

Entwicklung des Gruppen-Merkmals „Führungstechnische Fähigkeiten“ (in Prozent aller Anzeigen)

1993	1994	1995
50	41,5	52,1

lenanzeigen kein ausdrücklicher Wert auf Führungsfertigkeiten gelegt wird, obwohl der Arzt die Leitung über das medizinische Personal seiner Abteilung ausübt.

Weiterhin wurde ermittelt, inwieweit in den Stellenanzeigen eine Angabe der Versorgungsstufe wie zum Beispiel Grund-, Regel- oder Schwerpunktversorgung des Krankenhauses erfolgte:

Entwicklung der Angabe der Versorgungsstufe des Krankenhauses (in Prozent aller Anzeigen)

1993	1994	1995
41,2	49,4	53,1

Daraus ergibt sich ein anhaltend steigender Trend bei der Angabe der Versorgungsstufe der stellensuchenden Krankenhäuser. Da jedoch in 1995 fast die Hälfte aller Stellenanzeigen für leitende Ärzte ohne Angabe zur Versorgungsstufe ist, stellt sich die Frage, warum diese Angabe so häufig fehlt. Denn durch die Angabe der Versorgungsstufe kann dem potentiellen Stellenbewerber die Möglichkeit gegeben werden, Rückschlüsse auf den medizinischen „Standort“ des Krankenhauses zu ziehen.

Zitierweise dieses Beitrags:
Dt Ärztebl 1996; 93: A-2204–2206
[Heft 36]

Anschrift der Verfasser:

Prof. Dr. phil. Günther E. Braun
Dipl.-Wirtsch.-Inf. Dirk Egner
Lehrstuhl für Allgemeine Betriebswirtschaftslehre, insbesondere Öffentliche Verwaltungen und Öffentliche Unternehmen
Universität der Bundeswehr München
85577 Neubiberg

Qualitätssicherung im Krankenhaus

Trugschlüsse biometrischer Untersuchungen

Heino Kuhleemann, Jürgen Majerus, Johannes Möller

„Mathematische Unkenntnis zeigt sich in zu genauem Rechnen“ (C. F. Gauß) lautet eine fundamentale These, die in der Medizin wenig beachtet wird. Aussagen wie „bei einem von sieben Patienten trat eine Komplikation auf, das sind 14,29 Prozent“, untermauern die Gaußsche (1) Meinung. Im obigen Fall ist eine Prozentangabe nicht sinnvoll, insbesondere ist die Berechnung auf zwei Nachkommastellen nicht angemessen. Die Biometrie (2, 3, 4, 5, 6) ist in der empirischen Wissenschaft Medizin ein leistungsfähiges klassisches Verfahren, um medizinische Prozesse und Ergebnisse – zum Beispiel Therapiearten oder Operationstechniken – bewerten zu können. Sie ist im Rahmen von Qualitätssicherungsprojekten ein wichtiges Werkzeug und bietet durch die Möglichkeiten der EDV neue Chancen, birgt allerdings auch Risiken.

Anforderungen an die medizinische Behandlung eines Krankenhauses werden formuliert durch Patienten, Mitarbeiter, die Träger sowie die Kassen. Primäre Aufgabe der Qualitätssicherung ist die Feststellung der vorhandenen Qualität. Im nächsten Schritt ist das aktuelle Qualitätsniveau zu verbessern beziehungsweise ein bereits erzieltes hohes Niveau zu halten (7). Für diese Aufgaben spielt die Biometrie, die die medizinische Statistik beinhaltet, eine wichtige Rolle. Um dieses Werkzeug verantwortungsvoll einsetzen zu können, ist eine korrekte Methodologie die Voraussetzung.

