Medizinische Statistik – Mathematik oder Orakel?

Abschiedsvorlesung am 16. Juli 2004

Medical Statistics – Mathematics or Oracle?

Farewell Lecture

• Wilhelm Gaus[1]

In der Medizin ist fast nichts sicher - dies ist eine direkte Auswirkung der Individualität jedes Menschen. Deshalb ist medizinische Statistik notwendig. Aber eine naive Interpretation von Statistiken kann in die Irre führen: Abb. 1 zeigt das Maximum der Suizidrate in der Jugend, Abb. 2 in der Mitte des Lebens, Abb. 3 im Alter. Welche dieser widersprüchlichen Botschaften ist die richtige?
Nach einer Einführung in das Prinzip des statistischen Tests wird nach der Wahrscheinlichkeit gefragt, dass das Ergebnis eines statistischen Tests richtig ist. Dazu wird das Signifikanzniveau und die Power des Tests verglichen mit Sensitivität und Spezifität eines diagnostischen Verfahrens. Die Wahrscheinlichkeit für die Richtigkeit des Ergebnisses eines statistischen Tests ist wie die positive und negative Korrektheit eines diagnostischen Verfahrens und somit abhängig von der Prävalenz.
Zum Problem des multiplen Testens wird gezeigt, dass an jedem Datenmaterial von vernünftiger Größe mindestens ein Test signifikant wird - auch wenn alle Daten mit dem Zufallszahlengenerator erzeugt worden sind. Es ist eminent wichtig, ob eine Hypothese unabhängig von den Daten, an denen sie getestet wird, generiert worden ist. Nach diesen Ausführungen werden Steigerungen wie „Ausreden, Lügen, Statistiken" verständlich, ebenso der Unterschied zwischen reiner Wahrheit und voller Wahrheit. Abschließend wird noch über zwei historische Orakel berichtet.

Schlüsselwörter: Biometrie, Suizidrate, statistischer Test, Richtigkeit eines statistischen Tests, Richtigkeit eines diagnostischen Verfahrens, multiples Testen, a priori formulierte versus nachgeschobene Hypothesen, reine versus volle Wahrheit, persönliche Worte

Certainty is rare in medicine. This is a direct consequence of the individuality of each and every human being and the reason why we need medical statistics. However, statistics have their pitfalls, too. Fig. 1 shows that the suicide rate peaks in youth, while in Fig. 2 the rate is highest in midlife and Fig. 3 in old age. Which of these contradictory messages is right?
After an introduction to the principles of statistical testing, this lecture examines the probability with which statistical test results are correct. For this purpose the level of significance and the power of the test are compared with the sensitivity and specificity

[1] *Medizinische Fakultät der Universität Ulm, Abteilung Biometrie und Medizinische Dokumentation, Ulm, Deutschland*

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp/M2005081414-1)

of a diagnostic procedure. The probability of obtaining correct statistical test results is the same as that for the positive and negative correctness of a diagnostic procedure and therefore depends on prevalence.
The focus then shifts to the problem of multiple statistical testing. The lecture demonstrates that for each data set of reasonable size at least one test result proves to be significant - even if the data set is produced by a random number generator. It is extremely important that a hypothesis is generated independently from the data used for its testing. These considerations enable us to understand the gradation of "lame excuses, lies and statistics" and the difference between pure truth and the full truth. Finally, two historical oracles are cited.

Keywords: medical statistics, biometry, suicide-risk, hypothesis testing, diagnostic investigation, correctness of a statistical test result, multiple testing, a priori formulated hypotheses, pure versus whole truth, historical oracles, personal remarks

Vorlesung

• Warum medizinische Statistik?

Das zentrale Problem der Medizin - aus meiner biometrischen Sicht - ist, dass fast nichts sicher ist. Dies ist eine direkte Folge der Individualität des Menschen. Wir alle sind stolz darauf, dass wir einmalig auf der ganzen Welt und in der Zeit sind. Bei dieser Einmaligkeit kann man nicht erwarten, dass Gesundheit und Krankheit bei allen Menschen genau gleich verlaufen. Eine Diagnose kann richtig sein, ist wahrscheinlich richtig, aber manchmal auch nicht. Meistens hilft die eingesetzte Therapie, aber leider nicht immer. Manchmal entwickelt sich eine Komplikation, meistens nicht. Damit sind wir schon mitten in der Statistik, die sich mit Dingen beschäftigt, die selten, manchmal, häufig, aber nicht ganz sicher eintreten.

Medizinische Statistik ist ein schwieriges Fach, sowohl für Ärzte als auch für Mathematiker. Nachdem schon mehrere hundert Jahre mit Wahrscheinlichkeiten gerechnet worden war, hat erst 1933 Andrey Nikolayevich Kolmogorov seine axiomatische Definition der Wahrscheinlichkeit publiziert. Wenn in der Mathematik etwas so lange dauert, dann kann es nicht ganz einfach sein. Kolmogorov war Russe, hat aber in deutscher Sprache im Springer-Verlag Berlin unter dem völlig harmlosen Titel „Grundbegriffe der Wahrscheinlichkeitsrechnung" publiziert.

• Was ist ein Orakel?

Zunächst, es ist kein Horoskop, obwohl - nur nebenbei bemerkt - die Astrologie heute fast ausschließlich mit statistischen Verfahren arbeitet. Zum Beispiel interessiert die Frage, ob die Dauer und das Ende einer Ehe - ob durch Tod oder Scheidung - von der Kombination der Sternzeichen der Ehegatten abhängt. Diese Frage lässt sich mit Standesamtsdaten bearbeiten. Gegeben ist das Geburtsdatum des Mannes und der Frau, damit sind auch die Sternzeichen festgelegt, man kennt das Datum der Eheschließung und man kennt das Ende der Ehe, sei es nun durch Tod oder durch Scheidung. Damit lässt sich sehr wohl mit statistischen Verfahren ausrechnen, ob gewisse Kombinationen von Sternzeichen länger halten oder nicht. Falls es Sie interessiert, es besteht kein Zusammenhang.

Orakel möchte ich hier verstehen als Weissagung, als Vorhersage, und mit Prognosemodellen sind wir schon mitten in der Statistik. Nachher will ich auf zwei Orakel eingehen. Einmal war es Kroisos, König in Lydien, als er überlegte, ob er die Perser angreifen solle: Er befragte die Orakel von Delphi und Abai, wie ein solcher Krieg ausgehen würde. Vielleicht ist Ihnen auch MacBeth von Shakespeare bekannt. Die meisten Orakel haben sich im Nachhinein durchaus als richtig erwiesen, allerdings hatten die Orakel auch eine überraschende Komponente, wenn ich das so sagen darf.

• Anmerkung

Im Folgenden wird immer wieder von Begebenheiten zwischen Klinikern und Statistikern berichtet, fast so wie von Tünnes & Schäl zu Köln. Und da wird es fast immer so sein, dass der Statistiker der Schlaue, der Kluge ist und der Kliniker nur zweiter Sieger. Wenn Sie das verallgemeinern würden, würden Sie einen massiven Fehler machen. Die folgenden Beispiele sind alles andere als repräsentativ, weil wir sie aus biometrischer, nicht aus medizinischer Sicht betrachten.

• Die meisten Menschen sterben im Bett und unter ärztlicher Betreuung.

Wenn Sie also nicht sterben wollen, können Sie direkt den Schluss ziehen: „Gehe nicht ins Bett und vertraue dich nicht den Ärzten an." Ob dieser Schluss berechtigt

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

ist, darüber können wir zu einem späteren Zeitpunkt noch einmal philosophieren.

• Beispiel Operationsrisiko

Vor Jahren haben wir eine sehr detaillierte Operationen-Dokumentation durchgeführt. Unter anderem wurden erfasst: der Name des Operateurs, die Art der Operation, eingetretene Komplikationen und all diese Dinge. Bei der Auswertung mussten wir feststellen, dass die Komplikationsrate dann am kleinsten ist, wenn ein Anfänger in der Facharztausbildung operiert. Warum? Nun, ein Anfänger darf nur die leichten Fälle operieren, bei schwierigen Fällen müssen die Oberärzte oder der Chef ran und bei komplizierten Fällen ist die Komplikationsrate eben höher. Den Nichtmedizinern kann ich überzeugt empfehlen, sollten Sie einmal einen Leistenbruch haben oder eine akute Blinddarmentzündung und der Operateur stellt sich bei Ihnen vor und sagt: „Guten Tag, Sie sind mein erster Patient, den ich operieren werde.", dann seien Sie guten Mutes! Die Komplikationsrate ist minimal.

• Beispiel perinatale Todesrate

Die perinatale Todesrate ist wohldefiniert als der Tod des Kindes zwischen der vollendeten 22. Schwangerschaftswoche und dem 7. vollendeten Tag post partum. Diese WHO-Definition wird ziemlich einheitlich angewandt. Es ist eine Binsenweisheit, dass die perinatale Todesrate in Universitäts-Frauenkliniken höher ist als in Kreis-Krankenhäusern. Aber natürlich nicht, weil Universitätsklinika eine schlechte Geburtshilfe leisten, sondern weil alle Risikoschwangerschaften in eine Klinik der Maximalversorgung verlegt werden. Schon an diesen einfachen Beispielen können Sie sehen, die Statistik ist nicht ganz ohne Tücken.

• Beispiel Suizidgefährdung

Angenommen, ein Jugendpsychiater möchte Maßnahmen gegen die Suizidgefahr bei Jugendlichen erforschen, dazu einen Förderantrag stellen und würde mich fragen: „Haben Sie eine Statistik, die die Suizidgefährdung Jugendlicher zeigt?" Ich würde ihm Abbildung 1 mit dem Titel **„Traurige Jugend"** geben. Die Abbildung zeigt ein höheres Suizidrisiko bei jungen Männern als bei jungen Frauen. Bei Männern liegt der Gipfel mit 30 Jahren etwas später als bei Frauen mit 20 Jahren, aber es ist ja oft so, dass die Männer etwas später dran sind als die Frauen. Das Bild ist überzeugend, denke ich, die Daten sind von 2002 und damit aktuell, sie stammen vom Statistischen Landesamt Baden-Württemberg, eine solide Quelle.

Wenn nun jemand aus einem Psychiatrischen Landeskrankenhaus käme und sagte: „Wir wollen etwas tun gegen Suizide. Wie groß ist das Suizidrisiko bei Erwachsenen?" Jetzt würde ich Abbildung 2 liefern. Sie sehen auch hier einen minimalen Unterschied zwischen Männern und Frauen, aber jetzt ist der Gipfel doch deutlich auf ein höheres Alter verschoben. Deshalb habe ich diesem Bild den Titel **„Midlife Crisis"** gegeben.

