und Wales im Zeitraum von 1838 - 1978 (Silverman, 1985; McKeown, 1976).*

Wir werden im Folgenden anhand einer konkreten Studie, der Salk-Polio-Studie, die wesentlichen Punkte der Bedeutung klinischer Studien ansprechen. Die Salk-Polio-Studie ist zwar keine klinische Studie im engeren Sinne; sie ist eher als Präventionsstudie im öffentlichen Gesundheitswesen zu bezeichnen. Diese Studie ist jedoch in historischer Hinsicht interessant, da sie als größtes Experiment gilt, das je im öffentlichen Gesundheitswesen durchgeführt wurde, und sie ist darüber hinaus vom methodischen Standpunkt auch heute noch aktuell (Francis et al., 1955; Meier, 1985; Meier und Pringle Smith, 1998).

1.1 Die Salk-Polio-Studie

In den frühen fünfziger Jahren war in den USA die Frage zu klären, ob durch eine Impfung mit dem von Jonas Salk entwickelten Impfstoff eine Reduzierung der Inzidenz (Neuerkrankungsrate) der Poliomyelitis erreicht werden kann. Dazu standen verschiedene Vorgehensweisen zur Debatte:

Die einfachste Möglichkeit schien in der Durchführung eines historischen Vergleichs zu bestehen, d.h. im Jahre 1954 möglichst viele Kinder in den entsprechenden Altersgruppen (im Wesentlichen der Primary School) zu impfen und die Polio-Inzidenz dieses Jahres mit den Inzidenzen der Vorjahre zu vergleichen. Die Neuerkrankungsrate an Polio betrug zu dieser Zeit in den Vereinigten Staaten etwa 50 pro 100000, unterlag jedoch beträchtlichen jährlichen Schwankungen wie aus Abbildung 2 deutlich zu erkennen ist.

Man konnte also nicht ausschließen, dass aufgrund dieser Schwankungen die Inzidenz des Jahres 1954 auch bei Wirkungslosigkeit des Impfstoffes geringer als im Vorjahr sein würde. Das Auftreten einer Epidemie hätte andererseits die Inzidenz derart erhöhen können, dass eine tatsächliche Wirkung des Impfstoffes nicht erkannt worden wäre. Daher musste diese retrospektive Vorgehensweise des historischen Vergleichs a priori verworfen werden.

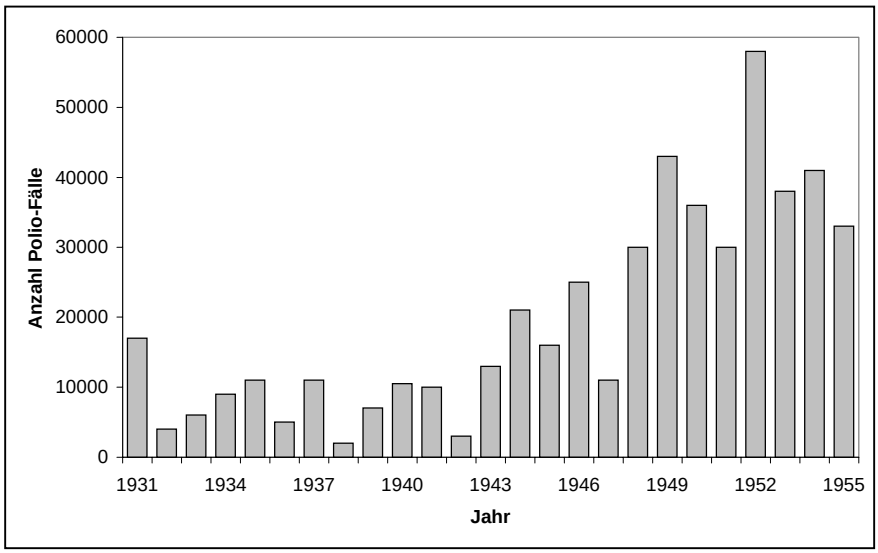

***Abbildung 2**: Anzahl Poliofälle in den USA während der Jahre 1931 bis 1955 (Francis et al., 1955).*

Als nächste Möglichkeit bot sich die Durchführung einer prospektiven Beobachtungsstudie an, bei der etwa den Eltern der Kinder in den entsprechenden Altersgruppen eine freiwillige Teilnahme an der Impfaktion angeboten würde. Die Inzidenz der Poliomyelitis in der Kohorte der geimpften Kinder hätte dann mit der Polio-Inzidenz in der Kohorte der nicht geimpften Kinder verglichen werden können. Bei diesem Vorgehen - so wurde befürchtet - würde die Zustimmung zur Teilnahme an der Impfaktion wesentlich vom sozio-ökonomischen Status und dem Gesundheitsbewusstsein der Eltern bestimmt sein, das wiederum, wie man wusste oder zumindest vermutete, auch einen Einfluss auf das Auftreten der Poliomyelitis hatte. Man musste also - wie beim historischen Vergleich - davon ausgehen, dass Impf- und Kontrollgruppe nicht vergleichbar sein würden.

Man entschied sich schließlich für eine randomisierte kontrollierte Studie, bei der die Kinder der Eltern, die ihre Einwilligung zur Teilnahme an der Studie gegeben hatten, der Impf- und der Kontrollgruppe randomisiert, d.h. zufällig zugeteilt wurden. Auf diese Weise sollten offenkundige und weniger offenkundige Selektionsmechanismen ausgeschlossen werden, die das Studienergebnis hätten verfälschen können. Darüber hinaus befürchtete man eine Verfälschung der Studienergebnisse durch mögliche Voreingenommenheit der untersuchenden Ärzte bei der Diagnose der Poliomyelitis bei geimpften und nicht-geimpften Kindern. Um auch diese Verzerrungsquelle auszuschließen, entschied man sich für die Durchführung einer doppel-blinden Studie, bei der weder die Kinder (bzw. deren Eltern) noch die impfenden und untersuchenden Ärzte wussten, ob mit dem Salk-Impfstoff oder aber nur mit Plazebo geimpft worden war. (Mit Plazebo wird ein Stoff bezeichnet, der wirkungslos - z.B. Kochsalzlösung - mit dem eigentlichen Wirkstoff jedoch äußerlich identisch ist).

