ÜBERSICHTSARBEIT

Wie bewertet man die p-Wert-Flut?

Hinweise zum Umgang mit dem multiplen Testen

Teil 10 der Serie zur Bewertung wissenschaftlicher Publikationen

Anja Victor, Amelie Elsäßer, Gerhard Hommel, Maria Blettner

ZUSAMMENFASSUNG

Hintergrund: Beim Lesen einer Publikation medizinischer Forschungsergebnisse trifft man zumeist auf p-Werte. In der Regel enthält eine Publikation nicht nur einen einzigen p-Wert, sondern die Autoren liefern eine ganze Flut, zumeist in Verbindung mit dem Wort „signifikant“. In diesem Artikel soll dem Leser die Problematik solcher p-Wert-Fluten erläutert und der Umgang damit aufgezeigt werden.

Methoden: Das Auftreten mehrerer p-Werte in einer Studie entsteht in der Regel durch das sogenannte „multiple Testen“. Es werden verschiedene Möglichkeiten zum korrekten Umgang mit diesem Problem vorgestellt. Der Artikel basiert auf klassischen Methoden der Statistik, wie sie in vielen Lehrbüchern dargestellt werden, und auf ausgewählter Spezialliteratur.

Ergebnisse: Generell sollte man Ergebnisse aus Publikationen mit vielen „Signifikanzen“, in denen der Autor nicht das Problem des „multiplen Testens“ durch adäquate Methoden berücksichtigt hat, vorsichtig bewerten. Forscher sollten vor Beginn ihrer Untersuchungen die Ziele klar definieren und, wenn möglich, ein einziges Hauptzielkriterium a priori definieren. Im Falle explorativer/hypothesengenerierender Studien ist darauf hinzuweisen, dass die Ergebnisse häufig zufälliger Natur sein könnten und in weiteren gezielten Studien bestätigt werden müssen.

Schlussfolgerungen: Insgesamt wird ein vorsichtiger Umgang mit dem Wort „signifikant“ und der Bewertung des Wortes empfohlen. Leser sollten Artikel im Hinblick auf das Problem des multiplen Testens kritisch bewerten. Autoren sollten die Anzahl der durchgeführten Tests angeben. Artikel sollten nach ihrer Qualität bewertet werden und nicht nach dem Auftreten des Begriffes „signifikant“.

Zitierweise: Dtsch Arztebl Int 2010; 107(4): 50–6
DOI: 10.3238/arztebl.2010.0050

Institut für medizinische Biometrie, Epidemiologie und Informatik Mainz:
Dr. rer. physiol. Victor, Dipl.-Stat. Elsäßer, Prof. Dr. rer. nat. Hommel,
Prof. Dr. rer. nat. Blettner

Autoren medizinischer Publikationen untermauern ihre Aussagen gerne mit p-Werten und dem Wort „signifikant“. Wie sind solche p-Werte und das gehäufte Auftreten von „signifikant“ überhaupt zu bewerten? Dazu wird zunächst erläutert, was ein p-Wert ist, und was „signifikant“ bedeutet.

Generell sollte jeder Untersuchung eine Hypothese zugrunde liegen. Es ist in der Praxis nicht möglich, diese Hypothese an allen sie betreffenden Personen zu überprüfen. Vielmehr wird eine medizinische Hypothese (zum Beispiel ein neues Medikament ist dem bisherigen Standardpräparat in der Senkung des systolischen Blutdrucks nach 16 Wochen Behandlung überlegen) nur an einer möglichst repräsentativen Stichprobe von Patienten überprüft. Die Ergebnisse aus dieser Stichprobe werden herangezogen, um über die Gültigkeit der Hypothese zu entscheiden. Die Möglichkeit eines Fehlers (Irrtumswahrscheinlichkeit) bleibt bei Annahme der Gültigkeit der Hypothese immer, denn es wurde nur eine Stichprobe untersucht. Die Entscheidung könnte sich zufällig genau entgegen der tatsächlichen Sachlage verhalten. In der Regel legt man die maximal tolerable Irrtumswahrscheinlichkeit, das sogenannte Signifikanzniveau (α), auf 5 % fest. Das Verfahren, das zur Bestätigung der Hypothese (mit maximaler Irrtumswahrscheinlichkeit α), oder zu ihrer Nichtbestätigung anhand der Ergebnisse der Stichprobe führt, ist der statistische Test. Dieser liefert als Ergebnis den sogenannten p-Wert, welchen man zur Entscheidungsfindung mit dem Signifikanzniveau vergleicht. Liegt der p-Wert unterhalb des Signifikanzniveaus, so gilt die Hypothese mit maximaler Irrtumswahrscheinlichkeit von α als abgesichert. Liegt er darüber, so kann die Hypothese nicht als bestätigt gelten *(Kasten 1)*.

Die meisten Publikationen beschränken sich nicht auf eine einzige Hypothese, stattdessen werden an derselben Stichprobe mehrere Hypothesen untersucht. Im genannten Beispiel könnte es sein, dass nicht nur die Wirkung der Blutdrucksenkung nach 16 Wochen zwischen Standard- und neuem Präparat untersucht wird, sondern

- auch die Blutdrucksenkung nach vier und acht Wochen
- zusätzlich der diastolische Blutdruck, ebenso wie Blutfettwerte
- noch weitere zwei Präparate und ein Placebo mit untersucht werden (dies steigert die Zahl der möglichen Vergleiche zwischen den Medikamenten auf zehn Vergleiche)

- anschließend „Subgruppenanalysen“ durchgeführt werden: Wie sind die Ergebnisse, wenn man Männer und Frauen oder wenn man ältere und jüngere Patienten getrennt betrachtet.