Möchte jemand aus der Geriatrie ein Projekt gegen Suizide einwerben und möchte die Notwendigkeit von Maßnahmen gegen den Suizid alter Menschen darstellen, so würde ich ihm Abbildung 3 präsentieren. Sie sehen, auch hier sind Männer durchgehend deutlich gefährdeter als die Frauen. Das Suizidrisiko nimmt mit dem Alter zu, deshalb der Titel **„Elendes Alter"**. Jetzt, zu Beginn meiner Pension, trifft mich dies, aber lassen wir das mal.

In Abbildung 4 sind die 3 Kurven übereinander gezeichnet. Offensichtlich widersprechen sich die Kurven und Sie fragen sich: „Ist die Jugend so traurig, ist die Midlife-Crisis so schlimm oder das Alter so elend?" Damit sind Sie in der Situation wie die Frau des Rabbiners in der Geschichte von Salcia Landmann. Falls Sie die Geschichte nicht kennen, sie geht folgendermaßen:

Zwei Juden hatten miteinander Streit. Einer der beiden geht zum Rabbiner und trägt den Streit vor. Der Rabbiner denkt längere Zeit darüber nach und entscheidet dann: „Du hast Recht." Einige Tage später kommt der Kontrahent in der gleichen Sache zum gleichen Rabbiner und trägt den Streit (vielleicht aus seiner Sicht) vor. Und wieder klärt, so heißt tiefes Nachdenken im Jiddischen, der Rebbe lange und sagt dann: „Du hast Recht." Dann kommt die Frau des Rabbiners und sagt: „Rebbe-Leben! Hat nun der eine Recht oder der andere?"

In dieser Situation sind Sie jetzt auch. Wenn Sie mich jetzt fragen würden „Ist das Risiko für Suizid in der Jugend, in der Mitte des Lebens oder im Alter am größten? Verflixt, es kann doch nicht zu allen drei Zeiten am größten sein." dann würde ich antworten wie jener Rebbe, der zu seiner Frau sagte: *„Du hast auch Recht."*

Zur Wiederholung: Alle drei Abbildungen beziehen sich auf 2002. Alle beruhen auf Daten des Statistischen Landesamtes Baden-Württemberg, eine durchaus solide Datenquelle. Warum diese Widersprüche? Hätte ich Ihnen nur eine der Abbildungen - egal welche - gezeigt, so hätten Sie mir die Aussage vermutlich ganz gut abgenommen und wären zur Tagesordnung übergegangen. Aber welches dieser drei Bilder ist nun richtig?

Letztendlich sind unterschiedliche Sachverhalte dargestellt. In der Abbildung 1 **„Traurige Jugend"** ist die

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp(M2005081-1)

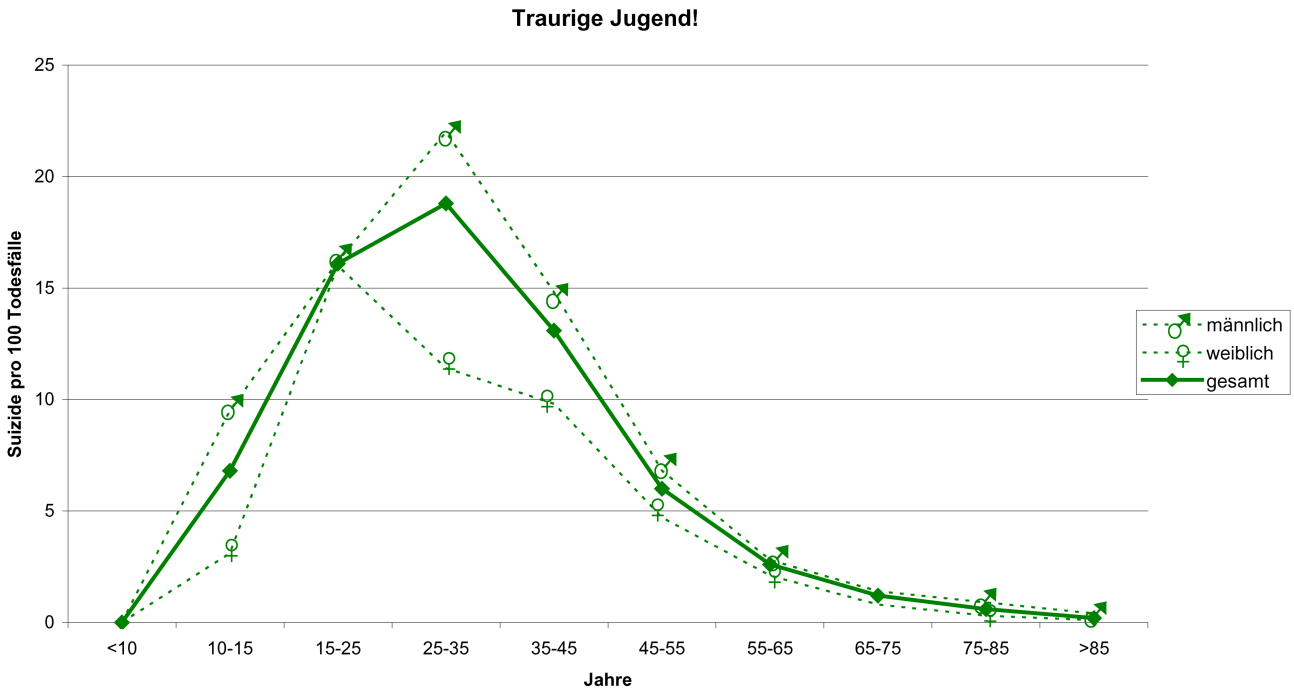

Abbildung 1: Suizide je 100 Todesfälle nach Altersgruppe und Geschlecht in Baden-Württemberg 2002 (Quelle: Statistisches Landesamt Baden-Württemberg)

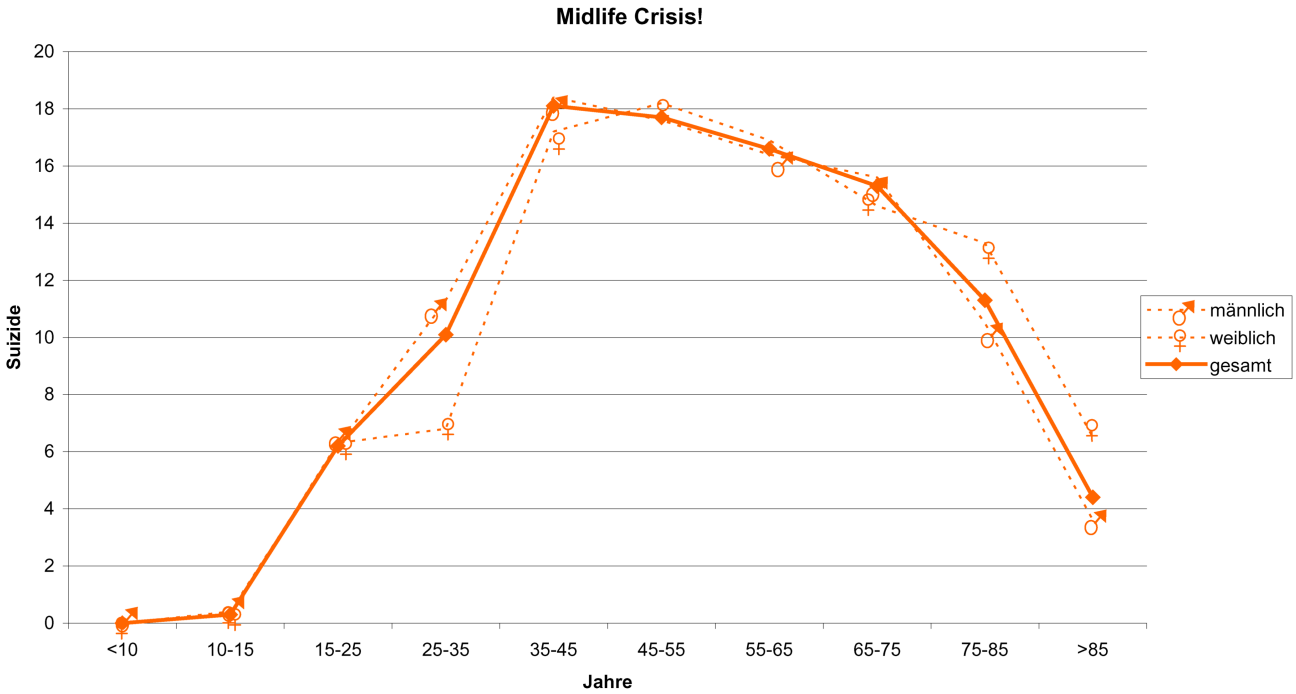

Abbildung 2: Von 100 Suiziden entfielen in Baden-Württemberg 2002 auf die betroffene Altersklasse ... (Quelle: Statistisches Landesamt Baden-Württemberg)

Anzahl der Suizide geteilt durch die Anzahl der Todesfälle. Im Alter zwischen 15 und 35 Jahren sind die Kinderkrankheiten vorbei, die Herz-Kreislauf-Krankheiten und die Tumoren sind noch nicht relevant und wer nicht gerade Motorrad fährt, der stirbt im Alter zwischen 15 und 35 nicht. Damit sind die Suizidtodesfälle relativ häufig, einfach, weil es wenig andere Todesursachen gibt. Das Minimum an anderen Todesursachen ist der Grund, warum beim Anteil der Suizide an den Todesfällen ein Gipfel in der Jugend entsteht.