„Biometrie“ mit „Statistik für die Medizin“ zu übersetzen wäre zu einfach. Die vielschichtige Problematik, die mit der wissenschaftlichen Untersuchung medizinischer Fragestellungen und mit dem unkritischen Einsatz von Statistiksoftware in der Medizin zusammenhängt, verlangt nach tiefergehenden Einblicken in die Bewertungsprinzipien medizinischer Tätigkeit. Die Biometrie ist umfassender als die reine Statistik und auf die medizinischen Anforderungen ausgerichtet. Dazu gehören zum Beispiel

- ▶ die kritische Betrachtung potentieller Meßgeräte
 - ▶ die Datenakquisition
 - ▶ die Festlegung der Stichproben
 - ▶ die Wahl der mathematischen Auswertung und der statistischen Schlußweisen
 - ▶ die Methode der Ergebnispräsentation und die Schlußfolgerungen.
- Dabei sind Kenntnisse der Meßtechnik ebenso erforderlich wie praktische mathematische Erfahrungen.
- Grundsätzlich ist in der Biometrie die deskriptive von der schließenden Statistik zu unterscheiden.
1. Die *deskriptive* Statistik beschreibt entweder eine Stichprobe oder eine Vollerhebung. Sie liefert – oft hochinteressante – interne Erkenntnisse und unterstützt damit das hausinterne Qualitätsmanagement. Die Wahl der grafischen Darstellung von zum Beispiel Mittelwerten oder Prozentanteilen ist entscheidend für die Interpretation und die Rezeption der Inhalte durch die Adressaten (Patienten, Ärzte). Hierbei können Fehlinterpretationen nicht adäquater Darstellungen Trugschlüsse herbeiführen. ▷

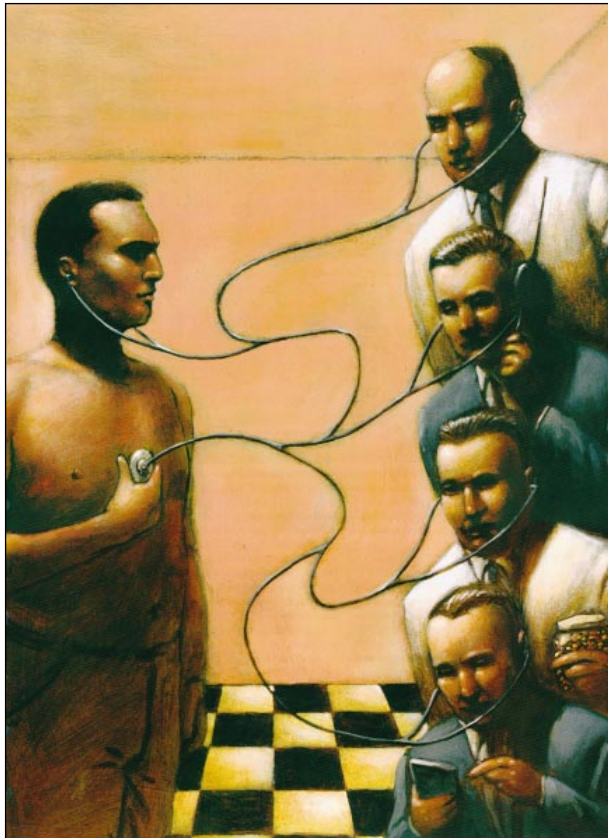
In der Werbung werden sogar durch gezielt inadäquate Darstellung Trugschlüsse beim Patienten ausgelöst, der dadurch beeinflusst bestimmte Wünsche beim Arzt artikuliert. Sollen allgemeine medizinische Erkenntnisse gewonnen werden, bedient man sich üblicherweise der schließenden Statistik. Deskriptive Statistik, also die Beschreibung einer Stichprobe, hat bezüglich des Schlusses über die Stichprobe oder Vollerhebung hinaus nur eine sehr eingeschränkte Aussagekraft.

Nicht selten sind Aussagen wie „Mit OP-Technik A ($n = 28$) betrug die Aufenthaltsdauer in der Stichprobe im Mittel 4,2 Tage, mit OP-Technik B ($n = 31$) 5,1 Tage“, denen dann unsinnige Interpretationen wie „OP-Technik A ist um 0,9 Tage besser und bezüglich der Aufenthaltsdauer also als besser anzusehen“ folgen. Man darf im allgemeinen mittels deskriptiver Statistik – hier als Beschreibung von Mittelwerten einer Stichprobe – keine Schlußfolgerungen auf die Grundgesamtheit formulieren.

2. Die *schließende* Statistik dahingegen gelangt genau dann zu solchen Folgerungen, wenn mathematische Verfahren eine OP-Technik als signifikant besser identifizieren. Sie berücksichtigt im vorliegenden Beispiel, daß sich beide (randomisierten) OP-Gruppen der Stichprobe bezüglich ihrer Aufenthaltsdauer zwar im Mittel unterscheiden, aber dennoch möglicherweise aufgrund der Werteverteilung kein derartiger Schluß auf ein größeres Patientengut möglich ist.