An diesem Beispiel wird deutlich, dass in einer Studie eine Vielzahl von Tests möglich sind.

Was passiert, wenn mehrere Hypothesen an demselben Kollektiv gleichzeitig getestet werden? Die Wahrscheinlichkeit, eine Falschaussage zu machen, steigt mit der Anzahl der durchgeführten Tests, denn man kann sich bei jedem Test irren. Vergleicht man den p-Wert jedes Tests weiterhin mit α, so kann man sich bei jeder der Aussagen mit der Wahrscheinlichkeit α irren. In der Summe über alle Tests steigt die Wahrscheinlichkeit, dass man mindestens eine Falschaussage macht, dramatisch an. Im Falle unabhängiger Tests lässt sich diese Gesamt-Irrtumswahrscheinlichkeit leicht ausrechnen *(Kasten 2)*. Bei 20 Tests, deren p-Werte jeweils mit α = 5 % verglichen werden, erwartet man bereits, dass rein zufällig ein p-Wert unter α liegt.

Das Problem des multiplen Testens tritt besonders häufig und ausgeprägt in genetischen Studien oder Prognosestudien auf.

In genetischen Assoziationsanalysen (Untersuchungen, ob eine Krankheit mit genetischen Markern verbunden ist, zum Beispiel mit „single nucleotide polymorphisms“ [SNPs]) wird in der Regel nicht nur ein genetischer Marker untersucht, sondern eine ganze Reihe gleichzeitig. Denkt man an die genomweiten Assoziationsstudien (1–3), bei denen in einer Studie Marker, die das ganze Genom repräsentieren, untersucht werden, gehen die Zahlen der untersuchten Marker schnell in

KASTEN 1

Der statistische Test und der p-Wert *(siehe auch Tabelle 1)*

1. Aufstellung der Hypothese:

Hypothese: Das neue Medikament ist dem bisherigen Standardpräparat in der Senkung des systolischen Blutdrucks nach 16 Wochen Behandlung überlegen

Zugehörige Nullhypothese: Das neue Medikament ist dem bisherigen Standardpräparat in der Senkung des systolischen Blutdrucks nach 16 Wochen Behandlung nicht überlegen

Die Nullhypothese ist das Gegenteil der Hypothese. Die Hypothese wird in der statistischen Nomenklatur auch als Alternativhypothese oder Gegenhypothese bezeichnet.

2. Datenerhebung an Stichprobe von Patienten

3. Statistischer Test zur Auswertung der Stichprobendaten

Ein statistischer Test geht davon aus, dass die Nullhypothese vorliegt und prüft, ob unter dieser Annahme die in der Stichprobe gemessenen Werte plausibel (im Beispiel: Werte des neuen Medikaments sind nicht besser als die des Standardpräparats), oder eher unplausibel sind (im Beispiel: Werte des neuen Medikaments deutlich besser als die des Standardpräparats). Aus den Daten der Stichprobe wird anhand eines statistischen Tests eine Plausibilitätsgröße berechnet („Wie wahrscheinlich ist das Ergebnis bei Gültigkeit der Nullhypothese?“). Diese liegt als Wahrscheinlichkeit zwischen 0 und 1 und wird p-Wert genannt. Je unwahrscheinlicher die beobachteten Daten bei Gültigkeit der Nullhypothese sind, desto mehr spricht gegen die Nullhypothese und desto mehr für die Hypothese, desto kleiner wird der p-Wert.

4. Testentscheidung: Annahme der Hypothese?

Ist der p-Wert klein, so sind die beobachteten Daten bei Gültigkeit der Nullhypothese unwahrscheinlich; das spricht gegen die Gültigkeit der Nullhypothese. Ein kleiner p-Wert spricht somit für das Gegenteil, die Hypothese. Ist der p-Wert kleiner als eine vorgegebene Schranke α (das Signifikanzniveau), so wird die Hypothese angenommen. Das Problem ist, dass in einer Stichprobe zufällig auffällige Werte zustande kommen können, auch wenn die Nullhypothese tatsächlich wahr ist. Nimmt man die Hypothese an, obwohl sie nicht wahr ist, begeht man einen Irrtum, man spricht vom Fehler erster Art. Diese Irrtumswahrscheinlichkeit wird mittels des Signifikanzniveaus α des Tests begrenzt. Man spricht von einem statistischen Test zum Niveau α. In der Regel setzt man α = 5 %. Das bedeutet, dass nur in 5 % der Fälle die Hypothese fälschlich als zutreffend bezeichnet wird. Die Testentscheidung wird mittels Vergleich des berechneten p-Werts mit dem vorgegebenen α gefällt. Das Ergebnis eines statistischen Tests wird dann als signifikant bezeichnet, wenn der berechnete p-Wert kleiner ist als das vorher festgelegte α; in diesem Fall wird die Hypothese mit maximal der Irrtumswahrscheinlichkeit α angenommen.