Wie entsteht der Suizidgipfel in der Abbildung 2 **„Midlife Crisis"**? Dargestellt ist die Altersverteilung der Suizidanden. Diese lässt sich nur beurteilen, wenn gleichzeitig die Altersverteilung der Bevölkerung bekannt ist. Erkenntnisse zum Suizidrisiko kann man nur

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

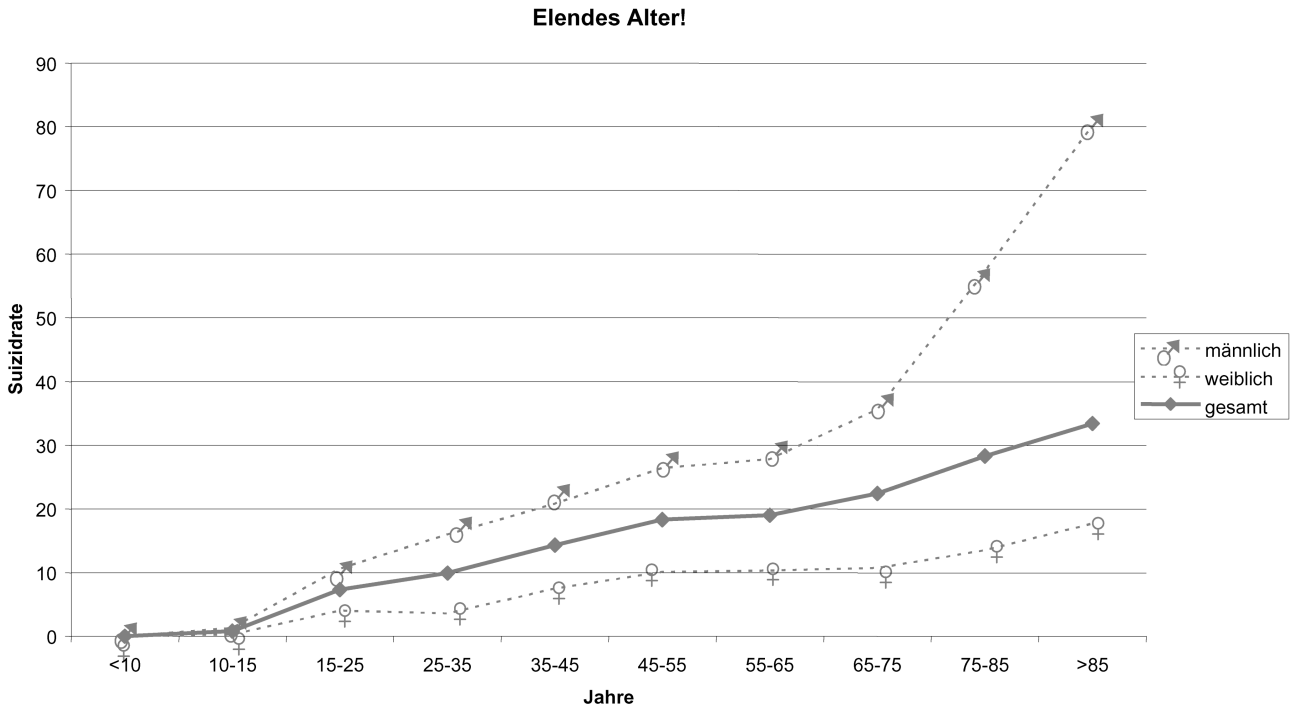

Abbildung 3: Suizide pro 100 000 Lebende der mittleren Bevölkerung in Baden-Württemberg 2002 (Quelle: Statistisches Landesamt Baden-Württemberg)

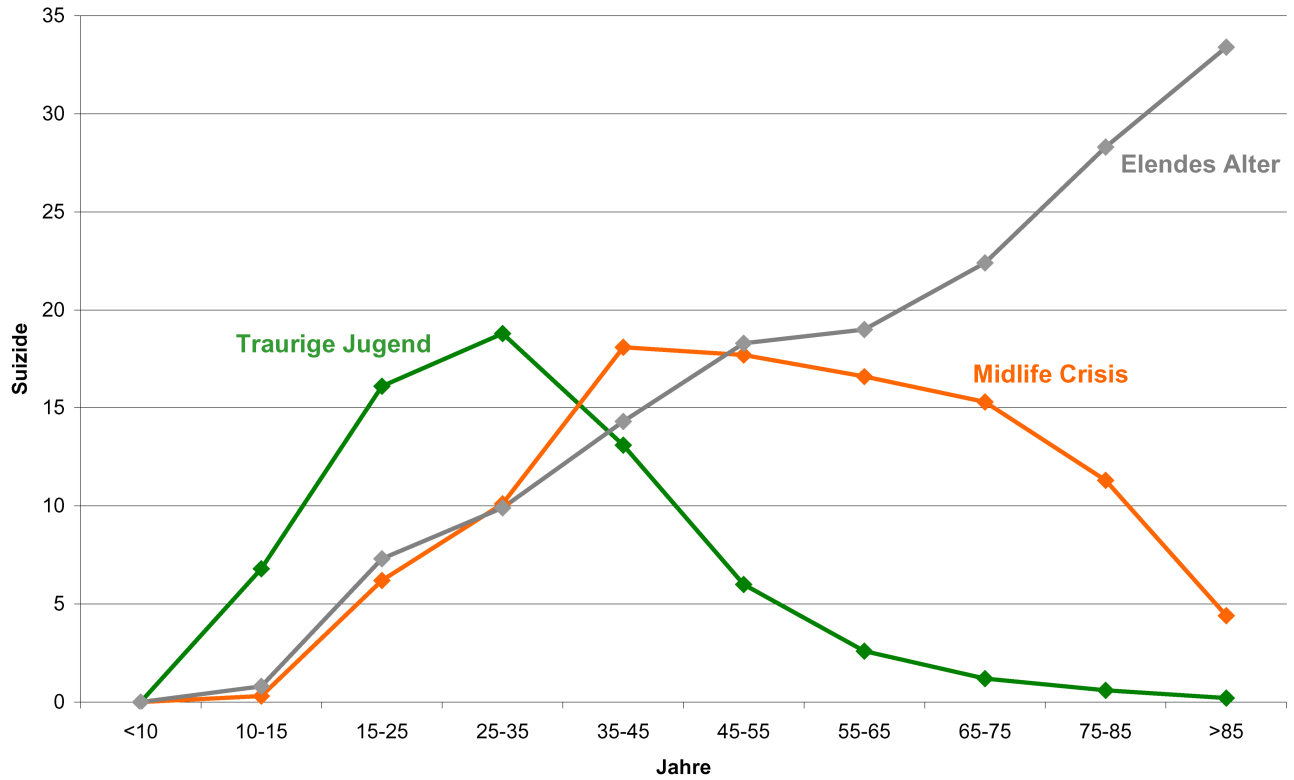

Abbildung 4: Suizide in Baden-Württemberg 2002 (Quelle: Statistisches Landesamt Baden-Württemberg)

aus den Unterschieden der Altersverteilung der Suizidanden (Abbildung 2) und der Altersverteilung der Bevölkerung (nicht dargestellt) gewinnen.

In der Abbildung 3 **„Elendes Alter"** ist nun die Anzahl der Suizide auf die Lebenden bezogen. In der Epidemiologie ist es durchaus gängig, dass Risiken auf die unter Risiko Stehenden bezogen werden. Nur wer lebt,

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

steht unter dem Risiko, einen Selbstmord zu begehen, und ich denke, das ist die richtige Abbildung.

Sie sehen, was mir als Pensionär bevorsteht. Ich tröste mich damit, dass Abbildung 3 nur eine von mehreren sinnvollen Darstellungen ist. Würden die durch Suizid verlorenen Lebensjahre dargestellt, so würde sich ein völlig anders Bild ergeben. Nochmals ein anderes Bild würde entstehen, wenn die durch Suizid verlorenen Lebensjahre mit der Lebensqualität gewichtet würden, also die Quality Adjusted Life Years (QUALYs) dargestellt würden.

• Der statistische Test trifft Entscheidungen

Mit der Statistik treffen wir auch **Entscheidungen**. Letztendlich wird mit statistischen Verfahren entschieden, ob eine Therapie wirksam ist, ob eine neue Therapie besser ist als die bisherige, ob die eine operative Vorgehensweise weniger Komplikationen bewirkt als die andere. Die Statistiker treffen diese Entscheidungen mit einem **statistischen Test**.

• Nullhypothese und Alternative

Traditionell wird das, was ein statistischer Test entscheidet, als **Nullhypothese** und deren **Alternative** formuliert.

Die Nullhypothese H_0 behauptet das Unerwünschte

typisch: Neue Behandlung ist auch nicht besser als die bisherige - Effekt einer Behandlung ist null - Kein Wirksamkeitsunterschied zwischen A und B

Die Alternative H_1 behauptet das Erwünschte

typisch: Neue Behandlung ist erfolgreicher als die bisherige - Behandlung hat einen Effekt - Es besteht ein Wirksamkeitsunterschied zwischen A und B

• Der statistische Test ist eine Wahrscheinlichkeitsrechnung

Berechnet wird die Wahrscheinlichkeit, dass die tatsächlich beobachteten Daten rein zufällig unter der Nullhypothese entstanden sind. Das ist ein Gedankengang, wie Sie ihn alle im Alter von etwa 12 Jahren nachvollzogen haben, als Sie in die Algebra eingeführt wurden. Man tut so, als würden wir das, was wir suchen, schon kennen und bezeichnet es mit x oder mit irgendeinem anderen Buchstaben. Dann rechnen wir damit und irgendwann einmal - wenn alles gut geht - kommt dann heraus x = soundso viel. Der statistische Test hat den gleichen Gedankengang. Angenommen, die Nullhypothese wäre richtig, wie groß ist dann die Wahrscheinlichkeit, dass die tatsächlich beobachteten Daten (oder ein noch extremeres Ergebnis) rein zufällig unter der Nullhypothese entstanden sein können? Diese berechnete Wahrscheinlichkeit heißt **Überschreitungswahrscheinlichkeit**, im Englischen kurz „p-value". Die Bezeichnung „Überschreitungs"-Wahrscheinlichkeit zeigt an, dass die tatsächlich beobachteten Daten als Grenze betrachtet werden, die hin zu einem noch extremeren Ergebnis überschritten wird.

Überschreitungs-Wahrscheinlichkeit (p-Wert) = Wsk (tatsächlich beobachtete Daten oder ein noch extremeres Ergebnis | H_0)

Wenn dieser p-Wert, diese Überschreitungswahrscheinlichkeit sehr klein ist, kleiner als eine vorab gewählte Grenze, die man als **Signifikanzniveau** bezeichnet, dann glaubt man nicht an den Zufall, sondern nimmt an, dass die Voraussetzung, die man in die Rechnung hineingesteckt hat, nämlich die Nullhypothese, falsch ist. Damit hat man die unerwünschte Nullhypothese als falsch - zumindest als unwahrscheinlich - erkannt und nimmt die erwünschte Alternative als richtig an. Das Signifikanzniveau - also die Grenze, ab der man nicht mehr an den Zufall glauben will - wird gewählt. In der Medizin wird meistens 5% gewählt.

Es gibt eine **Legende, wie der statistische Test entstanden ist**: Ein Mathematiker ging auf die Spielbank. Er spielte und hat in relativ wenigen Spielen relativ viel Geld verloren. So etwas soll auf der Spielbank passieren. Die wenigen Spiele und der große Verlust haben ihn geärgert. Das kann ich gut verstehen. Wieder zuhause hat er sich hingesetzt und berechnet, wie groß die Wahrscheinlichkeit ist, dass man in so wenigen Spielen, wie er gespielt hat, so viel Geld verliert, wie er verloren hat. Diese Wahrscheinlichkeit hat er unter der Annahme, unter der Nullhypothese berechnet, dass die Spielbank fair spielt. Die von ihm errechnete Wahrscheinlichkeit war sehr klein und er sagte zu sich selbst: „Würde die Spielbank fair spielen, so hätte ich - von extremem Pech abgesehen - in so wenigen Spielen nicht so viel verloren. Also betrügt die Spielbank." So ist, habe ich gehört, der statistische Test entstanden.