Bei den über 400 000 Kindern, die an der randomisierten Studie teilnahmen, zeigte sich, dass die Inzidenz in der Gruppe der geimpften Kinder nur etwa halb so groß war wie in der Gruppe der unbehandelten Kinder: nur 82 Kinder der geimpften Gruppe gegenüber 162 Kinder der Kontrollgruppe erkrankten an Polio (Tabelle 1).

Tabelle 1: *Ergebnis der Salk-Polio-Studie: Anzahl der an Polio erkrankten und der gesunden Kinder in den Interventionsgruppen (Francis et al., 1955).*

Impfung	Polio	
	Ja	Nein
Ja	82	200 663
Nein	162	201 067

Die Stichprobenumfänge für diese Studie scheinen auf den ersten Blick immens hoch - wir werden in Kapitel 10 auf die Begründung eingehen. Sicher ist jedoch das überzeugende Ergebnis dieser Studie, die ohne jeden Zweifel einen Meilenstein in der Bekämpfung der Poliomyelitis darstellte, auch wenn schon einige Jahre später der Salk-Impfstoff durch neue, bessere Vakzine abgelöst wurde.

1.2 Die Problematik historischer Vergleiche

Am Beispiel der Salk-Polio-Studie haben wir bereits Gründe für die Durchführung randomisierter Studien angeführt. Hier wollen wir mögliche Verzerrungsquellen und deren Auswirkungen bei der Verwendung historischer Kontrollen in Therapiestudien näher untersuchen. Die scheinbaren Vorteile bei der Verwendung historischer Kontrollen liegen darin, dass die zeitgleiche Kontrollgruppe eingespart wird, so dass ein geringerer Stichprobenumfang benötigt wird, weniger Kosten entstehen und Patienten nicht einer möglicherweise schlechteren Kontrollbehandlung ausgesetzt werden müssen. Neben der bewussten und auch unbewussten Selektion von Patienten wird als wichtigstes Argument gegen historische Kontrollen meist der sogenannte Zeittrend angeführt. Green (1982) und später Dupont (1985) haben einen Aspekt dieses Zeittrends in einer Graphik dargestellt, die in etwas abgewandelter Form in Abbildung 3 wiedergegeben ist. Das Stadium der Krankheit ist dabei repräsentiert durch den vertikalen Abstand zwischen den beiden divergierenden Linien.

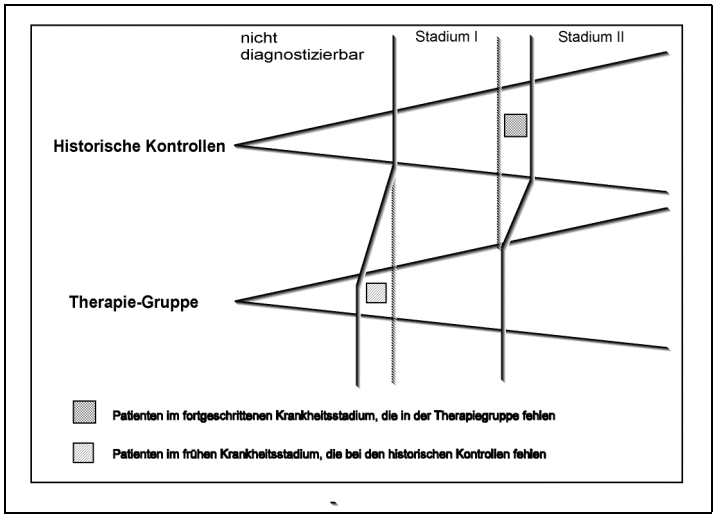

***Abbildung 3**: Schematische Darstellung der Auswirkungen des Zeittrends nach Dupont (1985).*

Ist die Definition der Krankheitsstadien einer bestimmten Erkrankung seit Jahrzehnten gleichgeblieben, könnte man zu dem Schluss gelangen, dass frühere Patienten in Stadium I oder II dieselbe Prognose wie heutige Patienten im gleichen Krankheitsstadium haben. Die Stadien bezeichnen dabei Abschnitte eines Fortschreitens der Krankheit. Entsprechend dem Fortschreiten der Krankheit wird die Prognose schlechter und das wahre Krankheitsstadium kann leichter erkannt werden. Die Möglichkeit, das wahre Krankheitsstadium zu erkennen, hat sich jedoch durch die Entwicklung neuer Diagnoseverfahren wesentlich verbessert. So kann eine früher nicht diagnostizierbare Krankheit heute bereits als Stadium I erkannt werden. Ebenso können fortgeschrittenere Krankheitsstadien früher diagnostiziert werden. Würde man daher heute eine klinische Studie an Patienten mit Stadium I durchführen und beispielsweise die Patienten von vor 10 Jahren als Kontrollen verwenden, so wird auch bei gleicher Wirkung der beiden Behandlungen die neue Behandlungsgruppe besser abschneiden als die historischen Kontrollen. Denn zur Therapiegruppe gehören nun Patienten des frühen Stadiums I mit sehr guter Prognose und es fehlen Patienten im späten Stadium I mit schlechter Prognose, die heute bereits als Stadium II diagnostiziert werden würden.

Das erfreuliche drastische Absinken der Brustkrebsmortalitätsraten seit Beginn der neunziger Jahre in England und Wales hat großes Aufsehen erregt und könnte als Indiz für die Auswirkung des Einsatzes verbesserter Therapien angesehen werden (Beral et al., 1995; Peto, 1998). Abbildung 4 zeigt die beobachtete Anzahl Todesfälle an Brustkrebs pro 100 000 Frauen von 1950 bis 1995 in England und Wales.