Ist der p-Wert größer als das Niveau α, so kann die Hypothese nicht angenommen werden. Es kann aber in diesem Fall nicht von der Gültigkeit der Nullhypothese ausgegangen werden, da man den Fehler zweiter Art im Gegensatz zu dem Fehler erster Art nicht kontrollieren kann. Der Fehler zweiter Art beschreibt in diesem Beispiel den Fehler, dass in Wahrheit das neue Medikament besser ist, aber die Nullhypothese beibehalten wurde. Die sogenannte Power ist eins minus der Fehler zweiter Art (1–Fehler zweiter Art), also die Wahrscheinlichkeit, mit der man die Hypothese korrekterweise annimmt (zeigen kann). Damit eine Studie Erfolg hat, sollte die Power möglichst groß sein. Sie kann aber nach Datenerhebung nicht mehr kontrolliert werden.

TABELLE 1

Entscheidung aufgrund der Stichprobe

	Nullhypothese beibehalten	Hypothese annehmen
Nullhypothese ist wahr	korrekte Entscheidung	Fehler 1. Art
Hypothese ist wahr	Fehler 2. Art	korrekte Entscheidung

die Tausende. Ähnliches gilt für Genexpressionsanalysen, bei denen auf einem Mikroarray die Expression mehrerer Tausend Gene untersucht wird. Bei 1 000 Tests, die alle zu $\alpha = 0{,}05$ durchgeführt werden, erwartet man rein zufällig 50 p-Werte kleiner als 0,05, welche dann zu falschpositiven Aussagen führen. Gerade für genetische Assoziationsstudien hat man daher festgestellt, dass sich viele Assoziationsaussagen nicht reproduzieren lassen, also höchstwahrscheinlich falschpositive Aussagen sind (4, 5).

Bei Studien auf Prognose untersucht man häufig eine Vielzahl potenzieller Einflussfaktoren. Bei einer Studie zur Prognose von Brustkrebs beispielsweise können neben den klassischen Faktoren unter Anderem zahlreiche histologische Tumoreigenschaften mit berücksichtigt werden. Bei Studien zur Prognose der koronaren Herzkrankheit wird oft eine immense Zahl von Labormarkern zusätzlich zu den klassischen Faktoren untersucht.

Multiples Testen entsteht aber auch in vielen anderen Gebieten durch multiple Endpunkte, Subgruppenanalysen, den Vergleich mehrerer Gruppen oder Zwischenanalysen bei sequenziellen Studiendesigns.

Dieser Artikel basiert auf den klassischen Methoden der Statistik, wie sie in vielen Lehrbüchern dargestellt werden, und auf ausgewählter Spezialliteratur.

KASTEN 2

Wahrscheinlichkeit, mindestens eine Nullhypothese fälschlich abzulehnen (= eine Hypothese fälschlich anzunehmen = ein falschpositives Ergebnis zu postulieren) wenn man zehn unabhängige Tests jeweils zum lokalen Niveau 5 % durchführt

= 1 – Wahrscheinlichkeit keine Nullhypothese aller zehn Tests fälschlich abzulehnen
= 1 – (Wahrscheinlichkeit keiner falschen Ablehnung je Test)^{10}
= 1 – (1 – Wahrscheinlichkeit einer falschen Ablehnung je Test)^{10}
= $1 - (1 - \alpha)^{10}$
= $1 - (0{,}95)^{10}$
= 1 – 0,60
= 0,4 = 40 %

Methodik multiplen Testens

Um die Flut falschpositiver Ergebnisse in der medizinischen Forschung einzudämmen, sind Maßnahmen zur Kontrolle der Irrtumswahrscheinlichkeit bezogen auf alle untersuchten Hypothesen nötig.

Anstatt das Niveau nur jedes einzelnen Tests zu betrachten, gibt es daher die Definition der FWER („familywise error rate“). Sie beschreibt die Wahrscheinlichkeit, dass man mindestens eine von allen untersuchten Nullhypothesen fälschlich ablehnt. Wenn man diese „Gesamt“-Wahrscheinlichkeit mit einer kleinen Größe (zum Beispiel $\alpha = 5$ %) kontrolliert, so kann man sich recht sicher (bei Gesamt-$\alpha = 5$ % mit 95 %) sein, keine falschpositive Aussage zu machen. Man spricht bei Kontrolle der FWER vom multiplen Niveau α, um zu verdeutlichen, dass man die Irrtumswahrscheinlichkeit aller Tests gleichzeitig beschränkt. Im Gegensatz dazu bedeutet lokales Niveau, dass keine Gesamt-Fehler-Betrachtung erfolgt.

Wie kontrolliert man die FWER? Anstatt jeden p-Wert mit dem Gesamtniveau α zu vergleichen, muss man eine kleinere Grenze für jeden einzelnen p-Wert ansetzen. Es gibt zahlreiche verschiedene Verfahren, wie diese kleinere Grenze zu wählen ist. Analog kann man auch umgekehrt vorgehen und den p-Wert nach solchen Verfahren vergrößern (adjustieren) und diesen dann mit dem multiplen Gesamtniveau α vergleichen. Der Vorteil der adjustierten p-Werte liegt in der einfacheren Verständlichkeit für den Leser, denn dieser kann den adjustierten p-Wert wie gewohnt mit α (zum Beispiel = 5 %) vergleichen. Man vermeidet so, dass ein Leser sich wundert, warum ein „so schön kleiner“ p-Wert nicht signifikant ist.