• Testergebnis

Das Ergebnis eines statistischen Tests heißt entweder signifikant oder nicht signifikant. Auch die Kliniker kennen diese Ausdrücke. Signifikant bedeutet man ist glücklich, die Alternative wird angenommen, das Erwünschte trifft zu, so behauptet wenigstens der Test. Wenn der Test nicht signifikant ist, konnte das Erwünschte leider nicht nachgewiesen werden. Falls der Test genügend Power hatte, kann man auch behaupten, das Erwünschte trifft nicht zu, zumindest ist das Erwünschte schwächer, als wir bei der Fallzahlplanung angenommen hatten. Die Power hat direkt etwas mit

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp0(2005)014.1)

der Fallzahl zu tun. Haben wir viele Fälle in unserer Statistik, haben wir eine große Power, dann können wir Unterschiede - falls sie wirklich vorhanden sind - auch entdecken. Haben wir nur wenige Fälle, na ja, dann haben wir auch nur wenig Information, wenig Power und dann können wir auch nicht so richtig entscheiden.

Ergebnis eines statistischen Tests:

signifikant: H_1 wird angenommen - Das Erwünschte trifft zu (behauptet der Test)

nicht sign.: H_0 wird beibehalten - Das Erwünschte konnte leider nicht nachgewiesen werden. Falls der Test genügend Power hatte: Das Erwünschte trifft nicht zu, es ist kleiner als bei der Fallzahlberechnung angenommen (behauptet der Test)

• Testfehler

Jeder, der Entscheidungen trifft, macht gelegentlich Fehler, auch ein statistischer Test. Es gibt 4 Möglichkeiten, die in Tabelle 1 gezeigt werden.

Tabelle 1: Testfehler

Wirklichkeit	Testergebnis	Facit
• Das Erwünschte (H_1) stimmt	signifikant	richtige Entscheidung
• Das Unerwünschte (H_0) stimmt	nicht sign.	richtige Entscheidung
• Das Unerwünschte (H_0) stimmt	signifikant	α-Fehler
• Das Erwünschte (H_1) stimmt	nicht sign.	β-Fehler

• α-Risiko und Power

Natürlich gehen wir jetzt weiter und beschäftigen uns nicht nur mit den Fehlern, sondern auch mit den Wahrscheinlichkeiten, mit denen diese Fehler auftreten. Die Wahrscheinlichkeit für einen α-Fehler heißt einfach **α-Risiko**. Das α-Risiko ist die Wahrscheinlichkeit, dass der Test (falsch) signifikant wird, unter der Bedingung, dass in Wirklichkeit die Nullhypothese richtig ist. Das ist aber genau dieser p-Wert, diese Überschreitungswahrscheinlichkeit, die wir ausgerechnet haben, und diese Wahrscheinlichkeit ist ja kleiner als das von uns gewählte Signifikanzniveau. Wurde als Signifikanzniveau 5% gewählt und der Test ist signifikant geworden, so ist das α-Risiko höchstens diese 5%.

α-Risiko = Wsk (α-Fehler) = Wsk (Test signifikant | unerwünschte H_0 stimmt) = errechneter p-Wert ≤ gewähltes Signifikanzniveau (meist wird 5% gewählt)

Nun zum **β-Risiko**. Wenn das Erwünschte gilt und der Test dies nicht merkt, dann hat er einen β-Fehler gemacht, die Wahrscheinlichkeit dafür heißt β-Risiko. Meist wird aber nicht das β-Risiko, sondern das Gegenteil, die **Power** betrachtet. Man kann ja sagen, das Glas ist zu 20% leer oder es ist zu 80% voll.

β-Risiko = Wsk (β-Fehler) = Wsk (Test nicht sign. | erwünschte H_1 stimmt)

Power (1 - β) = Wsk (Test signifikant | erwünschte H_1 stimmt)

Die Berechnung der Power ist schwieriger als die Berechnung des α-Risikos, aber unter bestimmten Bedingungen möglich.

Bei der Planung einer Studie wird nicht nur das Signifikanzniveau gewählt, sondern auch die Power. Meistens wird 80% Power gewählt, d.h. wenn das Erwünschte in dem angenommenen Ausmaß wirklich zutrifft, so ist die Wahrscheinlichkeit, dass der Test dies erkennt und ein signifikantes Ergebnis liefert, 80%. Eine Power von 80% hört sich nicht sehr groß an. Wird aber mehr Power gewünscht, dann sind mehr Fälle notwendig, und mehr Fälle, das macht den Klinikern schon manchmal Schwierigkeiten.

• Interpretation des Testergebnisses

Wenn eine klinische Studie abgeschlossen ist, man hat zwei Therapien verglichen, der Statistiker rechnet den Test und hurra, es kommt signifikant heraus, dann sagt man einfach, das Erwünschte, das was wir nachweisen wollten, ist richtig. Zumindest behauptet der Test das, und man sagt dann auch, die Wahrscheinlichkeit, dass wir uns irren, ist klein. Schon jetzt möchte ich die Frage stellen: Wie klein?

Wenn der Test nicht signifikant geworden ist, dann gibt es diese weiche Formulierung: Die beobachteten Daten können unter der Nullhypothese entstanden sein. Wohlgemerkt, sie können, sie müssen aber nicht! Das Unerwünschte kann richtig sein, das Erwünschte konnte leider nicht nachgewiesen werden. Nur falls der Test genügend Power hatte, falls wir genügend Fälle in unsere Studie aufgenommen hatten, dann können wir bei einem nicht signifikantem Test sagen: Das Erwünschte trifft nicht zu, zumindest ist sein Effekt kleiner als wir dachten. Und auch hier ist die Wahrscheinlichkeit, dass wir uns irren, klein. Bald werden wir der Frage nachgehen: Wie klein?

signifikant: Das Erwünschte (die H_1) ist richtig (behauptet der Test). - Die Wahrscheinlichkeit, dass wir uns irren, ist klein. - **Wie klein?**

nicht sign.: Die beobachteten Daten können unter H_0 entstanden sein. - Das Unerwünschte (die H_0) kann

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp(M2005081411)

richtig sein. - Das Erwünschte (die H_1) konnte leider nicht nachgewiesen werden. Falls der Test genügend Power hatte: Das Erwünschte trifft nicht zu, es ist kleiner als bei der Fallzahlberechnung angenommen (behauptet der Test). - Die Wahrscheinlichkeit, dass wir uns irren, ist klein. - **Wie klein?**

• Ist das Testergebnis richtig?

Wenn der Test signifikant geworden ist, wie groß ist dann, so frage ich jetzt, die Wahrscheinlichkeit, dass tatsächlich das Erwünschte gilt? Ist aber der Test nicht signifikant geworden, wie groß ist dann die Wahrscheinlichkeit, dass in Wirklichkeit das Unerwünschte gilt? Diese Fragen sind für jemand, der das Testergebnis hat, viel relevanter als α-Risiko und Power. Bei diesen Fragen ist gegenüber der Power und dem Komplement des α-Risikos die Bedingung und das Ereignis vertauscht.

Wsk (erwünschte H_1 stimmt │ Test signifikant) = ?

Zum Vergleich: Power = Wsk (Test signifikant │ H_1 stimmt)

Wsk (unerwünschte H_0 stimmt │ Test nicht sign.) = ?

Zum Vergleich: (1 - α-Risiko) = Wsk (Test nicht sign. │ H_0 stimmt)

• Vergleich: Diagnostisches Verfahren - Statistischer Test

Ein ähnliches Problem haben die Kliniker, wenn sie die **Zuverlässigkeit eines diagnostischen Verfahrens** beurteilen. Deshalb jetzt eine kleine Exkursion über die Zuverlässigkeit diagnostischer Verfahren: Bei diagnostischen Verfahren unterscheiden wir **Sensitivität** und **Spezifität**. Ein diagnostisches Verfahren ist sensitiv, hat eine hohe Sensitivität, Empfindlichkeit, wenn es alle Erkrankten erkennt. Sensitivität ist die Wahrscheinlichkeit, dass der Befund die Krankheit anzeigt, unter der Bedingung, dass der Untersuchte die gesuchte Krankheit tatsächlich hat. Ein diagnostisches Verfahren ist für eine bestimmte Krankheit spezifisch, hat eine hohe Spezifität, wenn es nur auf die eine, nur auf die gesuchte Krankheit anspricht, nicht aber auf andere, vielleicht ähnliche Krankheiten. Fieber ist z.B. ein unspezifischer Befund, weil viele und sehr verschiedene Krankheiten mit Fieber einhergehen.

Sensititvität = Wsk (Befund pathologisch | Patient hat gesuchte Krankheit)

Spezifität = Wsk (Befund normal | Patient hat gesuchte Krankheit nicht, aber vielleicht eine andere)

Sensitivität und Spezifität sind für die Wissenschaft interessant, der praktisch tätige Arzt will jedoch das Umgekehrte wissen, nämlich: Wenn ein Befund pathologisch ist, wie groß ist dann die Wahrscheinlichkeit, dass der Patient tatsächlich die gesuchte Krankheit hat? Der Fachausdruck dafür heißt **positive Prädiktion**. Ist ein Befund unauffällig, normal, so will der praktisch tätige Arzt wissen: Wie groß ist die Wahrscheinlichkeit, dass der Patient die Krankheit tatsächlich nicht hat? Diese bedingte Wahrscheinlichkeit heißt die **negative Prädiktion**.

positive Prädiktion = Wsk (Patient hat gesuchte Krankheit │ Befund pathologisch)

negative Prädiktion = Wsk (Patient hat gesuchte Krankheit nicht │ Befund normal)

Jetzt verbinde ich den statistischen Test und das diagnostisches Verfahren. Ich sage einfach - und das ist für die Kliniker gedacht - ein statistischer Test ist gar nichts anderes als ein diagnostisches Verfahren. So wie ein diagnostisches Verfahren herausfindet, ob der untersuchte Patient die gesuchte Krankheit hat oder nicht hat, so findet der statistische Test heraus, ob in dem untersuchten Datenmaterial das Erwünschte oder das Unerwünschte gilt. Gilt das Erwünschte, dann soll der Test signifikant sein, gilt das Unerwünschte, dann soll der Test nicht signifikant sein.

