Die statistische Bewertung prophylaktischer Methoden

P. J. Vitek
Computer Centre
Royal Postgraduate Medical School
Hammersmith Hospital
Ducane Road
GB London W 12

Uns wurden heute bereits verschiedene Behandlungsmethoden dargestellt. In meinem Beitrag möchte ich Ihre Aufmerksamkeit auf statistische Begriffe und Prinzipien lenken, die in der Beurteilung von Prüfungen und Studien zum Tragen kommen.

Bevor beschrieben wird, wie ein klinischer Versuch und die Durchführung und Analyse der Ergebnisse anzulegen ist, soll zuerst die Definition von Schwartz und Lellouch (1) hinsichtlich zweier Arten klinischer Versuche angeführt werden. Eine neuartige Behandlung muß, sobald die Sicherheit am Menschen erwiesen ist, dahingehend überprüft werden, ob sie tatsächlich die behauptete Wirkung besitzt. Diese Prüfung ist eine unmittelbare Fortführung der Laborexperimente und soll die vorangegangenen Befunde stützen. Eine zweite Art klinischer Prüfungen, die ich als pragmatisch bezeichne, soll in erster Linie den praktischen Wert einer neuen Therapie gegenüber anderen bereits eingeführten Behandlungsformen beurteilen. Die dabei gewonnenen Resultate dienen als Empfehlung für therapeutische Entscheidungen in der klinischen Praxis.

Der Grund für die Unterscheidung dieser beiden Vorgangsweisen liegt nicht nur darin, daß die Zielsetzungen unterschiedlich sind. Es geht dabei auch darum, daß verschiedene Aspekte der Durchführung, der Auswertung und der Interpretation von Ergebnissen jeweils davon abhängen, welche dieser Vorgangsweisen gewählt wurde. Bei der pragmatischen Vorgangsweise ist eine exakte Beschreibung und Auswahl des Patientenguts sowie der Behandlungsformen besonders wichtig, da die Befunde für die praktische Anwendung herangezogen werden sollen.

Es muß dem Kliniker, der die Entscheidung zu treffen hat, bewußt sein, welche Charakteristika des Patientenguts wichtig sind, um sinnvolle Schlußfolgerungen ziehen zu können. In diesem Zusammenhang ist es vielleicht sinnvoll festzuhalten, daß eine Arzneimittelprüfung immer im Hinblick auf eine bestimmte Dosierung und Verabreichungsform erfolgt und daß die daraus abgeleiteten Empfehlungen jeweils nur für Gruppen von Patienten gelten, die sich nicht wesentlich vom Patientengut der Prüfung unterscheiden.

Einer der häufigsten Gründe für nicht aussagekräftige Prüfungen liegt darin, daß die Prüfer sich nicht von vornherein darüber im klaren sind, wie groß eine Studie

angelegt und wie lang der Beobachtungszeitraum sein muß. Manchmal ist man sich auch nicht darüber im klaren, daß relativ große Unterschiede zwischen Behandlungen in einer klinischen Prüfung gar nicht darstellbar sein können, weil die randomisierten Unterschiede zwischen den Patientengruppen häufig viel größer sind, als ursprünglich erwartet wurde. Die Unterscheidbarkeit zweier klinischer Prüfungen hängt von verschiedenen Faktoren ab. Am wichtigsten sind die Fallzahl und die Wirksamkeit der zu prüfenden Therapieform. Je kleiner der Unterschied zwischen den Behandlungen ist, desto größer muß die Fallzahl sein, die benötigt wird. Das bedeutet natürlich entsprechend hohe Kosten. Bei Prüfungen, in denen es darum geht, den zeitlichen Ablauf bis zu einem bestimmten Ereignis zu vergleichen (z. B. der Nachweis der ersten Thrombose), hängt die Unterscheidungsfähigkeit der Prüfung von der Anzahl der Patienten ab, die von dem entsprechenden Ereignis betroffen sind, weniger von der Zahl der Prüfungsteilnehmer. Darauf wurde bereits von Peto et al. (2) hingewiesen. Sie haben eine Tabelle erstellt, die den Zusammenhang zwischen der Aussagefähigkeit von klinischen Prüfungen und einer bestimmten Anzahl von Ereignissen herstellt, die bei randomisierten Patienten in zwei Behandlungskollektiven zu gleichen Teilen auftreten. Wenn die erwartete Wirkung einer Therapie 2:3 beträgt, werden bereits mehr als 100 Ereignisse benötigt. Im Patientengut, das in eine Prüfung einbezogen wird, kommt natürlich ein bestimmtes Ereignis nicht in jedem Fall zum Tragen.

Das bedeutet, daß die Anzahl der Patienten, die zu einer Untersuchung herangezogen werden, wesentlich größer sein muß; um wieviel größer hängt von der Häufigkeit des zu erwartenden Ereignisses ab. Eine Lösung des Problems hinsichtlich der Patientenzahlen, die erforderlich sind, um einen Unterschied zwischen Therapieformen signifikant nachweisen zu können, wurde von Schwartz et al.(3) für verschiedene Formen der klinischen Prüfung entwickelt.