Eine sinnvolle Interpretation desselben Sachverhalts würde zum Beispiel lauten: „Der Unterschied der Liegedauer wurde auf einem vor Studienbeginn festgelegten Signifikanzniveau (Irrtumswahrscheinlichkeit) von $\alpha = 0,01$ getestet. Der ungepaarte t-Test ergab dabei keine signifikant niedrigere Liegedauer (errechnetes $p = 0,06 > \alpha = 0,01$) für OP-Technik A (im Mittel 4,2 Tage), verglichen mit OP-Technik B (im Mittel 5,1 Tage),

aufgrund der Verteilung der Werte in der Stichprobe. Die Hypothese der Gleichheit der OP-Techniken, bezogen auf die Liegedauer, kann nicht falsifiziert werden. Der in der Stichprobe errechnete Unterschied ist nicht signifikant.“ Die Folgerung also, daß OP-Technik A der anderen überlegen ist, läßt sich hieraus nicht ableiten.



Fallstricke klinischer Studien: Verschiedene Untersucher können durch Anwendung unterschiedlicher statistischer Methoden zu differenten Ergebnissen kommen.

Abbildung: Odyssey/R. Blommesteijn

Die obige Aussage, wonach die OP-Technik „A“ der alternativen OP-Technik „B“ überlegen ist, ist auf der Basis einer Stichprobenbeschreibung (deskriptiv) anhand von Mittelwerten schlicht falsch. Bei schließen der Statistik, die für derartige Fragestellungen das geeignete Verfahren ist, bedarf es eines trugschlußfreien Designs und der vollständigen Angabe der zugrundegelegten Annahmen und Methoden.

Mehr Information als die Angabe eines p-Wertes allein enthält die – mathematisch äquivalente – zusätzliche Angabe von Konfidenzintervallen.

Die oftmals vorzufindende Verkürzung „Je kleiner das p, um so besser“ vernichtet Informationen und ist in dieser Verkürzung falsch. Eine Kombination verschiedener Angaben erst (deskriptive Werte, p-Werte, Konfidenzintervalle, ranges etc.) und eine geprüfte Durchsicht der Rohdaten sind hilfreich bei der Vermeidung irrtümlicher Kurzinterpretationen.

Das Schätzen von Intervallgrenzen, zwischen denen der reale Erwartungswert (zum Beispiel der unbekannte Mittelwert der Grundgesamtheit) liegen kann, ist dabei ein adäquates Verfahren: „Die angegebenen Mittelwerte sind dabei Schätzer der unbekannteren Erwartungswerte, die durch folgende Intervallgrenzen zur statistischen Sicherheit $1 - \alpha = 0,99 = 99$ Prozent angegeben werden können: (...). Das Konfidenzintervall für die Differenz lautet $[-0,2; 2,0]$. Die Zahl 0 liegt zwischen diesen Konfidenzgrenzen; auch daraus ist ersichtlich, daß kein signifikanter Unterschied angegeben werden kann.“ Grundlage der schließenden Statistik ist immer das jeweilige mathematische Modell, das nur eine theoretische Annahme darstellt. Die Gaußsche Normalverteilung (1) ist beispielsweise ein Modell für eine Verteilung, die derart in der Medizin nicht vorkommt. Die Annahme eines Modells aber ermöglicht erst die nächsten Rechenschritte und damit den statistischen Rückschluß auf dasjenige Patientenkollektiv, für welches zukünftige Entscheidungen getroffen werden sollen. Ähnlich werden erst unter bestimmten Vorbedingungen statistische Tests möglich.

Ein wichtiges Grundprinzip der schließenden Statistik ist ferner, daß sie keine abschließende verifizierende Beweiskraft besitzt. Man kann lediglich eine vorläufige Schlußfolgerung ziehen, die mit einer Wahrscheinlichkeitsaussage über einen möglichen Irrtum dieser Schlußfolgerung belegt wird. Die Wahrscheinlichkeitsaussagen gelten aufgrund möglichst trugschlußfreier Methodik und

Bildung mathematischer Modelle, die dennoch nur eine begründete Annahme darstellen. Diese Wahrscheinlichkeit ist das bekannte Signifikanzniveau. $p < 0,01$ bedeutet also, daß die Wahrscheinlichkeit, daß der statistische Schluß ein irrtümlich gezogener ist, 1 Prozent = 0,01 beträgt.