Die wohl bekannteste Methode zur Kontrolle der FWER ist die nach Bonferroni. Damit der Gesamtfehler (die Wahrscheinlichkeit, mindestens eine falschpositive Aussage zu machen = die FWER) nicht α (zum Beispiel 5 %) überschreitet, teilt man das multiple Gesamtniveau durch die Anzahl durchgeführter Tests und vergleicht jeden p-Wert mit dieser kleineren Schranke. Ist zum Beispiel die Anzahl untersuchter Hypothesen 100 und das gewählte Gesamtniveau 5 %, so ist der p-Wert jedes Tests (jeder Hypothese) mit 5 %/100 = 0,0005 zu vergleichen. Zur FWER von 5 % können bei Verwendung dieser Prozedur dann nur die Hypothesen angenommen werden (als signifikant benannt werden), deren p-Werte kleiner gleich 0,0005 sind. Ein Rechenbeispiel findet man in *Kasten 3b*, weiteres zur Prozedur in *Kasten 3a*. Diese Prozedur hält das gewählte Niveau der FWER auch bei allen Formen der Abhängigkeit zwischen den Hypothesen ein. Die Prozedur ist allerdings sehr strikt, das heißt, es können Ergebnisse übersehen werden.

Eine Abwandlung dieser Prozedur mit Steigerung der Power ist die Bonferroni-Holm-Prozedur. Dabei werden alle p-Werte der Größe nach sortiert und mit wachsenden Schranken verglichen *(Kasten 3a, Kasten 3b)*.

Eine andere Möglichkeit zur Kontrolle der FWER ist das Verfahren der hierarchischen Ordnung. Dabei werden die Hypothesen a priori (vor Versuchsbeginn) ihrer

Wichtigkeit nach geordnet und die zugehörigen p-Werte dieser Ordnung folgend (angefangen bei der wichtigsten Hypothese) mit dem gewählten multiplen FWER-Niveau verglichen. Hypothesen können so lange abgelehnt werden, bis zum ersten Mal der p-Wert einer Hypothese in der absteigenden Ordnung nicht mehr kleiner als das gewählte Niveau ist. Diese Prozedur bietet den Vorteil, dass man alle p-Werte mit dem vollen Niveau (zum Beispiel 5 %) vergleichen kann. Jedoch ist nach erstmaliger Überschreitung des Niveaus keine weitere Annahme von Hypothesen möglich, unabhängig von der Größe aller noch folgenden p-Werte (auch wenn viele weitere p-Werte deutlich kleiner als das Niveau sind) *(Kasten 3a und b)*. Diese Prozedur eignet sich speziell bei klinischen Studien, in denen klar geordnete Hauptzielkriterien vorliegen (zum Beispiel Wirksamkeit als oberste Hypothese und anschließend geringere Nebenwirkungshäufigkeiten als zweite Hypothese). Für Untersuchungen, die explorativer Natur sind (zum Beispiel genetische Studien) und bei denen daher keine Ordnung der Hypothesen vorher festgelegt werden kann, eignet sich diese Prozedur nicht.

Für zwei weitere oft angewandte Prozeduren zur Kontrolle der FWER soll hier nur auf häufige Fehler bei ihrer Verwendung hingewiesen werden. Fisher's LSD-Test kontrolliert nur die FWER, wenn maximal drei Gruppen jeweils paarweise miteinander verglichen werden. Vergleicht man mehr als drei Gruppen, so ist dieser Test kein adäquates Mittel zur Kontrolle der FWER. Beim Vergleich verschiedener Dosierungen gegen eine Kontrolle wird häufig die Dunnett-Prozedur verwendet. Diese Prozedur hält die FWER nur für die Vergleiche gegen die Kontrolle; Vergleiche unter den verschiedenen Dosierungen dürfen nicht einbezogen werden.

Zu allen genannten und weiteren Prozeduren vergleiche man auch das Buch von Horn und Vollandt (6).

Die Wahrscheinlichkeit, mindestens eine Hypothese fälschlich abzulehnen (die FWER) steigt schnell mit der Anzahl der Tests. Daher müssen, um die FWER zu kontrollieren, sehr strikte Ablehnkriterien erfüllt wer-

KASTEN 3a

Bonferroni-Holm- und explorativer Simes (Benjamini-Hochberg)-Verfahren und Adjustieren von p-Werten beim Bonferroni-Verfahren

Variante des Bonferroni-Verfahrens zum Adjustieren der p-Werte

Analog zum im Text beschriebenen Bonferroni-Verfahren kann man auch die p-Werte adjustieren. Dies geschieht, indem man sie mit der Anzahl der Hypothesen multipliziert. Sollte sich bei dieser Adjustierung ein Wert größer 1 ergeben (zum Beispiel bei 30 Tests und einem p-Wert von 0,04 : 30 × 0,04 = 1,2), so wird der adjustierte p-Wert auf 1 gesetzt, da p-Werte als Wahrscheinlichkeiten nicht größer 1 sein können. Anschließend werden die adjustierten p-Werte mit dem Gesamtniveau α verglichen.