Um systematische Unterschiede zwischen Therapiegruppen auszugleichen, die Befunde entwerten können, müssen die Patienten den verschiedenen Kollektiven randomisiert zugeführt werden. Die Formulierung eines Protokolls muß relativ detaillierte Kriterien enthalten. Darunter fallen auch ethische Kriterien, die die Aufnahme in eine klinische Prüfung bestimmen. Patienten, die nicht in Frage kommen, dürfen von vornherein nicht berücksichtigt werden. Das ist besonders wichtig, denn ein Ausscheiden von Probanden aus der Untersuchung nach der Randomisierung führt zu einer Verzerrung der Ergebnisse. Manchmal müssen vor der Randomisierung Patientenkollektive geschichtet werden, um Verzerrungen, die durch prognostische Variablen eingebracht wurden, zu entfernen. Damit wird die eigentliche Zuordnung relativ kompliziert. Das ist aber nicht zwingend, denn es gibt statistische Methoden, mit denen der Verteilung prognostischer Variablen Rechnung getragen werden kann. Man spricht hier in den meisten Fällen von einer retrospektiven Schichtung. Es handelt sich dabei um eine rein statistische Methode, die auf vorliegende Daten angewendet werden kann. Der Nachteil einer vollständigen Randomisierung, insbesondere bei kleineren Prüfungen bzw. bei Prüfungen mit ungleichen Verhältnissen, liegt darin, daß die Zahl der Patienten in den

Behandlungsgruppen anders als geplant ausfallen kann. In diesen Fällen ist eine Pseudorandomisierung, die eine ausgeglichene Zuordnung entsprechend den Prüfanforderungen erlaubt, vorzuziehen.

Zur Auswertung der Ergebnisse von Vergleichsprüfungen gibt es verschiedene Methoden. Ihre Wahl hängt von der jeweiligen Problemstellung ab. Wenn beispielsweise die Zeit bis zum Eintritt eines Ereignisses irrelevant ist (Untersuchungen im Hinblick auf prophylaktische Anwendungen), brauchen die einzelnen Ereignisse nur gezählt zu werden. Die Signifikanz der Unterschiede zwischen den Behandlungen wird mit Hilfe des Chi-Quadrat-Tests beurteilt. Wenn aber ein erheblicher Teil der Patienten zu unterschiedlichem Zeitpunkt erkrankt, läßt sich eine empfindlichere und aussagekräftigere Beurteilung des Therapiewerts nicht nur durch Berücksichtigung des erkrankten Patientenanteils, sondern auch des Zeitpunkts, zu dem die Erkrankung nach der Zuordnung zur Patientengruppe eintrat, erzielen. Die zwei Methoden, die sich in diesem Zusammenhang als exakt und aussagekräftig erwiesen haben, sind die Life Table Methode und die logarithmischen P-Werte (2).

Die angeführten Methoden gehen davon aus, daß die erforderliche Fallzahl vor Beginn der Untersuchung berechnet und festgelegt wurde und die Auswertung erst dann vorgenommen wird, wenn die Ergebnisse vorliegen. Es muß betont werden, daß die Resultate verfälscht sein können, wenn diese Grundvoraussetzungen nicht erfüllt werden. Die Versuchsanforderung kann aber so geplant werden, daß die Beurteilung in Abständen während der Laufzeit der Untersuchung durchgeführt wird. Die Untersuchung kann abgebrochen werden, sobald eine Differenz festgestellt wird. Diese sequentiellen Analysen haben einen großen Vorteil, der von Prof. Armitage (4) herausgestellt wurde. Die Untersuchung läßt sich automatisch abbrechen, sobald sich eine Behandlungsart als wesentlich ungünstiger erweist. Allgemein ist es jedoch so, daß sich ein statistisch signifikanter Unterschied bei der Verwendung der sequentiellen Methode kaum so rasch zeigen wird, wenn eine Behandlungsart nur geringfügig ungünstiger ist (2).

Ethische Erwägungen spielen in jedem Stadium der klinischen Prüfung eine große Rolle und setzen der Untersuchung bestimmte Grenzen. Besonders wenn der Kliniker davon überzeugt ist, daß eine Therapieform für einen Patienten besser geeignet ist als eine andere, kann er die Auswahl nicht randomisieren.

Doppelblindversuche können dieses Problem zwar abschwächen, sind jedoch nicht anwendbar, wenn es darauf ankommt, daß der Arzt die Behandlung kennen muß, um Nebenwirkungen zu vermeiden. Wenn der Prüfungsablauf für einen Patienten völlig ungeeignet ist, muß aus ethischen Erwägungen davon abgegangen werden, selbst wenn dadurch die Befunde der Prüfung insgesamt abgeschwächt werden. Das Ausscheiden eines Patienten ist allerdings bei pragmatischer Handhabung einer Prüfung tatsächlich weniger schwerwiegend. Die Gegebenheiten, die auch in der Praxis zu erwarten sind, können von vornherein in die Definition der

Behandlung eingeführt werden (3). Auf keinen Fall dürfen ausgeschiedene Patienten jedoch von der Auswertung ausgeschlossen werden, weil damit das Ergebnis verzerrt und die Untersuchung ihren Sinn verlieren würde.

Literatur

1. Schwartz, D., und Lellouch, J. (1967), Explanatory and pragmatic attitudes in clinical trials, J. Chron. Dis., 20, 637–648.
2. Peto, R., Pike, M. C. et al. (1976), Design and analysis of randomized clinical trials requiring prolonged observations of each patient, Br. J. Cancer, 34, 585–612.
3. Schwartz, D., Flamant, R., und Lellouch, J. (1980), Clinical Trials, Academic Press Inc. (London) Ltd.
4. Armitage, P. (1975), Sequential Medical Trials, 2nd ed., Oxford: Blackwell Scientific Publications.