Eine Stichprobe hat üblicherweise charakteristische Werte (zum Beispiel die mittlere Liegedauer von Patienten), die mit einer bestimmten Technik operiert worden sind. Ein Mittelwert von 5,3 Tagen sagt nur etwas über die Stichprobe, also die untersuchten Patienten, aus – mehr nicht! Sie sind im mathematischen Sinne „Schätzer“ des unbekanntes Erwartungswertes.

Der Erwartungswert ist lediglich durch modellhafte Annahme in Form von Grenzen eines sogenannten Konfidenzintervalles anzugeben. Das Intervall [4,7; 5,9] für die Liegedauer gibt an, daß mit einer statistischen Sicherheit von zum Beispiel 99,9 Prozent = 0,999 (also Irrtum 0,1 Prozent = 0,001) der Erwartungswert für die Liegedauer von den Intervallgrenzen eingeschlossen wird. Wie der reale Erwartungswert aber lautet, wird unbekannt bleiben.

Vernetzung von Daten

Bereits in der früheren Geschichte der Medizin wurde die Biometrie zur Erstellung trugschlußfreier Metriken angewendet. Über ihre Nutzung als qualitätsbewertendes Werkzeug hinaus ist die Biometrie selbst als Objekt einer Qualitätsprüfung anzusehen, um einen methodisch korrekten Einsatz – frei von Trugschlüssen – gewährleisten zu können.

Trägt man die im menschlichen Körper vorhandene Vitamin-C-Konzentration über die Zeit grafisch auf, ohne aber die Zeitachse mit Einheiten wie Minuten, Stunden oder Tagen zu beschriften, kann man sogar suggerieren, man müsse stündlich Vitamin C einnehmen (deskriptiver Trugschluß). Erst die vollständige deskriptive Darstellung ermöglicht die Einsicht, wie häufig und in welcher Dosis Vitamin C verabreicht werden sollte – vorausgesetzt, daß Vitamin-C-Gaben überhaupt sinnvoll sind. Der immer

häufigere Einsatz von Statistiksoftware führt derzeit zu einer Renaissance klassischer Trugschlüsse. Der Trend zur Vernetzung von Patientendatenbanken und Informationssystemen innerhalb einer Organisation (Intranet) oder weltweit (Internet) bringt weitere Gefahren mit sich, die gravierende Folgen haben, wenn man sich nicht um Ausschaltung systematischer Fehlerquellen bemüht.

Die Trugschlüsse reichen dabei von einfachen kognitiven Insuffizienzen wie im obigen Prozentbeispiel bis zu komplizierteren Fehlern bei multizentrischen retrospektiven Auswertungen von Datenbanken und Registern. Der Trend zur weltweiten Vernetzung bietet zwar große Chancen für die wissenschaftliche Auswertung von Daten. Die anwenderfreundliche Bedienung von Statistiksoftware aber verleitet, zu geringen Aufwand in die Vermeidung methodischer Fehler zu investieren.

Die Verantwortung für die Veröffentlichung von Ergebnissen ist groß, wenn man bedenkt, daß dadurch die Behandlung von Tausenden von Patienten beeinflusst werden kann. Diese Verantwortung kann ohne die korrekte Anwendung der Biometrie im allgemeinen nicht mehr getragen werden. Es ist zu beobachten, daß allzu oft Studien und Auswertungen veröffentlicht werden, die schon im Design systematische Fehler aufweisen, wodurch die Ergebnisse einen unkalkulierbaren Bias (Verzerrung) enthalten und die Interpretation geradezu gefährlich ist.

Selbst abstracts von Wissenschaftlern enthalten nicht selten Fehler, die für den Arzt oft schwer erkennbar sind. Diese Trugschlüsse sind aber standardisierbar, und viele können schon durch Grundkenntnisse weitgehend vermieden werden (1), ohne die mathematischen Hintergründe im Detail verstehen zu müssen.