Bonferroni-Holm

Zunächst werden alle p-Werte der Größe nach sortiert und anschließend mit wachsenden Schranken verglichen. Die kleinste Schranke ist wie bei Bonferroni das Niveau geteilt durch die Anzahl der Hypothesen. Die Schranke für den nächsten p-Wert ist dann das Niveau geteilt durch die Anzahl der Hypothesen minus 1, die darauf folgende das Niveau geteilt durch Anzahl Hypothesen minus 2 und so fort. In einem Beispiel mit Niveau 5 % und 100 Hypothesen wäre

- der kleinste p-Wert mit 5 %/100 = 0,0005
- der zweitkleinste p-Wert mit 5 %/99 ~ 0,000505
- der drittkleinste mit 5 %/98 ~ 0,00051 und so weiter

zu vergleichen. Abgelehnt werden können Nullhypothesen mit p-Werten, die kleiner als die zugehörige Schranke waren, jedoch nur bis zum ersten Mal die Schranke überschritten wird. Diese Prozedur kontrolliert die FWER (Familywise Error Rate) auch bei allen Formen der Abhängigkeit zwischen den Hypothesen.

Explorative Simes-Prozedur / Benjamini-Hochberg-Prozedur

Auch für diese Prozedur sind die p-Werte der Größe nach zu ordnen. Der kleinste p-Wert ist wiederum mit der Bonferroni-Schranke zu vergleichen: gewähltes FDR-Niveau (False Discovery Rate) geteilt durch Anzahl der Hypothesen. Der zweitkleinste p-Wert ist mit dem Niveau multipliziert mit 2, geteilt durch Anzahl der Hypothesen, zu vergleichen; der drittkleinste mit dem Niveau multipliziert mit 3, geteilt durch die Anzahl der Hypothesen, und so weiter. Im Beispiel mit gewähltem FDR-Niveau 5 % und 100 Hypothesen wachsen die Schranken somit wie folgt:

- der kleinste p-Wert ist mit 5 %/100 = 0,0005
- der zweitkleinste p-Wert mit 5 % × 2/100 = 0,001
- der drittkleinste mit 5 % × 3/100 = 0,0015 und so weiter zu vergleichen.

Bei dieser Prozedur kann man nicht nur die Nullhypothesen zu p-Werten bis zur ersten Überschreitung der Grenzen ablehnen, wie das bei Bonferroni-Holm der Fall ist. Man kann bei dieser Prozedur alle Nullhypothesen ablehnen, die zu p-Werten gehören, die kleiner als der größte p-Wert sind, der eine zugehörige Schranke unterschreitet. Diese Prozedur kontrolliert im Falle der Unabhängigkeit und der sogenannten „positive regression dependency" (PRDS, eine spezielle Form positiver Abhängigkeit) der Hypothesen die FDR zum gewählten Niveau.

KASTEN 3b

Beispiel zur Anwendung der vorgestellten multiplen Test-Prozeduren

Es werden vier Hypothesen untersucht zur FWER 5 % (Familywise Error Rate) beziehungsweise zur FDR 5 % (Familywise Discovery Rate). Die sich ergebenden p-Werte seien $p_1 = 0{,}03$, $p_2 = 0{,}01$, $p_3 = 0{,}035$ und $p_4 = 0{,}30$.

Vorgehen bei der Bonferroni-Korrektur (Kontrolle der FWER)

Vergleich der p-Werte mit 5 %/Anzahl der Tests = 5 %/4 = 1,25 % = 0,0125. Nur p_2 ist kleiner als diese Schranke und nur die dazugehörige Hypothese kann als signifikant bezeichnet werden. Umgekehrt könnte man auch die p-Werte adjustieren (mit der Anzahl der Tests = 4 multiplizieren). Adjustiert ergäbe sich adj. $p_1 = 0{,}03 \times 4 = 0{,}12$, adj. $p_2 = 0{,}01 \times 4 = 0{,}04$, adj. $p_3 = 0{,}035 \times 4 = 0{,}14$ und adj. $p_4 = 0{,}30 \times 4 = 1{,}2$. Letzterer Wert ist größer als 1 und wird somit als adj. $p_4 = 1$ gesetzt. Die adjustierten p-Werte kann man dann mit dem Gesamtniveau 5 % vergleichen, was zum selben Ergebnis wie mit den angepassten Schranken führt.

Vorgehen bei der Bonferroni-Holm-Korrektur (Kontrolle der FWER)

Zunächst müssen die p-Werte der Größe nach sortiert werden:
$p_2 = 0{,}01$; $p_1 = 0{,}03$; $p_3 = 0{,}035$; $p_4 = 0{,}30$
Die aufsteigenden Schranken sind: 5 %/4 = 0,0125; 5 %/3 = 0,0167; 5 %/2 = 0,025; 5 %/1 = 5 %
Vergleich des kleinsten p-Werts mit der untersten Schranke $p_2 = 0{,}01 < 0{,}0125$: zugehörige Nullhypothese ablehnbar
Vergleich des zweitkleinsten p-Werts mit der zweiten Schranke $p_1 = 0{,}03 > 0{,}0167$: Ende der Prozedur, keine weitere Nullhypothese ablehnbar, keine weitere Hypothese annehmbar