Ein **diagnostisches Verfahren** soll herausfinden, ob die untersuchte Person

eine bestimmte Krankheit hat → pathologischer Befund

oder diese Krankheit nicht hat → normaler Befund

Ein **statistischer Test** soll herausfinden, ob für die untersuchten Daten

das Erwünschte (die H_1) gilt → signifikant

oder das Unerwünschte (die H_0) → nicht signifikant

Die Kliniker wissen sehr gut, dass die positive und negative Prädiktion von der Häufigkeit, der **Prävalenz** der Krankheit abhängt. Wenn eine Krankheit sehr häufig ist, dann ist es leicht, eine gute positive Prädiktion zu erreichen. Ist eine Krankheit selten, dann ist es einfach, eine gute negative Prädiktion zu erreichen. In Abbildung 5 ist links die Prävalenz dargestellt, d.h. die Wahrscheinlichkeit, dass der Untersuchte die gesuchte Krankheit hat, bevor er untersucht wurde. Das diagnostische Verfahren kann einen pathologischen (= krankhaften) oder einen normalen (= unauffälligen) Befund liefern. Auf der rechten Seite ist die Wahrscheinlichkeit, dass dieser Befund richtig ist, darge-

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

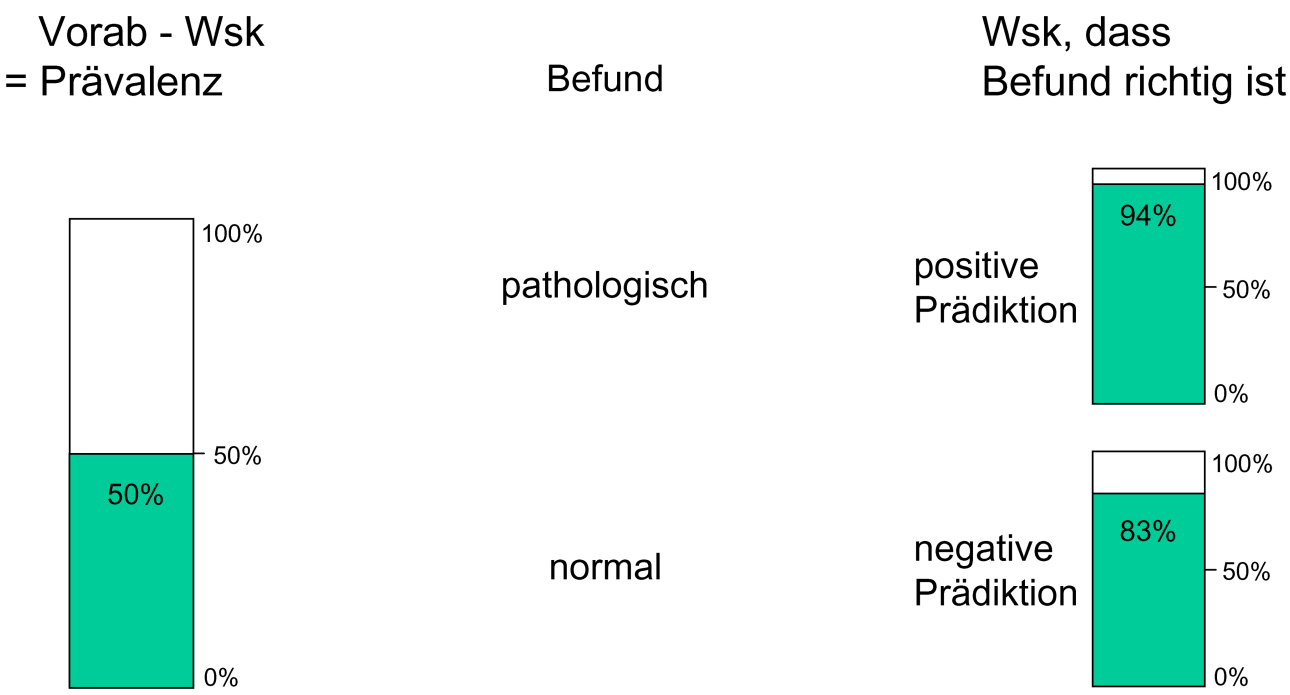

Abbildung 5: Diagnostisches Verfahren

stellt. Wie gesagt, die Fachausdrücke heißen positive Prädiktion und negative Prädiktion. Damit es nicht zu kompliziert wird, wird für Abbildung 5 und Tabelle 2 durchgehend angenommen, dass das diagnostische Verfahren eine Sensitivität von 80% hat, also die, die wirklich krank sind, werden zu 80% erkannt, und eine Spezifität von 95% besitzt, das heißt, die Personen, welche die gesuchte Krankheit nicht (aber vielleicht irgend eine andere Krankheit) haben, werden zu 95% richtig erkannt. Abbildung 5 zeigt: Wenn jede zweite untersuchte Person die zu diagnostizierende Krankheit hat (Prävalenz = 50%), dann sind 94% der pathologischen Befunde und 83% der normalen Befunde richtig. Die etwas unterschiedlichen Werte für positive und negative Prädiktion entstehen trotz 50% Prävalenz, weil Sensitivität und Spezifität etwas unterschiedlich gewählt wurden.

In Tabelle 2 wird die Prävalenz sukzessiv kleiner und Sie sehen, wie die Wahrscheinlichkeiten, dass die Befunde richtig sind, sich verändern. Die letzte Zeile der Tabelle 2 zeigt die Situation mit nur noch 1% Vorabwahrscheinlichkeit. Wir betrachten also eine relativ seltene Krankheit. Jeder 100. Patient, den die Kliniker auf diese Krankheit untersuchen, hat diese Krankheit. Wenn jetzt ein pathologischer Befund entsteht, dann ist die Richtigkeit dieses Befunds noch lumpige 14%. Umgekehrt, wenn die betrachtete Krankheit einigermaßen selten und der Befund unauffällig ist, dann ist es ziemlich sicher, dass der Patient diese Krankheit nicht hat. Die negative Prädiktion ist nicht - wie in Tabelle 2 angegeben - genau 100%, aber so nahe an 100%, dass wir, wenn wir korrekt auf ganze Prozent runden, auf 100% aufrunden.

Mit diesem Wissen kann man nette Spielchen treiben. Ich behaupte, ich sei ein phantastischer Diagnostiker und ich könne Ihrer Nasenspitze ansehen, dass Sie nicht HIV-positiv sind und weniger als eine Million Euro im Jahr verdienen. Bei diesen Beispielen ist die Prävalenz so klein, dass meine Aussage fast immer richtig ist, auch wenn meine Diagnostik gar nichts taugt.

• Wahrscheinlichkeit, dass ein Testergebnis richtig ist

Abbildung 6 und Tabelle 3 enthalten die gleichen Zahlen wie Abbildung 5 und Tabelle 2, unterscheiden sich aber in den Überschriften. Wir betrachten jetzt nicht ein diagnostisches Verfahren, sondern wir betrachten einen statistischen Test, und zwar einen

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

Tabelle 2: Bei einem diagnostischen Verfahren hängt die Richtigkeit der Befunde von der Prävalenz der gesuchten Krankheit ab. Für das diagnostische Verfahren wurde 80% Sensitivität und 95% Spezifität angenommen.

Vorab – Wsk = Prävalenz	Befund	Wsk, dass Befund richtig ist
70%	pathologisch	97% positive Prädiktion
	normal	67% negative Prädiktion
50%	pathologisch	94% positive Prädiktion
	normal	83% negative Prädiktion
40%	pathologisch	91% positive Prädiktion
	normal	88% negative Prädiktion
30%	pathologisch	87% positive Prädiktion
	normal	92% negative Prädiktion
20%	pathologisch	80% positive Prädiktion
	normal	95% negative Prädiktion
10%	pathologisch	64% positive Prädiktion
	normal	98% negative Prädiktion
5%	pathologisch	46% positive Prädiktion
	normal	99% negative Prädiktion
1%	pathologisch	14% positive Prädiktion
	normal	100% negative Prädiktion

Anmerkung: In der Vorlesung wurde anstatt dieser Tabelle eine Serie von 8 Bildern gezeigt, die alle wie Abb. 5 aufgebaut waren. Die Bildsequenz zeigte wie diese Tabelle, dass mit abnehmender Prävalenz die positive Prädiktion abnimmt und die negative Prädiktion zunimmt.

konfirmatorischen, also einen beweisenden Test. Anstatt 80% Sensitivität heißt es jetzt 80% Power, und anstatt 95% Spezifität heißt es jetzt 5% Signifikanzniveau. Das Wort Prävalenz muss ich weglassen da es nur mit Krankheiten gebräuchlich ist. Der statistische Fachausdruck für die Vorabwahrscheinlichkeit heißt **a priori Wahrscheinlichkeit**.

Abbildung 6 hat wie Abbildung 5 die a priori Wahrscheinlichkeit von 50%. Angenommen, jemand hat eine neue Therapie entwickelt und seine Chancen, dass diese Therapie seine Hoffnungen erfüllt, stehen 50:50. Dies ist beachtlich gut. Jetzt wird eine klinische Studie zum Wirksamkeitsnachweis dieser Therapie durchgeführt. Weiter angenommen, der Test wird signifikant. Unter diesen Annahmen ist die Wahrscheinlichkeit, dass diese Signifikanz richtig ist, nicht 100%, sondern 94%. Auch wenn der Test am Ende der Studie nicht signifikant wird, obwohl wir eine ausreichende Fallzahl hatten, um eine Power von 80% zu garantieren, dann ist die Wahrscheinlichkeit dafür, dass dieses „nicht signifikant" richtig ist, auch nur 83%.

In Tabelle 3 wird analog zu Tabelle 2 die a priori Wahrscheinlichkeit, die Vorabwahrscheinlichkeit, immer kleiner und Sie können beobachten, wie die Wahrscheinlichkeit, dass das Ergebnis richtig ist, sich verändert. In der letzten Zeile sind wir wieder bei der a priori Wahrscheinlichkeit von 1% angekommen. Vielleicht ist das ein bisschen wenig, etwas pessimistisch, aber manchmal ist es schon so, dass die Chance, eine bessere Therapie gefunden zu haben, vorab nicht dramatisch über 1% liegt. Nehmen wir mal an, vorab sei die Wahrscheinlichkeit, dass das erwünschte Ergebnis zutrifft, 1% gewesen, die Therapie wird in einer Studie erprobt, der Test wird signifikant, große

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp0(2005)014.1)