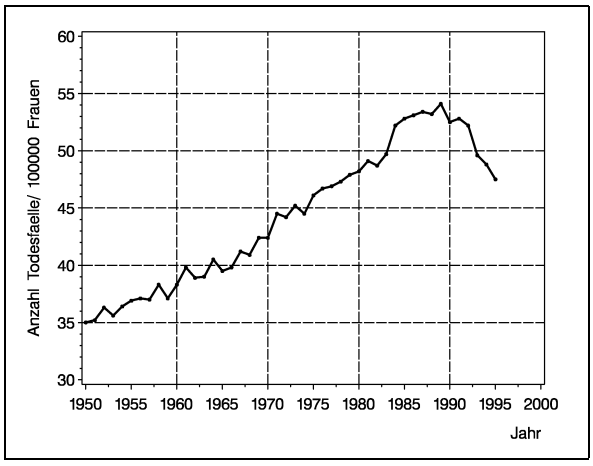

Abbildung 4: *Brustkrebsmortalität in England und Wales in den Jahren 1950 bis 1995 (Anzahl Todesfälle pro 100000 Frauen). Die Daten wurden der WHO Mortality Database in 1998 entnommen, die im Internet verfügbar ist (http://www.who.int/whosis).*

In den vergangenen Jahren konnte in einer Vielzahl klinischer Studien die Wirksamkeit neuer Therapien, insbesondere systemischer Therapien wie Tamoxifen, zur Behandlung des Brustkrebs nachgewiesen werden (Early Breast Cancer Trialists' Collaborative Group, 1992). Der historische Vergleich der Mortalitätsraten ist dennoch mit großer Vorsicht zu interpretieren, da auch andere Faktoren wie die Verbesserung der diagnostischen Möglichkeiten des Brustkrebs und die Durchführung regelmäßiger Vorsorgeuntersuchungen und somit der frühzeitige Einsatz operativer Therapien eine Rolle spielen können. Weiterhin können Veränderungen der Lebensbedingungen, das Absinken der Neuerkrankungsrate oder selbst Neuerungen in Definitionen bei der Erstellung von Todesursachenstatistiken einen Einfluss auf die ursachen-spezifischen Mortalitätsraten haben. Zwar spricht einiges dafür, dass durch die in klinischen Studien nachgewiesene Verbesserung der Therapiemöglichkeiten Todesfälle verhindert werden konnten und zu erwarten ist, dass sich diese Entwicklung auch in der Mortalitätsstatistik niederschlägt, dennoch ist das Ausmaß dieses Effektes nur unter großen Vorbehalten darin ablesbar.

Von welcher Größenordnung die Verzerrung bei der Verwendung historischer Kontrollen nämlich sein kann, zeigt eine Arbeit von Chalmers et al. (1977) über verschiedene klinische Studien zur Behandlung von Infarktpatienten mit Antikoagulantien, die im Zeitraum von 1948 bis 1975 veröffentlicht wurden. Die Ergebnisse dieser Untersuchung, die auch von Peto (1978) eingehend diskutiert wurden, sind zusammenfassend in Tabelle 2 dargestellt. In den randomisierten Studien ergibt sich eine deutliche Überlegenheit der Antikoagulantien ($P<0.01$). Für die Patienten, die mit Antikoagulantien behandelt wurden, ergab sich ein relatives Risiko von 0.80, d.h. 20% der Todesfälle, die ohne Behandlung eingetreten wären, konnten verhindert werden. Aufgrund der großen Stichprobenumfänge sind wir hier in der Lage, die Größenordnung der Verzerrung bei den Studien mit historischen Kontrollen abzuschätzen. Es stellt sich heraus, dass die Verzerrung so groß ist, dass sie selbst einen hypothetischen adversen Effekt der Antikoagulantien überdeckt und sogar in einem solchen Fall die Behandlung mit Antikoagulantien als die überlegene dargestellt hätte.

In einer Übersichtsarbeit über empirische Vergleiche randomisierter und nicht-randomisierter klinischer Studien kommen Kunz und Oxman (1998) zu dem Schluss, dass im allgemeinen nicht-randomisierte Studien den Effekt neuer Therapien überschätzen. Sie beobachteten, dass die Verzerrung jedoch prinzipiell in jede Richtung gehen kann; sie kann einen Effekt auch umkehren oder verschleiern. Zwei neuere Untersuchungen zum Vergleich randomisierter klinischer Studien mit nicht-randomisierten Beobachtungsstudien in verschiedenen therapeutischen Bereichen fanden keine Unterschiede zwischen den geschätzten Behandlungseffekten in den Beobachtungsstudien und den randomisierten kontrollierten Studien (Benson und Hartz, 2000; Concato et al., 2000).

***Tabelle** 2: Ergebnisse von 32 Studien zur Behandlung von Herzinfarktpatienten mit Antikoagulantien (Chalmers et al., 1977).*

Studientyp/ Kontrollgruppe	Anzahl Studien	Anzahl Kontrollen	Anzahl behandelte Patienten	Todesfälle Kontrollen	Todesfälle behandelte Patienten	Relatives Risiko (RR)
Studien mit historischen Kontrollen	18	4460	4194	1381 (31%)	640 (15%)	0.49
Studien mit externen Kontrollen	8	1627	1517	462 (28%)	308 (20%)	0.71
Kontrollierte randomisierte Studien	6	1748	2106	313 (18%)	301 (14%)	0.80

Diese Beobachtungen könnten darauf hindeuten, dass sich die Qualität nicht-randomisierter Studien im Laufe der Zeit verbessert hat und Beobachtungsstudien mit hoher Qualität durchaus zu ähnlichen Ergebnissen wie randomisierte Studien kommen können. Dennoch werden randomisierte klinische Studien, die sorgfältig durchgeführt wurden, das Standardinstument des Wirksamkeitsnachweises bleiben (Barton, 2000). Denn Beobachtungsstudien haben den wesentlichen Nachteil, dass ihr Design nicht experimentell ist. Die Behandlung jedes einzelnen Patienten wurde bewusst gewählt und nicht zufällig zugeteilt, so dass das Risiko niemals ausgeschlossen werden kann, dass systematische Unterschiede im Behandlungsergebnis auf andere Faktoren als die Behandlung zurückzuführen sind (Pocock und Elbourne, 2000), wie wir im nächsten Abschnitt näher ausführen werden.

1.3 Beobachtungsstudien und Registerdaten

Die immer größer werdende Flut von Registern, allgemeinen Dokumentationen und Datenbanken bringt den verständlichen Wunsch mit sich, diese Daten auch zu einem retrospektiven Therapievergleich einzusetzen und auf die kontrollierte prospektive Durchführung einer Therapiestudie zu verzichten. Nun besteht bei der Bewertung von Therapien, anders als etwa im epidemiologischen Bereich bei der Beurteilung der Auswirkungen von Risikofaktoren, die zusätzliche Schwierigkeit, dass die Folgen einer bewussten Handlung, mit der ja der Therapieerfolg hinsichtlich eines Kriteriums positiv beeinflusst werden sollte, beurteilt werden müssen. Bei dieser retrospektiven Vorgehensweise ist allerdings allenfalls der Spielraum des ärztlichen Ermessen nachträglich zu erkennen, nicht aber der Grund, weshalb ein Patient etwa mit einer hohen, ein anderer aber mit einer niedrigen Dosierung behandelt worden ist.