Vorgehen bei hierarchischer Ordnung (Kontrolle der FWER)

1. Fall: Man hatte die Hypothesen folgendermaßen angeordnet: Als wichtigste Hypothese H_1 (zu p_1 gehörig), dann H_2, dann H_3, als letztes H_4.
Da $p_1 \leq 0{,}05$ kann das Ergebnis für H_1 als signifikant bezeichnet werden. Gleiches gilt für H_2 und H_3, aber nicht mehr für H_4, da $p_4 > 0{,}05$.
2. Fall: Man hatte die Hypothesen folgendermaßen angeordnet: Als wichtigste Hypothese H_4 (zu p_4 gehörig), dann H_3, dann H_2, als letztes H_1.
Da $p_4 > 0{,}05$, kann H_4 nicht als signifikant bezeichnet werden. Gleiches gilt für alle weiteren Hypothesen, da der p-Wert der hierarchisch am höchsten angeordneten Hypothese zu groß war.

Vorgehen bei der explorativen Simes-Prozedur (auch Benjamini-Hochberg-Prozedur genannt, Kontrolle der FDR)

Zunächst müssen die p-Werte der Größe nach sortiert werden:
$p_2 = 0{,}01$; $p_1 = 0{,}03$; $p_3 = 0{,}035$; $p_4 = 0{,}30$
Die aufsteigenden Schranken sind: 5 %/4 = 0,0125; 5 %/4 × 2 = 0,025; 5 %/4 × 3 = 0,0375; 5 %/4 × 4 = 5 %
Vergleich des kleinsten p-Werts mit der untersten Schranke $p_2 = 0{,}01 < 0{,}0125$: zugehörige Hypothese annehmbar
Vergleich des zweitkleinsten p-Werts mit der zweiten Schranke $p_1 = 0{,}03 > 0{,}025$: trotzdem noch kein Ende der Prozedur
Vergleich des drittkleinsten p-Werts mit der dritten Schranke $p_3 = 0{,}035 < 0{,}0375$: somit kann die zu p_3 gehörende Nullhypothese abgelehnt werden, zusätzlich alle zu kleineren p-Werten gehörige Hypothesen, also neben der zu p_2 auch die zu p_1 gehörige, obwohl sie ihre eigene Schranke nicht geschafft hat
Vergleich des größten p-Werts mit der vierten Schranke $p_4 = 0{,}30 > 0{,}05$: keine weitere Ablehnung möglich

Zusammenfassender Vergleich in diesem Beispiel: abgelehnte Nullhypothesen = signifikante Hypothesen

- mit Bonferroni-Verfahren: H_2
- mit Bonferroni-Holm-Verfahren: H_2
- mit hierarchischer Ordnung, falls nach 1. Fall geordnet: H_1, H_2, H_3
- mit hierarchischer Ordnung, falls nach 2. Fall geordnet: keine
- mit explorativer Simes-Prozedur (Benjamini-Hochberg): H_1, H_2, H_3

In diesem Beispiel gibt es trotz größer werdender Schranken keine zusätzlichen Ablehnungen bei Bonferroni-Holm im Vergleich zum Bonferroni-Verfahren aber mehr Ablehnungen bei Kontrolle der FDR. Man erkennt sowohl den möglichen Gewinn als auch die Gefahr der hierarchischen Anordnung (im Vergleich zwischen a priori Ordnung nach Fall 1 beziehungsweise Fall 2). Die Ordnung sollte daher immer sachbezogen erfolgen.

den, wie man an obigen Beispielen sieht. So kann rigorose Auslegung des multiplen Testens in Verbindung mit vielen Tests eine geringe Power zur Folge haben, das heißt man übersieht wahre Aussagen. Dies wird dann fälschlicherweise oft als Negativbeweis interpretiert. In Studien, denen weitere Folge-Studien angeschlossen werden, kann es wichtiger sein, möglichst wenige potenzielle Ansatzpunkte zu verpassen und dafür einige fälschlich signifikante Hypothesen in Kauf zu nehmen. Für solche Situationen bietet sich die FDR („false discovery rate“) als weniger strikte Möglichkeit der Fehlerkontrolle an. Bei der FDR kontrolliert man den erwarteten Anteil der fälschlich abgelehnten an allen abgelehnten Hypothesen *(Tabelle 2)*.

Die am weitesten verbreitete Prozedur zur Kontrolle der FDR ist die sogenannte explorative Simes-Prozedur, von den meisten Autoren Benjamini-Hochberg-Prozedur genannt. Die Prozedur wurde von Simes (7) erwähnt, jedoch zeigten erst Benjamini und Hochberg (8), dass diese Prozedur die FDR kontrolliert. Das genaue Vorgehen ist in *Kasten 3a* beschrieben, ein Rechenbeispiel wiederum findet man in *Kasten 3b*. Diese Prozedur führt zu mehr Ablehnungen als die Bonferroni-Holm-Prozedur. Mit der FDR-Kontrolle wird ein weniger striktes Fehlerkriterium verwendet, mit welchem mehr Power erreicht wird, jedoch auch mehr falschpositive Aussagen akzeptiert werden. Die FDR sollte daher als Fehlerdefinition nicht in klinischen Studien, sondern nur in eher explorativ angelegten Untersuchungen verwendet werden.