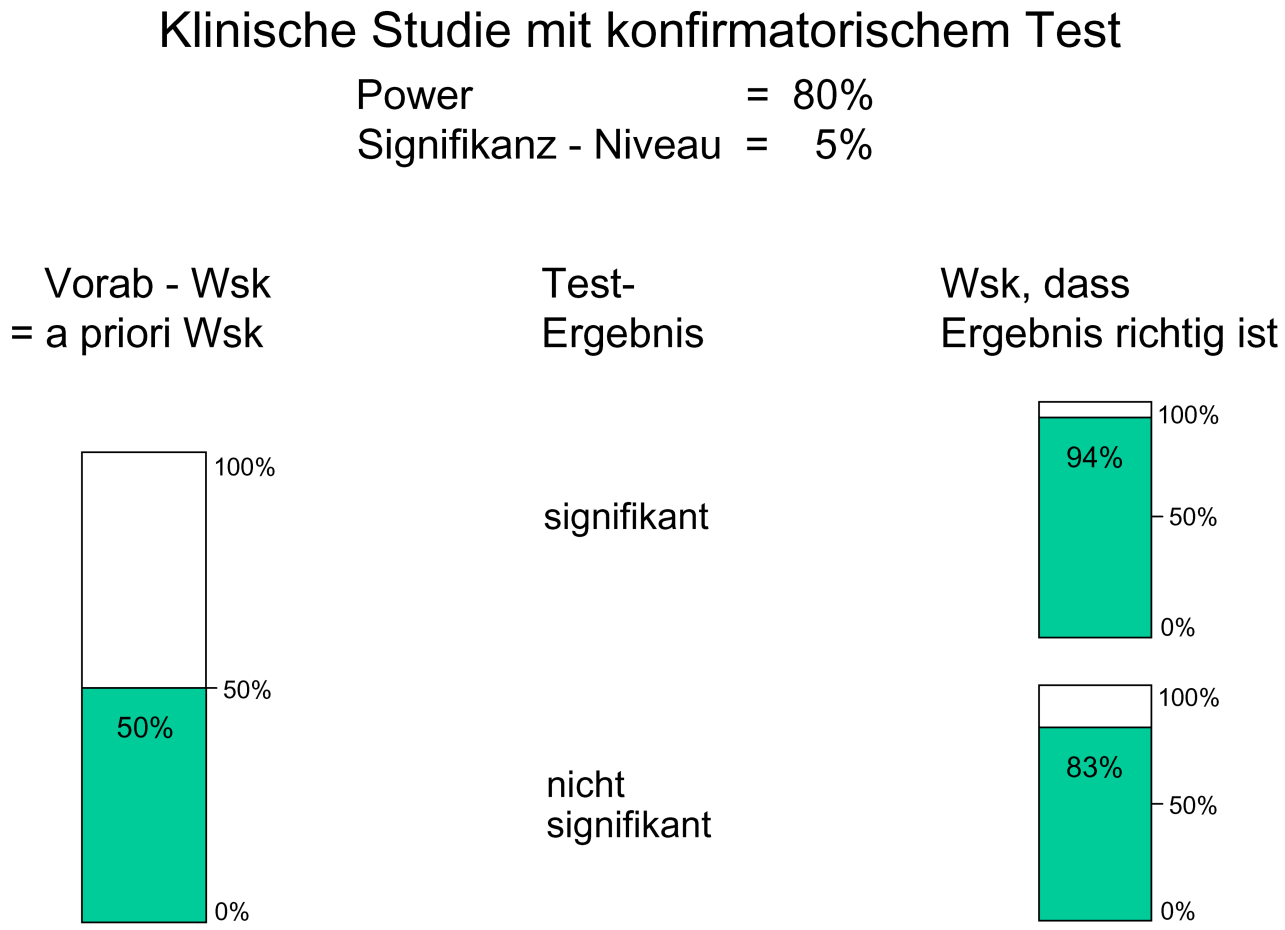

Abbildung 6: Klinische Studie mit konfirmatorischem Test

Freude. Jetzt muss der Statistiker schrecklich viel Wasser in den Wein gießen, diesen Sekt erheblich verdünnen und sagen: „Lieber Kliniker, die Wahrscheinlichkeit, dass der Test hier richtig entschieden hat, ist bescheidene 14%." Allerdings, wenn der Test nicht signifikant geworden ist, dann ist es in Verbindung mit der kleinen Vorabwahrscheinlichkeit ziemlich sicher, dass die neue Therapie auch nicht besser ist als die bisherige.

Das ist eine Botschaft, die ich Ihnen auf den Weg geben möchte: **Die Wahrscheinlichkeit für die Richtigkeit eines statistischen Testergebnisses hängt entscheidend davon ab, wie groß die Wahrscheinlichkeit vor Studienbeginn war.** Die Mathematik dazu ist harmlos. Trotzdem können Sie diesen Sachverhalt in keinem Lehrbuch nachlesen. Den meisten Leuten ist nicht bewusst: So wie bei einem diagnostischen Verfahren bei niedriger Prävalenz ein pathologischer Befund noch ziemlich unsicher ist, so unsicher ist bei geringer a priori Wahrscheinlichkeit ein signifikantes Testergebnis. Erst wenn ein weiteres diagnostisches Verfahren erneut einen pathologischen Befund liefert, erst wenn eine weitere klinische Studie ebenfalls zu einem signifikanten Ergebnis kommt, ist eine brauchbare Wahrscheinlichkeit erreicht.

• Wie viele Tests sind an einem Datenmaterial möglich?

Nun wende ich mich einem neuen Thema zu, dem Thema **Multiples Testen**. Was geschieht, wenn wir nicht einen Test rechnen, sondern wenn wir mehrere Tests durchführen? Dazu will ich Ihnen erst einmal ein **Datenmaterial** beschreiben.

Nehmen wir an, ein **Datenmaterial** besteht aus g **G**ruppen, an denen zu **t** Zeitpunkten m **M**erkmale bestimmt wurden.

Beispiel:

g = 2 Gruppen: Medikament A, Medikament B verordnet

t = 3 Zeitpunkte: präoperativ, intraoperativ, 24 h nach Op-Ende

m = 5 Merkmale: systolischer Blutdruck, diastolischer Blutdruck, Herzfrequenz, pO_2, pCO_2

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

Tabelle 3: Bei einem statistischen Test hängt die Richtigkeit des Testergenisses auch davon ab, wie wahrscheinlich die Alternative schon vor Studienbeginn war. Die Fallzahl ermöglichte eine Power von 80%, als Signifikanzniveau wurden 5% gewählt. Hinweis: Tab. 2 und 3 enthalten die gleichen Zahlen.

Vorab – Wsk = a priori Wsk	Test-Ergebnis	Wsk, dass Ergebnis richtig ist
70%	signifikant	97%
	nicht signifikant	67%
50%	signifikant	94%
	nicht signifikant	83%
40%	signifikant	91%
	nicht signifikant	88%
30%	signifikant	87%
	nicht signifikant	92%
20%	signifikant	80%
	nicht signifikant	95%
10%	signifikant	64%
	nicht signifikant	98%
5%	signifikant	46%
	nicht signifikant	99%
1%	signifikant	14%
	nicht signifikant	100%

Anmerkung: In der Vorlesung wurde anstatt dieser Tabelle eine Serie von 8 Bildern gezeigt, die alle wie Abb. 5 aufgebaut waren. Die Bildsequenz zeigte wie diese Tabelle, dass mit abnehmender Prävalenz die positive Prädiktion abnimmt und die negative Prädiktion zunimmt.

Wie viele **paarweise Vergleiche** sind an einem solchen Datenmaterial möglich? Zunächst kann jede Gruppe mit jeder anderen Gruppe verglichen werden. Bei 3 Gruppen kann man die erste mit der zweiten, die erste mit der dritten und die zweite mit der dritten Gruppe vergleichen. Das gibt g × (g - 1) mögliche Vergleiche. Weil aber der Vergleich der ersten mit der zweiten Gruppe dasselbe ist wie der Vergleich der zweiten mit der ersten Gruppe - der Vergleich A mit B ist dasselbe wie B mit A - kommt noch der Faktor ½ dazu. Diese Vergleiche zwischen den Gruppen sind zu jedem Zeitpunkt möglich, z.B. kann man die Gruppen präoperativ vergleichen, intraoperativ und postoperativ. Alle diese Vergleiche sind für jedes Merkmal möglich.

• Jede Gruppe lässt sich mit jeder anderen Gruppe zu jedem Zeitpunkt, bei jedem Merkmal vergleichen:
t m ½ g (g - 1) Vergleiche

Anstatt Gruppen lassen sich auch Zeitpunkte vergleichen, um zu sehen, ob sich im Lauf der Zeit etwas verändert. Nun kann jeder Zeitpunkt mit jedem anderen Zeitpunkt verglichen werden. Wiederum muss die Anzahl der Vergleiche halbiert werden, weil z.B. der Vergleich prä - post die gleichen Ergebnisse liefert wie der Vergleich post - prä. Damit sind wir bei ½ t (t - 1) möglichen Vergleichen zwischen Zeitpunkten. Diese Vergleiche sind möglich für jede Gruppe und für jedes Merkmal.

• Jeder Zeitpunkt lässt sich mit jedem anderen Zeitpunkt für jede Gruppe, bei jedem Merkmal vergleichen:
g m ½ t (t - 1) Vergleiche

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

Werden die Anzahl der Vergleiche zwischen den Gruppen und die Anzahl der Vergleiche zwischen den Zeitpunkten addiert, so ergibt sich nach etwas Vereinfachung:

Anzahl der an einem Datenmaterial möglichen paarweisen Vergleiche = ½ g t m (g + t - 2)

Beispiele:

g = 2, t = 1, m = 1 → 1 Vergleich möglich

g = 2, t = 1, m = 10 → 10 Vergleiche möglich

g = 2, t = 3, m = 10 → 90 Vergleiche möglich

g = 4, t = 5, m = 100 → 7000 Vergleiche möglich

Einige Anmerkungen zu den Beispielen: Wenn ich 2 Gruppen habe, 1 Zeitpunkt und 1 Merkmal, dann kann ich wirklich nur die 2 Gruppen und sonst nichts vergleichen, dann habe ich nur 1 Vergleichsmöglichkeit. Wurden 10 Merkmale erhoben, so ist dieser Vergleich der 2 Gruppen bei jedem der 10 Merkmale möglich. Wurde nicht zu einem Zeitpunkt, sondern an 3 Zeitpunkten beobachtet, jeweils die 2 Gruppen und 10 Merkmale, so ergeben sich schon ½ × 2 × 3 × 10 (2 + 3 - 2) = 90 Vergleichsmöglichkeiten. Hat jemand z.B. 4 Gruppen, 5 Zeitpunkte und 100 Merkmale, so kann er schon 7000 paarweise Vergleiche durchführen, eine ganz beachtliche Zahl. Trotzdem, 4 Gruppen, 5 Zeitpunkte und 100 Merkmale ist für klinische Verhältnisse ein kleiner Datenbestand. Einhundert Merkmale erfordern etwa 2 Seiten Datenerhebungsbogen. Manche Studien haben Case-Books, die über 100 Seiten dick sind, dann sind praktisch jede Menge Vergleiche möglich.

• Wahrscheinlichkeit, dass von mehreren Tests mindestens einer signifikant wird

Jetzt eine neue Frage: Wie groß ist die Wahrscheinlichkeit, dass von mehreren Tests mindestens einer signifikant wird? Dazu möchte ich mit Ihnen einen **Gedankenversuch** machen. Stellen Sie sich vor, wir lassen von einem **Zufallszahlengenerator** Zufallszahlen erzeugen. Als Zufallszahlengenerator nehmen wir ein Computerprogramm, ein Würfel wäre zu mühsam und würde auch nur ganze Zahlen von 1 bis 6 liefern. Wir lassen uns z.B. 2 × 80 reelle Zahlen geben, die wir als Werte der Zielgröße von 2 Gruppen zu je 80 Fällen verwenden. Da es reine Zufallszahlen sind, wissen wir definitiv, dass zwischen den Gruppen kein Unterschied besteht, dass die Nullhypothese gilt.