Wir wollen diese Problematik an dem Beispiel einer Studie verdeutlichen, die an der Universitätsfrauenklinik in Freiburg zur Evaluierung prognostischer Faktoren beim frühen Zervixkarzinom durchgeführt wurde (Pfisterer et al., 1996). Es konnten 212 nicht vorbehandelte Patientinnen mit der Diagnose eines Zervixkarzinoms im FIGO-Stadium IB und II, die im Zeitraum von 1982 bis 1989 in der Klinik behandelt worden waren, für die Studie berücksichtigt werden. Daten über den Verlauf der Erkrankung, die Behandlungsmodalitäten und potentiell wichtige prognostische Faktoren wie beispielsweise das Alter, Tumorgrading und den Ploidiestatus wurden retrospektiv den Krankenakten entnommen und waren in einer Datenbank verfügbar.

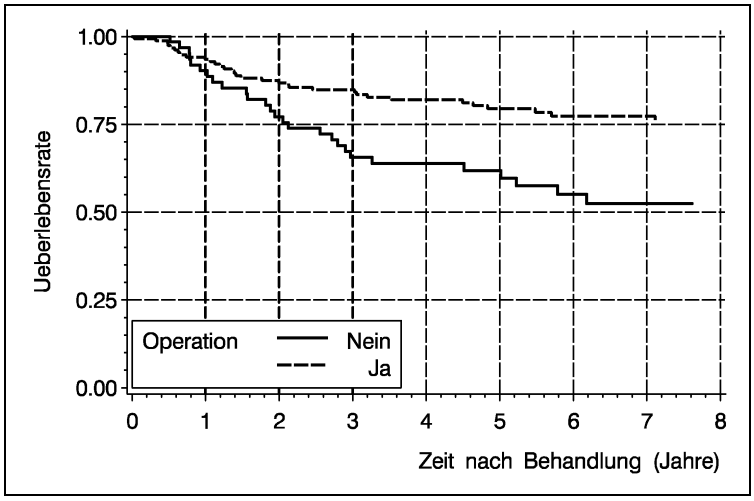

Abbildung 5: *Überlebensraten gemäß Operationsstatus von 212 Zervixkarzinom-Patientinnen im Stadium IB - II der Universitätsfrauenklinik Freiburg nach Behandlung in den Jahren 1982 – 1989 (151 operierte und 61 nicht-operierte Patientinnen).*

Abbildung 5 zeigt die Überlebensraten der 212 Patientinnen aufgeteilt in die Gruppe der operierten (N=151) und der nicht-operierten Patientinnen (N=61). Insgesamt waren bei einer medianen Beobachtungszeit von 5.3 Jahren nach Behandlung (1139 Personenjahre) 61 Todesfälle zu verzeichnen. Man erkennt, dass die operierten Patientinnen in diesem Kollektiv eine günstigere Prognose haben, im Vergleich zu den nicht- operierten Patientinnen.

Abbildung 6 untersucht den Einfluss der Strahlentherapie in diesem Kollektiv. Die Überlebensraten der Patientinnen, die bestrahlt wurden (N=150) sind deutlich geringer als die der Patientinnen, die keine Bestrahlung erhielten (N=62).

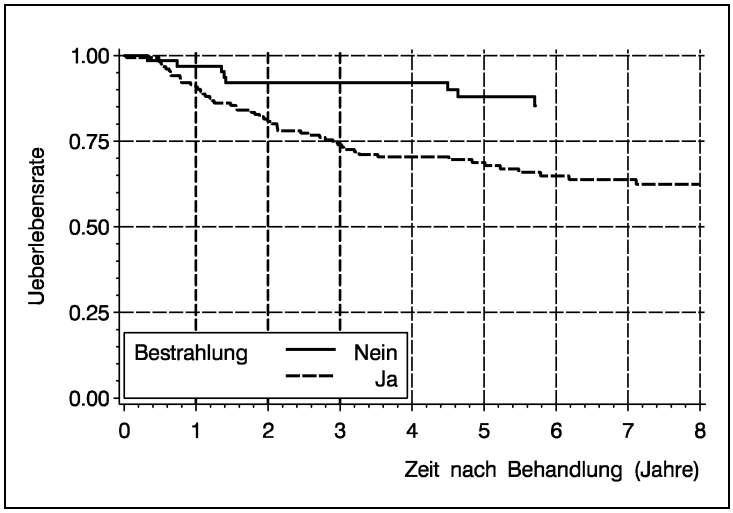

Abbildung 6: *Überlebensraten gemäß Bestrahlungsstatus von 212 Zervixkarzinom-Patientinnen im Stadium IB - II der Universitätsfrauenklinik Freiburg nach Behandlung in den Jahren 1982 – 1989 (150 bestrahlte und 62 nicht-bestrahlte Patientinnen).*

Ähnliche Ergebnisse erhält man, wenn man statt der Überlebenszeit die rezidivfreie Überlebenszeit betrachtet. Auch die Berücksichtigung prognostischer Faktoren wie Alter, Stadium und Grading ändert nichts an den beobachteten Effekten. Kann man daher aus diesen Ergebnissen eine Therapieempfehlung ableiten, etwa in dem Sinne, dass auf eine Bestrahlung verzichtet werden und stattdessen der Operation der Vorzug gegeben werden sollte? Diese Frage ist mit einem klaren NEIN zu beantworten. Auch wenn man für einige wenige bekannte einflussreiche Faktoren adjustieren kann, wird es eine Vielzahl unbekannter Faktoren mit großem Einfluss auf die Prognose geben, die in die Therapieentscheidung des behandelnden Arztes eingeflossen sind. In dem hier betrachteten Kollektiv der Patientinnen mit frühem Zervixkarzinom erhielten vielfach Patientinnen, die bei Diagnose in einem nicht mehr operablen, prognostisch ungünstigen Zustand waren, die Strahlentherapie. Daher sind die beobachteten Effekte der Operation und Bestrahlung nicht der jeweiligen verabreichten Therapie, sondern der Auswahl eines prognostisch günstigen bzw. ungünstigen Kollektivs zuzuschreiben.