Um das Problem des multiplen Testens generell zu umgehen oder gering zu halten, sollte man vor allem bei klinischen Studien eine oder sehr wenige Haupthypothesen formulieren, welche man dann konfirmatorisch unter Anwendung einer Prozedur zur Kontrolle der FWER testet. Alle weiteren durchgeführten Tests dürfen dann nicht mit dem Wort „signifikant“ bezeichnet werden und sind vorsichtig zu interpretieren. Dieses Vorgehen wird auch von der EMEA (European Medicines Agency) vorgeschlagen (9). Bei rein explorativen Studien, die eher der Hypothesengenerierung dienen, kann entweder die FDR als Fehlerdefinition gewählt werden, oder auf die Korrektur für multiples Testen verzichtet werden. Im letzteren Fall darf nicht von signifikanten Ergebnissen gesprochen werden, sondern nur von auffällig kleinen p-Werten, die als Motivation für eventuelle weitere (dann vielleicht konfirmatorische) Studien dienen. Man muss deutlich machen, dass es sich auch um zufällige Ergebnisse handeln kann, da keinerlei Kontrolle der Irrtumswahrscheinlichkeit stattgefunden hat.

Resultate

Da heute die Untersuchungsmethoden hoch technisiert sind und es möglich ist, je Patient unzählige Daten zu erheben (Laborwerte, genetische Daten etc.) werden in einer Studie sehr viele Tests durchgeführt. Bei Nichtbeachtung des multiplen Testproblems führt dies zu zahlreichen falschpositiven Zufallsfunden, die anschließend publiziert werden. Sind erst einmal falschpositive

TABELLE 2

Zur Erläuterung der Fehlerraten. Die FWER ist die Wahrscheinlichkeit, dass $V > 0$, die FDR ist der Erwartungswert von (V/R).

	Beibehaltene Nullhypothesen	Abgelehnte Nullhypothesen	
Wahre Nullhypothesen	U	V	m_0
Falsche Nullhypothesen	S	T	m_1
		$R = V + T$	$M = m_0 + m_1$

Ergebnisse publiziert, dauert es lange, bis diese widerlegt sind und noch länger bis die Widerlegung allgemein bekannt ist. Man sollte sich darüber im Klaren sein, dass das Wort „signifikant“ häufig fälschlich verwendet wird und keineswegs ein „Gütekriterium“ ist. Bei Nichtbeachtung des multiplen Testens hat „signifikant“ nicht seine eigentliche Bedeutung der begrenzten Irrtumswahrscheinlichkeit und ein fälschlich als „signifikant“ bezeichnetes Ergebnis kann völlig wertlos für die Interpretation sein.

Für Forscher ist es daher unabdingbar, eine Studie gut zu planen und möglichst nur ein Hauptzielkriterium (oder wenige) zu wählen. In Artikeln sollte man ehrlich die Anzahl durchgeführter Tests angeben und entsprechende Prozeduren verwenden, um auf „Signifikanz“ zu entscheiden. Als Manipulation anzusehen ist das Vorgehen, viele Tests durchzuführen, die p-Werte jeweils mit α zu vergleichen und nur die Ergebnisse mit $p \leq \alpha$ zu erwähnen und als signifikant zu bezeichnen.

Auch die Zulassungsbehörden weisen auf dieses Problem hin und in klinischen Studien ist dem entsprechend Rechnung zu tragen (9).

Diskussion

Generell ist es immer notwendig, einen Artikel im Hinblick auf die Validität seiner Aussagen kritisch zu bewerten (10). Speziell das Problem der Nichtberücksichtigung des multiplen Testens ist weit verbreitet und wird zumeist unterschätzt. Diese Einschätzung basiert einerseits auf der persönlichen Erfahrung der Autoren durch die statistische Betreuung zahlreicher medizinischer Forschungsvorhaben und als Gutachter für Zeitschriften, wird andererseits aber auch durch Literatur (11, 12) gestützt. In zwei am IMBEI erstellten Literaturreviews zur Assoziation von Brustkrebs zu Polymorphismen im COMT beziehungsweise SULT1A1-Gen wurde festgestellt, dass in den meisten verwendeten Originalarbeiten mehrere Tests durchgeführt wurden. Beim Review für COMT war dies in 28 von 34 und bei SULT1A1 bei 10 von 14 Arbeiten der Fall. Allerdings wurde in den seltensten Fällen für das multiple Testen korrigiert. Das Problem wurde nur in 4 der 28 Arbeiten mit mehreren Tests für COMT und in einer der 10 Arbeiten mit mehreren Tests zu SULT1A1 berücksichtigt. Somit wurde in etwa 9 von 10 Originalartikeln das Pro-

blem des multiplen Testens nicht berücksichtigt, obwohl es auftrat.