Gedankenversuch: Alle Daten sind mit dem Zufallszahlengenerator erzeugt - Bei allen Tests gilt definitionsgemäß H_0 - d.h. alle Signifikanzen sind falsch!

Mit diesen Daten rechnen wir einen statistischen Test. Wie groß ist die Chance, dass wir mit diesem Test eine Signifikanz erhalten? Das haben wir schon eingangs besprochen. Wenn kein Unterschied besteht, d.h. die unerwünschte Nullhypothese gilt, dann ist die Wahrscheinlichkeit, dass der Test falsch signifikant wird, das α-Risiko, das üblicherweise auf 5% begrenzt wird.

Wiederholung: Wsk (Test sign. $\mid H_0$) = α-Risiko, meist wird 5% gewählt

Dieses Gedankenspiel mit dem Zufallszahlengenerator und dem statistischen Test treiben wir öfters, sagen wir z.B. n = 100 mal. Nun wollen wir die Formel herleiten für die Wahrscheinlichkeit, dass mindestens einer dieser Tests signifikant wird. Dazu betrachten wir die Sache von der anderen Seite und sagen: Die Wahrscheinlichkeit, dass mindestens 1 Test signifikant wird, ist 1 minus die Wahrscheinlichkeit, dass kein Test signifikant wird.

Wsk (≥ 1 Test sign.) = 1 - Wsk (kein Test sign.)

Nun zerlegen wir weiter und sagen, wenn kein Test signifikant wird, dann ist der 1. Test nicht signifikant, der 2. Test nicht signifikant, der 3. Test nicht usw. bis zum letzten Test. Weil die Daten für jeden Test mit dem Zufallsgenerator erzeugt sind, sind sie unabhängig und man darf die Wahrscheinlichkeiten multiplizieren.

Wsk (≥ 1 Test sign.) = 1 - Wsk (1. Test nicht sign.) × Wsk (2. Test nicht sign.) × … ... × Wsk (n. Test nicht sign.)

Für jeden einzelnen Test ist die Wahrscheinlichkeit, dass er signifikant wird, das α-Risiko. Die Wahrscheinlichkeit, dass er nicht signifikant wird, ist das Gegenteil, nämlich 1 - α.

Wsk (≥ 1 Test sign.) = 1 - (1 - α) × (1 - α) × (1 - α) × …… × (1 - α)

Die Wahrscheinlichkeit, dass von n Tests mindestens einer signifikant wird, ist somit

Wsk (≥ 1 Test sign.) = 1 - (1 - α)^n

n = Anzahl der Tests

α = α-Risiko, meist wird 5% gewählt

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp(M2005081414-1)

Diese Formel gilt für Zufallszahlen, d.h. bei allen Tests gilt definitionsgemäß die Nullhypothese und alle Signifikanzen sind falsch.

• Wahrscheinlichkeit, dass ein Datenmaterial mindestens eine Signifikanz liefert

Nun fügen wir die beiden hergeleiteten Formeln zusammen. Wir setzen die Anzahl der an einem Datenmaterial möglichen Tests ein in die Formel für die Wahrscheinlichkeit, dass von mehreren Tests mindestens einer signifikant wird.

An dem Datenmaterial sind ½ g t m (g + t - 2) Tests möglich

Die Wahrscheinlichkeit, dass an einem Datenmaterial aus Zufallszahlen mindestens 1 Test (falsch) signifikant wird = $1 - (1 - \alpha)^{1/2\ g\ t\ m\ (g + t - 2)}$

Die Anzahl der an einem Datenmaterial möglichen Tests wird als Exponent verwendet. So große Zahlen im Exponenten: Sie wittern Gefahr. Zurecht! Als Beispiele sind in Tabelle 4 einige Varianten durchgespielt. Wenn man 2 Gruppen hat, 1 Zeitpunkt, 1 Merkmal, na ja, dann ist nur 1 Vergleich möglich und die Wahrscheinlichkeit, dass dieser eine Test signifikant wird, ist unser ganz gewöhnliches Signifikanzniveau, die üblichen 5%. Hat jemand aber nicht 1 Merkmal, sondern 10 Merkmale erhoben, so sind 10 Vergleiche möglich. Dann ist die Wahrscheinlichkeit, dass einer signifikant wird, schon deutlich größer, schon 40%. Wurde nicht nur zu einem Zeitpunkt beobachtet, sondern zu 3 Zeitpunkten, so gibt es schon 90 Vergleichsmöglichkeiten. Dann ist die Chance, zumindest eine Signifikanz zu erhalten, schon 99%.

Mit ein paar wenigen Gruppen, Zeitpunkten und Merkmalen erhält man praktisch immer eine Signifikanz! Dabei wir sind immer noch in dem Gedankenversuch, dass die Daten mit dem Zufallszahlengenerator erzeugt sind, dass absolut nichts drin ist und für alle diese Tests definitiv die Nullhypothese gilt. Tabelle 4 zeigt (1.) wie mit steigender Anzahl von Gruppen, Zeitpunkten und Merkmalen die Anzahl der möglichen Tests zunimmt, (2.) dass die Wahrscheinlichkeit, dass mindestens ein Test (falsch) signifikant wird, ebenfalls steigt und (3.) dass sich diese Wahrscheinlichkeit ziemlich schnell den 100% nähert. Das heißt, wenn jemand nur ein bisschen Fleiß entwickelt, erhält er immer etwas signifikantes, auch dann, wenn in Wirklichkeit überhaupt nichts in den Daten drin ist.

• Unabhängig erzeugte Hypothesen

Stellen Sie sich jetzt bitte folgende Situation vor: Ein Kliniker kommt zu uns in die statistische Beratung. Er hat zu 3 Zeitpunkten 2 Gruppen untersucht und jeweils 10 Merkmale erhoben. Der Kliniker ist fleißig, hat sich abends hingesetzt und schon mal die gängigen deskriptiven Statistiken Mittelwert, relative Häufigkeiten usw. berechnet, und hat an einer Stelle etwas Interessantes gefunden. Er kommt nun mit diesen deskriptiven Statistiken zu uns und sagt: „Lieber Statistiker, kannst du mal bitte nachprüfen, ob dieser Unterschied signifikant ist oder nicht." Der Statistiker sieht mit Augenmaß - die Kliniker haben ihren klinischen Blick, die Statistker ihren statistischen Blick - dass dieser gezeigte Unterschied isoliert betrachtet signifikant ist, aber er kennt auch das Problem des multiplen Testens. Er sagt: „Lieber Kliniker, das ist gar nicht so einfach. Wie ist diese Fragestellung entstanden? Hätten Sie die Frage gestellt, bevor Sie die Daten gesehen und bevor Sie die Mittelwerte und relativen Häufigkeiten berechnet haben, dann wäre es jetzt signifikant. Weil Sie aber die Daten, Mittelwerte und Häufigkeiten gesehen haben, deshalb ist es jetzt nicht mehr signifikant." Wenn ich das einem Kliniker sage, kann der eigentlich nur einen Schluss ziehen, nämlich: „Der Statistiker ist meschugge."

Trotzdem versuche ich jetzt dafür zu werben, dass der Statistiker doch Recht hat. Der Kliniker, den ich vor Ihnen aufbaue, hat die Mittelwerte, Häufigkeiten usw. angeschaut und hat dann ganz vernünftig entschieden: „Da ist nichts drin, dort ist nichts drin und da auch nichts." Im Grunde genommen hat er bei jedem Vergleich per Augenmaß einen Test gemacht und erkannt, dass der Unterschied zu klein ist um signifikant zu sein. Dass aber bei so vielen Tests, bei mehr oder weniger allen an diesem Datenmaterial möglichen Tests der eine oder andere signifikant wird, haben wir eben gesehen. Wenn einer von vielleicht hundert Tests signifikant wird, so ist das ohne jede Bedeutung. Hätte der Kliniker seine Hypothese vorab, a priori formuliert und nur einen einzigen Test gerechnet, dann wäre es jetzt eine konfirmatorische, eine statistisch beweisende Signifikanz mit einem α-Risiko von maximal 5%. Weil der Kliniker aber seine Hypothese an den gleichen Daten entwickelt und getestet hat, ist es eine Art von Zirkelschluss und das α-Risiko ist nicht 5%, sondern - siehe Tabelle 4 siebte Zeile - 99%. Damit ist diese „Signifikanz" wertlos. Wir müssen also sehr sorgfältig unterscheiden zwischen vorab aufgestellten Hypothesen, der Fachausdruck heißt **a priori Hypothesen**, und solchen Hypothesen, die erst nach Studienbeginn, nach Kenntnis der Daten entwickelt wurden, der Fachausdruck heißt **nachgeschobene Hypothesen**.

• Verschiedene Wahrheiten

Nachdem wir so weit fortgeschritten sind, haben Sie vielleicht Verständnis dafür, wie es zu solchen Steigerungen kommen kann wie **Ausreden, Lügen, Statis-**

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp0(2005014.1)

Tabelle 4: Wahrscheinlichkeit, dass mindestens 1 Test signifikant wird, obwohl für alle Tests definitiv die Nullhypothese gilt. Nominelles Signifikanzniveau = 5%

Anzahl Gruppen	Anzahl Zeitpunkte	Anzahl Merkmale	Daraus: Anzahl Tests [1)]	Wsk (≥1 Test sign.) [2)]
2	1	1	1	5%
2	1	2	2	10%
2	1	4	4	19%
2	1	10	10	40%
2	2	5	20	64%
2	3	5	45	90%
2	3	10	90	99%
2	3	15	135	100%
4	5	100	7 000	100%

[1)] Anzahl der an diesem Datenmaterial möglichen paarweisen Vergleiche. Jeder Vergleich wird getestet

[2)] Wahrscheinlichkeit, dass mindestens 1 Test signifikant wird, obwohl die Daten mit dem Zufallszahlengenerator erzeugt worden sind und damit für alle Tests definitiv die Nullhypothese gilt. Alle Signifikanzen sind somit falsch. Wahrscheinlichkeit auf ganze Prozent gerundet.

Anmerkung: In der Vorlesung wurde anstatt dieser Tabelle eine Serie von 7 Bildern gezeigt, die veranschaulichte, wie mit zunehmender Anzahl von Gruppen, Zeitpunkten und Merkmalen die Wahrscheinlichkeit, dass mindestens 1 Test signifikant wird, sich schnell den 100% nähert.

tiken. Und Sie können sich vielleicht auch vorstellen, warum die **Juristen unterscheiden zwischen der reinen Wahrheit und der vollen Wahrheit**. Reine Wahrheit wäre z.B., irgendein Arzt hat 3 Patienten das Leben gerettet, aber vielleicht ist die volle Wahrheit, dass er auch 5 Patienten mit seinen Kunstfehlern zu Tode gebracht hat. In der Geschichte zwischen Kliniker und Statistiker wäre es die reine Wahrheit, zu sagen, dass der eine Test bei einem nominellen Signifikanzniveau von 5% signifikant geworden ist. Die volle Wahrheit wäre, hinzuzufügen, dass die Hypothese am gleichen Datenmaterial erzeugt worden ist und deshalb das echte (multiple) Signifikanzniveau eher bei 99% als bei 5% liegt. Meist sind medizinische Fragestellungen und Sachverhalte so komplex, dass man auf einem Kongress mit nur 10 Minuten Redezeit die volle Wahrheit kaum darstellen kann.