Das Phänomen, dass sich die Überlegenheit einer Therapie über eine andere bei der Analyse von Registerdaten und Beobachtungsstudien sogar umkehren kann, wird in der Literatur als Simpson's Paradoxon bezeichnet. Diese Bezeichnungsweise entstand aufgrund einer Veröffentlichung von Simpson (1951). Um dieses Phänomen zu verdeutlichen, verwenden wir zunächst ein hypothetisches Beispiel (Green und Byar, 1984): Bei einem Vergleich zweier Therapien (A und B) zeigt sich sowohl bei männlichen als auch bei weiblichen Patienten die Überlegenheit

der Behandlung B, was in den Risikoverhältnissen von 1.38 bzw. 2.50 für A zu B in Tabelle 3 zum Ausdruck kommt. Ignoriert man jedoch das Geschlecht der Patienten und fasst alle Daten in einer einzigen Vierfeldertafel zusammen so ergibt sich ein Risikoverhältnis von 0.80, das klar für eine Überlegenheit von A spricht.

Dieses Paradoxon findet seine Erklärung in der Unbalanciertheit der Randhäufigkeiten der beiden Kontingenztafeln: In unserem Beispiel haben die Männer eine schlechtere Prognose als die Frauen; die meisten männlichen Patienten in dieser Studie erhielten aber B. Die Frauen haben eine weitaus bessere Prognose, und die Majorität der Frauen erhielt A. Damit ist klar, dass das Geschlecht der Patienten ein bedeutender vermengender Faktor war, der das auf den ersten Blick erstaunliche entgegengesetzte Gesamtresultat der Studie erklärt.

***Tabelle 3**: Hypothetisches Beispiel für das Simpsonsche Paradoxon nach Green und Byar (1984).*

	Behandlung	**gestorben**	**überlebt**	**gesamt**
Männliche	A	20 (50%)	20	40
Patienten	B	40 (36%)	70	110
		60 (40%)	90	150
		Relatives Risiko: 20:40 / 40:110 = 1.38		
Weibliche	A	20 (10%)	180	200
Patienten	B	4 (4%)	96	100
		24 (8%)	276	300
		Relatives Risiko: 20:200 / 4:100 = 2.50		
Alle	A	40 (17%)	200	240
Patienten	B	44 (21%)	166	210
		84 (19%)	366	450
		Relatives Risiko: 40:240 / 44:210 = 0.80		

Man mag einwenden, dass ein solch extremes Resultat der Umkehrung des Therapieeffektes nur konstruiert ist und in Wirklichkeit nicht auftreten wird. Wie leicht jedoch ein solcher Effekt in Beobachtungsstudien möglich ist und scheinbar unsin-

nige Ergebnisse produzieren kann, zeigt ein reales Beispiel einer epidemiologischen Studie (Appleton et al., 1996).

In dieser Kohortenstudie zu Schilddrüsen- und Herzerkrankungen wurde der Überlebensstatus aller Teilnehmer 20 Jahre nach einer Basisuntersuchung Anfang der siebziger Jahre erhoben, bei der unter anderem die Rauchgewohnheiten erfasst worden waren. Tabelle 4 zeigt im oberen Teil den Überlebensstatus von 1314 Frauen, die gemäß der Basisuntersuchung als Raucherinnen bzw. Nichtraucherinnen eingestuft wurden. Bei den Nichtraucherinnen betrug die Mortalitätsrate 31% während nur 24% der Raucherinnen 20 Jahre nach der Basisuntersuchung verstorben waren, was einem relativen Risiko der Raucherinnen gegenüber den Nichtraucherinnen von 0.76 mit einem 95%-Konfidenzintervall von (0.64, 0.91) entspricht, also einer Reduktion der Mortalitätsrate um ca. 25% durch das Rauchen! Welche Erklärung gibt es für dieses verblüffende und wenig glaubwürdige Ergebnis? Es lässt sich erklären durch die Vernachlässigung einer wichtigen Variablen, nämlich dem Alter der Frauen zur Zeit der Basisuntersuchung, das sowohl mit den Rauchgewohnheiten als auch mit dem Mortalitätsrisiko zusammenhängt. Eine solche Variable, die sowohl mit der Exposition als auch mit der Erkrankungshäufigkeit zusammenhängt, bezeichnet man in der Epidemiologie als Confounder.

Der untere Teil von Tabelle 4 zeigt, dass der Anteil der Raucherinnen in der Altersgruppe über 65 deutlich geringer ist als in den anderen Gruppen, das Mortalitätsrisiko jedoch natürlicherweise mit dem Alter steigt. Mit Ausnahme der höchsten Altersgruppe, ist das relative Risiko zu versterben in jeder Altersgruppe für Raucherinnen erhöht, was auf einen gesundheitsgefährdenden Effekt des Rauchens hinweist. Eine geeignete gewichtete Zusammenfassung der 4 altersspezifischen relativen Risiken (vgl Kapitel 4.7 und 8.2) ergibt ein relatives Risiko von 1.21 mit einem 95%-Konfidenzintervall von [1.03, 1.41]. Wie schon in dem hypothetischen Beispiel führt die Berücksichtigung einer einflussreichen Kovariablen zu einer Umkehrung des Ergebnisses. Weiter Beispiele für das Simpsonsche Paradoxon geben Reintjes et al. (2000) mit den Daten einer multizentrischen Studie zu nosokomialen Infektionen sowie Julious und Mullee (1994) aus verschiedenen Bereichen der medizinischen Forschung.

Bei der Analyse von Beobachtungsstudien und Registerdaten mit dem Ziel eines Therapievergleichs liegt allerdings selten ein so offenkundiger Faktor vor, wie es in unseren obigen Beispielen das Geschlecht der Patienten bzw. das Alter der Frauen gewesen ist. Vielmehr werden solche Faktoren wesentlich subtiler und im Rahmen solcher Studien meistens auch nicht dokumentiert sein, so dass eine Überprüfung unmöglich ist (Byar, 1980; Dambrosia und Ellenberg, 1980).