Sowohl als Leser, wie auch als Editor oder Reviewer sollte man sein Augenmerk speziell darauf richten, dass das Wort „signifikant“ nicht unangebracht verwendet wird, sondern das Problem des multiplen Testens adäquat berücksichtigt wurde. Einer Inflation des Vorkommens von „signifikant“ sollte man kritisch gegenüberstehen. Ergebnisse, die „nur“ als explorativ dargestellt werden, sollte man nicht schlechter bewerten als solche, die ohne Beachtung des multiplen Testens als „signifikant“ bezeichnet wurden. Ein Ergebnis, was fälschlich mit „signifikant“ bezeichnet wurde, ist schlechter als eines, das korrekterweise vorsichtiger interpretiert wurde. Vielmehr geht der Leser wegen „signifikant“ fälschlicherweise von einer geringen Irrtumswahrscheinlichkeit aus. Leider ist es nicht immer möglich, multiples Testen zu erkennen. Verschweigt ein Autor, dass er eigentlich sehr viele Tests durchgeführt hat und veröffentlicht nur das Ergebnis seines auffälligsten Tests, so kann man dieses Ergebnis als Leser nicht in Relation zur Anzahl durchgeführter Tests bewerten. Man sollte darauf achten, ob es Hinweise gibt, dass mehr Tests als die aufgeführten durchgeführt wurden. Hinweise sind etwa Verweise der Autoren auf andere (eigene) Publikationen, in denen das Patientenkollektiv oder die Studie bereits beschrieben ist.

Selbst wenn Autoren erwähnen, dass sie Methoden zur Berücksichtigung des multiplen Testens verwendet haben, so ist es aufgrund der Vielzahl der vorhandenen Methoden für den nicht statistisch geschulten Leser schwierig zu bewerten, ob die verwendete Methode das Problem korrekt löst. Im Methodenteil wurden daher gängige, einfache Verfahren vorgestellt und häufige Fehler erwähnt. Generell sollten Ergebnisse aus einer Studie mit vielen Tests (wie zum Beispiel in genetischen Assoziationsstudien oder Prognosestudien) erst als wahrscheinlich wahr angesehen werden, wenn sie unabhängig reproduziert werden konnten.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt im Sinne der Richtlinien des International Committee of Medical Journal Editors besteht.

Manuskriptdaten
eingereicht: 9. 2. 2009, revidierte Fassung angenommen: 21. 7. 2009

LITERATUR

1. Sladek R, Rocheleau G, Rung J, et al.: A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007; 454: 881–5.
2. Samani NJ, Erdmann J, Hall AS, et al.: Genome-wide association analysis of coronary artery disease. NEJM 2007; 357: 443–53.
3. The Wellcome Trust Case Control Consortium: A genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. Nature 2007; 447: 661–78.
4. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG: Replication validity of genetic association studies. Nature Genetics 2001; 29: 306–9.
5. Lohmüller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genetics 2003; 33: 177–82.
6. Horn M, Vollandt R: Multiple Tests und Auswahlverfahren. Stuttgart 1995: Gustav Fischer Verlag.
7. Simes RJ: An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986; 73: 751–4.
8. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistic Society 1995; 57: 298–300.
9. EMEA: Points to consider on multiplicity issues in clinical trials, www.emea.europa.eu/pdfs/human/ewp/090899en.pdf
10. Prel JB du, Röhrig B, Blettner M: Kritisches Lesen wissenschaftlicher Artikel: Teil 1 der Serie zur Bewertung wissenschaftlicher Publikationen. Dtsch Arztebl Int 2009; 106(7): 100–5.
11. Cardon LR, Bell JI: Association study designs for complex diseases. Nature Reviews Genetics 2001; 2: 91–9.
12. Risch N: Searching for genetic determinants in the new millenium. Nature 2000; 405: 847–56.
13. Traunecker KB: Assoziation der genetischen Polymorphismen am Beispiel von SULT1A1 und COMT mit Brustkrebs. Promotionsschrift am Fachbereich Medizin der Universitätsmedizin der Johannes-Gutenberg Universität Mainz.

Anschrift für die Verfasser
Dr. rer. physiol. Anja Victor
Institut für medizinische Biometrie, Epidemiologie und Informatik
Universitätsmedizin der Johannes-Gutenberg-Universität Mainz
Obere Zahlbacher Straße 69
55101 Mainz
E-Mail: victor@imbei.uni-mainz.de

SUMMARY

Judging a Plethora of p-Values: How to Contend With the Problem of Multiple Testing

Part 10 of a Series on Evaluation of Scientific Publications

Background: When reading reports of medical research findings, one is usually confronted with p-values. Publications typically contain not just one p-value, but an abundance of them, mostly accompanied by the word “significant.” This article is intended to help readers understand the problem of multiple p-values and how to deal with it.

Methods: When multiple p-values appear in a single study, this is usually because of multiple testing. A number of valid approaches are presented for dealing with the problem. This article is based on classical statistical methods as presented in many textbooks and on selected specialized literature.

Results: Conclusions from publications with many “significant” results should be judged with caution if the authors have not taken adequate steps to correct for multiple testing. Researchers should define the goal of their study clearly at the outset and, if possible, define a single primary endpoint a priori. If the study is of an exploratory or hypothesis-generating nature, it should be clearly stated that any positive results might be due to chance and will need to be confirmed in further targeted studies.

Conclusions: It is recommended that the word “significant” be used and interpreted with care. Readers should assess articles critically with regard to the problem of multiple testing. Authors should state the number of tests that were performed. Scientific articles should be judged on their scientific merit rather than by the number of times they contain the word “significant.”

Zitierweise: Dtsch Arztebl Int 2010; 107(4): 50–6
DOI: 10.3238/arztebl.2010.0050

The English version of this article is available online:
www.aerzteblatt-international.de