Wir könnten noch über Vieles reden, auch z.B. über vorsätzliche und fahrlässige Fehler. **Vorsätzliche Fehler sind Fälschungen**, das weiß ich wohl, und ich darf erwähnen, dass Friedhelm Herrmann und ich zur gleichen Fakultät gehörten, er aber nie zu uns in die statistische Beratung kam. Darauf will ich jetzt nicht eingehen. Aber einen Gedanken möchte ich noch darlegen. Bei **fahrlässigen Fehlern** denkt jedermann zunächst, sie hätten keine Tendenz. Fahrlässige Fehler gehen mal in die erwünschte mal in die unerwünschte Richtung. Im Prinzip ja, aber …, Sie kennen diese Witze. Wenn jemand ein unerwünschtes Ergebnis erhält, dann sagt er: „Kann das sein? Das widerspricht meinen Vorstellungen." Er setzt sich hin und sucht nach einem Fehler. Vielleicht hat er ja Glück, findet seinen Fehler und korrigiert ihn. Wenn dagegen das Erwünschte herauskommt, dann freut er sich, lässt den Sektkorken knallen und verwendet das Ergebnis. Fahrlässige Fehler, die das Ergebnis in die unerwünschte Richtung lenken, haben eine größere Chance, entdeckt und korrigiert zu werden als Fehler, die das Ergebnis in die erwünschte Richtung lenken. Somit haben - statistisch betrachtet - fahrlässige Fehler die Tendenz in die erwünschte Richtung.

• Grenzen der medizinischen Statistik

Eines meiner Anliegen ist, Ihnen zu zeigen: **Auch wenn alle statistischen Berechnungen mathematisch solide sind, können die Annahmen, die wir in unsere Mathematik hineinstecken, sehr vage sein.** Auch kleine, zunächst unscheinbar aussehende Änderungen an diesen Annahmen können die Ergeb-

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

nisse völlig verändern. Deshalb kann, trotz solider Mathematik ein Ergebnis irreführend sein. Nicht weil die Statistiker schlechte Mathematiker sind, sondern weil die Annahmen, die Werte, die wir in unsere Berechnungen einsetzen, unsicher sind. Manchmal ist auch das Ergebnis schwierig zu interpretieren, siehe Suizidrisiko.

• Never use a statistical method unless …

Ich denke, nun habe ich Sie genug geplagt. Als kleinen Lohn will ich noch eine Geschichte zitieren, die ich vor etlichen Jahren in Lancet gefunden habe:

Three statisticians and three clinicians are travelling by train to a conference.

The statisticians ask the clinicians whether they have bought tickets. They have. "Fools!", say the statisticians. "We've only bought one between us!"

When the ticket inspector appears, the statisticians go together in the toilet. The inspector knocks and they pass the ticket under the door.

He clips the ticket and slides it back under the door to the statisticians.

The clinicians are very impressed, and resolve to adopt this technique themselves.

On the return they purchase only one ticket between them, and they share the journey with the statisticians, who again ask whether they have all bought tickets.

"No", they reply, "we've bought one to share."

"Fools!", say the statisticians, "we've not bought any ticket."

"But, what will you do when the inspector comes?"

"You'll see."

This time when the inspector appears, the clinicians hide together in the toilet. One of the statisticians walk to the door and knock on it. The clinicians slide their ticket under the door, and the statisticians take it and use it as before -

leaving the clinicians to be caught by the inspector.

Ja, liebe Kollegen und Freunde, das Interessanteste, das Wichtigste, die Moral von der Geschichte kommt noch:

The moral of the story is that you should never use a statistical method unless you are completely familiar with it.

(Adaptiert nach The Lancet Vol. 348, 16. November 1996, p 1392)

• Historische Orakel

Ich will auch noch die eingangs erwähnten Orakel zu Ende bringen. Herodot berichtet in seinen Historien:

Kroisos befragte die Orakel / ob er gegen die Perser in den Krieg ziehen solle.

Beide Orakel, das hellenische in Delphi und das lybische in Abai im Lande Phokis, verkündeten:

Anmerkung des Statistikers: Es ist bemerkenswert, dass sich die Orakel einig waren. Ich möchte gerne wissen, ob sie unabhängig waren (n = 2) oder ob sie sich abgesprochen hatten (n = 1).

Wenn er gegen die Perser in den Krieg zöge / würde er ein großes Reich zerstören.

Kroisos zog gegen die Perser in den Krieg und er zerstörte ein großes Reich, aber es war dummerweise sein eigenes Reich, das er dabei zerstörte!

Das andere Beispiel ist MacBeth von Shakespeare:

Von neuem sucht MacBeth die Hexen auf, / sie um seine Zukunft zu befragen.

Sie versichern ihm:

„Sei blutig, kühn und fest, lach aller Toren: / Dir schadet keiner, den ein Weib geboren; / Kein solcher kränkt MacBeth"

Da wir ja alle vom Weibe geboren worden sind, dachte MacBeth, er sei unbesiegbar. Einige Zeit später, im Kampf gegen MacDuff, wird es dann doch eng für MacBeth und er ruft - damals hat man ja noch Mann gegen Mann gekämpft - seinem Gegner zu:

„Verlorene Müh! / Mein Leben ist gefeit, / kann nicht erliegen, / einem vom Weib Geborenen."

Darauf antwortet MacDuff:

„So verzweifle! / Denn vor der Zeit / ward ich geschnitten aus des Mutters Leib."

Na ja, und dann ging's mit MacBeth zu Ende.

• Medizinische Statistik - Mathematik oder Orakel?

Nun möchte ich zu meinem Vortragstitel zurückkommen. Vorab hatte ich Spekulationen gehört: Einige sagten Medizinische Statistik sei (angewandte) Mathe-

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.

matik, andere spöttelten und sagten, medizinische Statistiken seien wie Orakel, wieder andere meinten, medizinische Statistik sei weder Mathematik, noch Orakel. Nachdem Sie so aufmerksam gelesen haben, werden Sie meiner Antwort auf die im Titel genannte Frage zustimmen können: **„Medizinische Statistik ist sowohl Mathematik als auch Orakel."** Ich kann Ihnen nur den Rat geben: **Betreiben Sie medizinische Statistik stets mit ernsthafter, mit anständiger, mit ordentlicher Mathematik, nach dem Motto: Never use a statistical method unless you are completely familiar with it. Fassen Sie dann aber das Ergebnis wie ein Orakel auf, damit es Ihnen nicht geht wie Kroisos oder MacBeth.**

• Persönliche Worte

Abschließend noch einige persönliche Worte. Mathematik und Medizin sind beides sehr alte, aber doch sehr verschiedene Wissenschaften. Die Arbeit an der Schnittstelle zwischen Mathematik und Medizin finde ich hoch interessant und sie war für mich überaus befriedigend. Oft werde ich gefragt, ob ich von Hause aus Mathematiker oder Mediziner sei. Von meinem Studium her bin ich Ingenieur. Zwar verstehe ich vom Ingenieurwesen nichts mehr, fühle mich aber mit meiner Ingenieurs-Mathematik in der Biometrie gar nicht so falsch am Platz. Ich habe genug Verständnis für die Sorgen der Kliniker und ihre Anwendungen, habe aber auch, denke ich, so viel Mathematik, dass ich zumindest mit den Mathematikern reden kann.

Für das Design einer klinischen Studie gibt es - wie für eine Konstruktion oder einen Entwurf im Ingenieurwesen - eine Fülle widersprüchlicher Anforderungen: Es gibt methodische, versuchsplanerische, biometrische Bedingungen, damit die Ergebnisse der Studie belastbar sind. Eine Studie muss aber auch im Klinikbetrieb praktikabel sein und man muss sie den Patienten mit gutem Gewissen zumuten können, d.h. sie muss ethisch verantwortbar sein. Diese doch sehr widersprüchlichen Anforderungen unter einen Hut zu bringen, und zwar so unter einen Hut zu bringen, dass nicht ein lauer Kompromiss, sondern eine gute Synthese entsteht, ist nicht ganz einfach.

Ich habe mich in der GMDS (und auch in der Biometrischen Gesellschaft und im Deutschen Verband Medizinischer Dokumentare) sehr wohl gefühlt. Den Kollegen danke ich für diese Zeit und für diese Zusammenarbeit. Der GMDS habe ich ein kleines Erinnerungsgeschenk: einen neuen, alten Zehnmarkschein. Einen neuen, weil er noch ziemlich ungebraucht ist, einen alten, weil unsere Währung jetzt der Euro ist. Auf diesem Zehnmarkschein ist der Mathematiker und Statistiker Carl Friedrich Gauß und die von ihm entwickelte Normalverteilung abgebildet. Dieser Carl Friedrich Gauß ist natürlich viel berühmter als ich, aber er ist schon 1855 gestorben, während ich noch lebe und auch ganz munter bin. Momentan ist mir mein Leben und auch meine bevorstehende Pension viel lieber als sein Ruhm, den vergönne ich Carl Friedrich Gauß von ganzem Herzen. Diesen 10 DM-Schein, dieses kleine Souvenir muss aber die GMDS nicht auf immer und ewig aufbewahren. Auch Erinnerungen, auch Souvenirs verblassen, liegen dann herum und werden lästig. Sollten einmal die Finanzen der GMDS knapp sein, dann darf - es muss nicht der Präsident sein - jemand auf die Landeszentralbank gehen, den 10 DM-Schein in Euro umtauschen und diese 5.11 Euro für die GMDS völlig frei verwenden. Allerdings bin ich mir bewusst, dass es ein winziger Betrag ist. Ich wünsche der GMDS, ich wünsche allen Kollegen und Freunden Gottes Segen und viele pfiffige Ideen. Dann wird es gut gehen. Herzlichen Dank für die Zeit mit Ihnen und alles Gute!

Korrespondenzadresse:

• Prof. Dr. Wilhelm Gaus, Medizinische Fakultät der Universität Ulm, Abteilung Biometrie und Medizinische Dokumentation, 89070 Ulm, Deutschland
wilhelm.gaus@medizin.uni-ulm.de

Artikel online frei zugänglich unter http://www.egms.de/en/journals/mibe/2005-1/mibe000008.shtml
Bitte zitieren als: Gaus W. Medizinische Statistik – Mathematik oder Orakel?: Abschiedsvorlesung am 16. Juli 2004. GMS Med Inform Biom Epidemiol. 2005;1(2):Doc08.
gbp(M2005081414-1)