Tabelle 4: *Reales Beispiel für das Simpsonsche Paradoxon nach Appleton et al. (1996). 20-Jahres Überlebensraten von 1314 Frauen gemäß Raucherstatus, insgesamt und in vier Altersgruppen. Relative Risiken mit 95%-Konfidenzintervallen sind pro Vierfeldertafel angegeben.*

	Rauchen	**gestorben**	**überlebt**	**Gesamt**
alle Frauen	ja	139 (24%)	443	582 (44%)
	nein	230 (31%)	502	732 (56%)
		369 (28%)	945	1314 (100%)
		Relatives Risiko: 0.76 (0.64 - 0.91)		
Alter unter 45	ja	19 (7%)	269	288 (46%)
	nein	13 (4%)	327	340 (54%)
		32 (5%)	596	628 (100%)
		Relatives Risiko: 1.73 (0.88 - 3.40)		
Alter 45 – 54	ja	27 (21%)	103	130 (62%)
	nein	12 (15%)	66	78 (38%)
		39 (19%)	169	208 (100%)
		Relatives Risiko: 1.35 (0.73 - 2.49)		
Alter 55 – 64	ja	51 (44%)	64	115 (49%)
	nein	40 (33%)	81	121 (51%)
		91 (39%)	145	236 (100%)
		Relatives Risiko: 1.34 (0.97 - 1.86)		
Alter über 65	ja	42 (86%)	7	49 (20%)
	nein	165 (85%)	28	193 (80%)
		207 (86%)	35	242 (100%)
		Relatives Risiko: 1.00 (0.88 - 1.14)		

Für die Zulassung neuer Therapien akzeptieren die zuständigen Behörden in aller Regel nur Wirksamkeitsnachweise basierend auf randomisierten kontrollierten Studien. In anderen Bereichen wie etwa der Überwachung der Arzneimittelsicherheit nach Zulassung oder der Evaluation von Risikofaktoren für das Entstehen von Krankheiten sind wir auf die Ergebnisse von Beobachtungsstudien und Auswertungen von Registerdaten angewiesen. In seltenen Fällen, in denen die zu vergleichenden Therapien sehr unterschiedlich sind, wie etwa der Vergleich einer operativen mit einer medikamentösen Therapie, können randomisierte Studien an der

Teilnahmeverweigerung der Patienten scheitern. Beobachtungsstudien sind deshalb trotz vieler damit verbundener Probleme unverzichtbar in den Bereichen, in denen der Einsatz der Randomisation nicht möglich ist.

1.4 Randomisierte klinische Studien

Ein Experiment - zumindest im engeren Bereich der Naturwissenschaften - ist dadurch gekennzeichnet, dass durch Variation der Einfluss eines Faktors auf ein zu untersuchendes Kriterium bestimmt wird. Beobachtete Unterschiede hinsichtlich dieses Kriteriums dem Einflussfaktor zuzuschreiben, ist jedoch nur dann möglich, wenn alle anderen Einflussgrößen während des Experiments fixiert werden können oder die Versuchseinheiten in allen Aspekten identisch sind.

Übertragen auf klinische Studien, in denen der Einflussfaktor die Behandlung und die Versuchseinheiten Patienten sind, ist es klar, dass die Forderung nach Fixierung aller anderen Einflussgrößen oder gar identischen Versuchseinheiten im Rahmen klinischer Studien unerfüllbar ist. Patienten variieren auch bei beschränktem Indikationsgebiet und restriktiven Ein- und Ausschlusskriterien für eine Studie in so vielen anderen bekannten und unbekannten Faktoren, dass der Versuch, diese alle bei der Auswertung berücksichtigen zu wollen, von vornherein zum Scheitern verurteilt wäre.

Um diesem Dilemma zu entgehen, wird seit den vierziger und fünfziger Jahren dieses Jahrhunderts - im Bereich der Medizin besonders gefördert durch Sir Austin Bradford Hill (1951, 1962) - die Randomisation, d.h. die zufällige Zuteilung der Patienten zu den Therapiegruppen, weithin akzeptiert und bei Therapiestudien eingesetzt (Doll, 1992). Unter randomisierter Therapiezuteilung verstehen wir, dass jeder Patient, der in die Studie eingeschlossen wird, eine vorgegebene, bekannte Wahrscheinlichkeit hat, jede der Behandlungen zu erhalten, die Behandlungszuteilung aber nicht vorhergesagt werden kann (Altman und Bland, 1999). Im einfachsten Fall zweier gleich großer Behandlungsgruppen könnte beispielsweise ein Münzwurf über die Therapiezuteilung entscheiden. Häufig ist es jedoch vorteilhaft, aufwendigere Zufallsmechanismen zu verwenden, wie wir in Kapitel 11 näher ausführen werden. Dort wird auch auf die Problematik der Geheimhaltung der Randomisation und der Verblindung der Therapien näher eingegangen.

Die Randomisation bietet den äußerst wichtigen Vorteil, dass die Gefahr einer Verzerrung der Ergebnisse durch Selektion, d.h. durch bewusste und auch unbewusste systematische Zuordnung von Patienten mit besonders guter oder schlechter Prognose zu den einzelnen Behandlungen, ausgeschlossen wird. Zwar bietet auch die Randomisation keinen sicheren Schutz vor Unbalanciertheit bezüglich wichtiger Einflussfaktoren in den Gruppen, jedoch ist die Wahrscheinlichkeit für ein solches Ungleichgewicht äußerst gering. Darüber hinaus gewährleistet die zu-

fällige Zuordnung der Patienten zu den Behandlungsgruppen die verzerrungsfreie Schätzung des Behandlungsunterschieds und die Validität statistischer Tests bei der Auswertung der Studie. Die Randomisation ermöglicht es, einen beobachteten Effekt auch tatsächlich der Therapie zuschreiben zu können. Zusätzlich ist aber auch zu bedenken, dass die Randomisation oft erst den Einsatz blinder Techniken ermöglicht. Im Beispiel der Salk-Polio-Studie ging es um die vorurteilsfreie Diagnose der Poliomyelitis. Wie wichtig dieser Punkt bei der Verwendung von subjektiv zu bewertenden Kriterien bei der Beurteilung von Therapien sein kann, ist klar, wenn man sich Kriterien wie Ansprechen von Tumoren auf eine Behandlung, die Befindlichkeit und verschiedene Aspekte der Lebensqualität von Patienten oder subjektive Einschätzung von Verbesserung und Verschlechterung des Krankheitszustandes vor Augen hält.

Trotz der vielen Vorteile, die die zufällige Zuteilung der Therapie in klinischen Versuchen bietet, verlief die Einführung der Randomisation im medizinischen Bereich nicht reibungslos und noch heute wird ihre Notwendigkeit und Einsetzbarkeit immer wieder in Frage gestellt. Gründe dafür sind zum Teil darin zu sehen, dass klinische Forscher nicht ausreichend in den Prinzipien klinischer Experimente ausgebildet werden, zum anderen in ethischen Bedenken gegenüber Experimenten am Menschen. Die Hürden, die es bei der Einführung der adäquaten Methodik klinischer Studien zu überwinden galt, werden deutlich, wenn man eine sehr eindrucksvolle, vielzitierte Argumentation von Jerome Cornfield liest, einem weiteren Pionier auf dem Gebiet der randomisierten klinischen Studien (Ederer, 1982; Gail, 1996): Im Rahmen der Planung klinischer Studien zum Vergleich zweier Bestrahlungsmodalitäten wurde seitens eines Radiologen der Vorschlag gemacht, alle geeigneten Patienten einer Klinik mit der neuen Therapie und alle geeigneten Patienten einer anderen Klinik mit der herkömmlichen Therapie zu behandeln. Die Randomisation individueller Patienten wurde als zu aufwendig betrachtet. Cornfield begegnete diesem Vorschlag, indem er als Beispiel den hypothetischen Vergleich zweier Medikamente gegen Seekrankheit anführte. In einer solchen Studie verabreichte man der Besatzung eines Bootes Tabletten der Sorte A, der Besatzung eines andern Bootes Tabletten der Sorte B. Im Verlauf dieser Studie stellte sich dann jedoch heraus, dass das eine Boot schwerer beladen war und dadurch mehr Turbulenzen auftraten. Bei der Besatzung dieses Bootes traten mehr Fälle von Seekrankheit auf. Es war infolge der unterschiedlichen Beladung unmöglich herauszufinden, ob die Unterschiede in der Krankheitshäufigkeit auf das Medikament oder den Ballast zurückzuführen war. Ein unverzerrter Vergleich der beiden Medikamente war dadurch nicht mehr möglich. Cornfields Argumente überzeugten die Radiologen, trotz des höheren Aufwands Studien durchzuführen, bei der die Patienten und nicht die Kliniken den Therapien randomisiert zugeteilt wurden.

1.5 Interne und externe Validität

Die interne Validität des Behandlungsvergleichs in einer klinischen Studie ist durch die Randomisation gegeben. Die externe Validität, d.h. die Übertragbarkeit der Studienresultate auf andere Patienten, ist nicht automatisch durch die Studie gesichert. Die Patienten einer klinischen Studie sind keine Zufallsstichprobe aus der „Grundgesamtheit" aller möglichen Patienten mit der untersuchten Erkrankung. In der Regel wird eine klinische Studie an einer oder mehreren Kliniken (Zentren) durchgeführt, die nicht zufällig ausgewählt werden. Die Patienten, die im Studienzeitraum in dem teilnehmenden Zentrum behandelt werden, die die Einschlusskriterien der Studie erfüllen und ihre Zustimmung geben, werden in die Studie eingeschlossen. Durch die Ein- und Ausschlusskriterien der Studie, die im Studienprotokoll festgelegt sind, wird die Zielpopulation festgelegt, auf die die Studienergebnisse verallgemeinert werden sollen. Die Beschreibung der Zusammensetzung der Studienpopulation anhand der erhobenen Patientencharakteristika bei Einschluss in die Studie gibt weiteren Aufschluss hinsichtlich der Übertragbarkeit der Ergebnisse.

In bestimmten Situationen ist das sogenannte „Comprehensive Cohort Design" (Scheurlen et al., 1984; Olschewski und Scheurlen, 1985) hilfreich, um die externe Validität der Studienergebnisse zu untersuchen. Dieses Design wurde beispielsweise in der Coronary Artery Surgery Study (CASS) zum Vergleich der Bypass-Operation mit einer konventionellen medikamentösen Therapie eingesetzt (Olschewski et al., 1992). Bei diesem Design werden alle für die Studie geeigneten Patienten in eine prospektive Kohortenstudie eingeschlossen, in der die Subkohorte der Patienten, die der Randomisation zustimmen, für den eigentlichen Behandlungsvergleich zur Verfügung stehen. Patienten, die der Randomisation nicht zustimmen, entscheiden sich für die eine oder andere Therapie im Rahmen der Studie. Ein Vergleich der randomisierten mit den nicht-randomisierten Patienten hinsichtlich der Zusammensetzung der Kollektive sowie des Behandlungserfolges gibt Aufschluss über die Übertragbarkeit der Ergebnisse des Therapievergleichs. Als Alternative für diese aufwändige Vorgehensweise wird häufig vorgeschlagen, zunächst eine Liste mit den wichtigsten Basisdaten der Patienten zu führen, die zwar die Ein- und Ausschlusskriterien der Studie erfüllen, aber aus irgendeinem Grund, z.B. fehlende Zustimmung zur Randomisation, nicht in die Studie aufgenommen wurden (Schmoor et al., 1996).

1.6 Entwicklungsstadien medizinischer Behandlungen

Nachdem eine neue Substanz die *prä-klinischen Phasen* der Labor- und Tierversuche erfolgreich durchlaufen hat, beginnen die klinischen Versuche am Menschen bei denen man in der Regel vier Phasen unterscheidet. In der *Phase I* der klinischen Studien wird die neue Substanz erstmalig an Menschen eingesetzt. Der Ein-

satz der Behandlung in dieser Phase hat keine therapeutischen Ziele und wird häufig an gesunden Freiwilligen durchgeführt. Das wesentliche Ziel dieser Studien ist es, Informationen über pharmakokinetische und –dynamische Eigenschaften der Substanz zu gewinnen (vgl. Kapitel 16). Darüber hinaus erhofft man sich erste Daten über Sicherheit und Verträglichkeit. In der *Phase II* ist das primäre Ziel, erste Hinweise auf die Wirksamkeit einer neuen Substanz bei Patienten zu erlangen. Phase I und II Studien dienen wesentlich dazu, Dosis und Darreichungsform der neuen Behandlung für die folgenden Phase III Studien zu bestimmen. *Phase III* Studien dienen dem Wirksamkeitsnachweis einer neuen Behandlung. Sie sind die Basis für einen formalen Zulassungsantrag pharmazeutischer Produkte bei den entsprechenden Behörden. Diese Studien sind in aller Regel randomisierte Studien. Die *Phase IV* Studien werden nach Zulassung eines Medikamentes durchgeführt. Sie dienen der Überwachung der Arzneimittelsicherheit und liefern zusätzliche Daten zur Wirksamkeit eines Medikamentes. Eine gute Übersicht über die verschiedenen Stadien der Medikamentenentwicklung bietet die Richtlinie E8 der International Conference on Harmonisation (ICH) (vgl. Kapitel 15).

1.7 Literatur

Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? *British Medical Journal* 1999; 318: 1209.

Appleton DR, French JM, Vanderpump MPJ. Ignoring a covariate: an example of Simpson's paradox. *The American Statistician* 1996; 50: 340-341.

Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *British Medical Journal* 2000; 321: 255-256.

Benson K, Hartz A. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine* 2000; 342: 1878-1886.

Beral V, Hermon C, Reeves G, Peto R. Sudden fall in breast cancer death rates in England and Wales. *Lancet* 1995; 345: 1642-1643.

Byar DP. Why data bases should not replace randomized clinical trials. *Biometrics* 1980; 36: 337-342.

Berkson J, Harrington SW, Clagett OT et al. Mortality and survival in surgically treated cancer patients of the breast. Proceedings of the Staff Meeting of the Mayo Clinic 1957; 32: 645-670.

Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence of favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New England Journal of Medicine* 1977; 297: 1091-1096.

Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine* 2000; 342: 1887-1892.

Dambrosia JM, Ellenberg JH. Statistical considerations for medical data base. *Biometrics* 1980; 36: 323-332.

Doll R. Sir Austin Bradford Hill and the progress of medical science. *British Medical Journal* 1992; 305: 1521-1526.

Dupont WD. Randomized vs. historical clinical trials: are the benefits worth the costs? *American Journal of Epidemiology* 1985; 122: 940-946.

Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet* 1992; 339: 1-15, 71-85.

Ederer F. Jerome Cornfield's contributions to the conduct of clinical trials. *Biometrics* 1982; 38 (Supplement), 25-32.

Ederer F. History of clinical trials. In: Armitage P, Colton T (eds). Encyclopedia of Biostatistics. Chichester: Wiley, 1998.

Edwards MJ, Gamel JW, Feuer EJ. Improvement in the prognosis of breast cancer from 1965 to 1984. *Journal of Clinical Oncology* 1998; 16: 1030-1035.

Francis TH Jr et al. An evaluation of the 1954 Poliomyelitis vaccine trials - Summary report. *American Journal of Public Health* 1955; 45: 1-63.

Gail MH. Statistics in action. *Journal of the American Statistical Association* 1996; 91: 1-13.

Green SB. Patient heterogeneity and the need for randomized clinical trials. *Controlled Clinical Trials* 1982; 3: 189-198.

Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Statistics in Medicine* 1984; 3: 361-370.

Hill AB. The clinical trial. *British Medical Bulletin* 1951; 7: 278-282.

Hill AB. Statistical methods in clinical and preventive medicine. Edinburgh: Livingstone, 1962.

ICH E8. General considerations for clinical trials. London, UK: International Conference on Harmonisation; 1997. Adopted by CPMP September 1997 (CPMP/ICH/291/95).

Julious SA, Mullee MA. Confounding and Simpson's paradox. *British Medical Journal* 1994; 309: 1480-1481.

Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal* 1998; 317: 1185-1190.

McKeown T. The modern rise of population. London: Edward Arnold, 1976.

Meier P. The biggest public health experiment ever: the 1954 field trial of the Salk Poliomyelitis vaccine. In: Tanur JM, Mosteller F, Kruskal WH, Lehmann EL, Link RF, Pieters RS, Rising GR (eds). Statistics: A guide to the unknown. Monterey: Wadsworth & Brooks, 1989.

Meier P, Pringle Smith R. Salk Vaccine. In: Armitage P, Colton T (eds). Encyclopedia of Biostatistics. Chichester: Wiley, 1998.

Olschewski M, Scheurlen H. Comprehensive cohort study: an alternative to randomized consent design in a breast preservation trial. *Methods of Information in Medicine* 1985; 24: 131-134.

Olschewski M, Schumacher M, Davis KB. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Controlled Clinical Trials* 1992; 13:226-239.

Peto R. Clinical trial methodology. *Biomedicine Special Issue* 1978; 28: 24-36.

Peto R. Mortality from breast cancer in UK has decreased suddenly. *British Medical Journal* 1998; 317: 476-477.

Pfisterer J, Kommoss F, Sauerbrei W, Baranski B, Kiechle M, Ikenberg H. DNA flow cytometry in stage IB and II cervical carcinoma. *International Journal of Gynecological Cancer* 1996, 6: 54-60.

Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *New England Journal of Medicine* 2000; 342: 1907-1909.

Reintjes R, de Boer A, van Pelt W, Mintjes-de Groot J. Simpson's paradox: an example from hospital epidemiology. *Epidemiology* 2000; 11: 81-83.

Scheurlen H, Olschewski M, Leibbrand D. Zur Methodologie kontrollierter klinischer Studien über die Primärbehandlung des operablen Mammakarzinoms. *Strahlentherapie* 1984; 160: 459-468.

Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statistics in Medicine* 1996; 15: 263-271.

Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society* 1951; B-13: 238-241.

Silverman WA. Human experimentation. A guided step into the unknown. Oxford: Oxford University Press, 1985.

Sutherland I. Medical Research Council Streptomycin trial. In: Armitage P, Colton T (eds). Encyclopedia of Biostatistics. Chichester: Wiley, 1998.