Es ist unbestritten, daß mathematische Modelle, die für eine Stichprobe mit beispielsweise acht Patienten angenommen werden, mehr als fraglich sind. Aber auch mit hohen Fallzahlen ist nicht gewährleistet, daß die getroffenen Annahmen einen Schluß auf die gewünschte Patientengruppe zulassen, wenn beispielsweise die Merkmalsstruktur der

Stichprobe nicht mit der Struktur der Gesamtpopulation übereinstimmt (Strukturungleichheit). Ein diagnostischer Test auf eine in bestimmten Bevölkerungsschichten besonders häufige Krankheit (M. Bechterew oder HIV) hat bei einem einzelnen getesteten Patienten eine unterschiedliche Aussagekraft (sogenannter positiver und negativer prädiktiver Wert) – je nachdem, ob der Patient zu einer Risikogruppe gehört oder nicht.

Geprüftes Studiendesign

Bei einer Prüfung der diagnostischen Eingangstests, für die sich ein Krankenhaus entscheidet, müssen derartige Effekte berücksichtigt werden. Neben der oft wenig spezifizierten „langjährigen Erfahrung eines Arztes“ gibt es also objektive Verfahren, die die diagnostische Handlungsweise und Entscheidungsfindung unterstützen und optimieren können. Die immer noch in einigen Bereichen zu hohe Zahl an testpositiven Patienten, bei denen durch die intraoperative Diagnostik festgestellt wird, daß die präoperativ festgestellte OP-Indikation ein Irrtum war, untermauert die Forderungen nach biometrischen Konzepten im Rahmen von Qualitätssicherungsprojekten.

Üblicherweise werden nach einem geprüften Studiendesign die Werte in einer Datenbank erfaßt und statistisch ausgewertet. Die moderne Tendenz der halb- oder vollautomatischen Dokumentations- und Auswertungsprogramme soll vielfach den Wunsch von Medizinern befriedigen, sich möglichst wenig mit der Materie „Biometrie“ auseinandersetzen zu müssen und „schnell Ergebnisse“ erhalten zu können. Doch genau in diesen Forderungen liegt die große Gefahr weiterer statistischer Trugschlüsse. Wird die konkrete Fragestellung nicht im Vorfeld prospektiv definiert, sondern werden nur „möglichst viele“ Patientendaten dokumentiert, um dann „mal zu sehen, was an Ergebnissen herauskommt“, begeht man bereits einen der wichtigsten Trugschlüsse [8], nämlich den des Multiplen Testens.

Dieser zwar einfache, aber dennoch weitverbreitete Trugschluß

Multiples Testen soll hier exemplifizierend erläutert werden, weil dessen Verbreitung mit dem Einsatz von EDV proportional zunimmt. Er tritt häufig auch in Kombination mit weiteren, zum Teil schwer erkennbaren Trugschlüssen auf, gehört aber zu denjenigen, die schon in der Designphase vermieden werden sollten, während andere erst in einer späteren Phase auffallen und dann zu einem Abbruch und Neustart der Studie führen müssen. Das Multiple Testen tritt inhaltlich meist in drei – hier beispielhaft vereinfachten – Varianten auf. Das Grundprinzip ist vergleichbar mit einem Glücksspiel, bei dem die Gewinnchance zum Beispiel 1 Prozent = 0,01 beträgt. Je öfter man spielt, um so größer ist die gesamte Wahrscheinlichkeit, in dieser Kette von Spielversuchen mindestens einen Gewinn zu erzielen. Spielt man zum Beispiel 500mal, gewinnt man „ziemlich sicher“, nämlich mit einer Wahrscheinlichkeit von 1 bis $0,99^{500} = 99,3$ Prozent. Anders als beim Glücksspiel, bei dem jeder Versuch (Ziehung) einen Einsatz kostet, dem Spieler also die finanziellen Mittel ausgehen würden, geht das Datenmaterial des Biometrikers niemals aus, und er kann „beliebig oft spielen“.

Trugschluß 1: Es werden „möglichst viele“ Werte dokumentiert, um hinterher (retrospektiv) möglichst „irgendein signifikantes Ergebnis“ erreichen zu können. Dieses retrospektive Ausprobieren hat folgenden Trugschlußmechanismus: Je mehr (stochastisch unabhängige) Parameter wie Schmerz, Zufriedenheit, Blutverlust etc. (zum Beispiel 100) auf dem Signifikanzniveau ($\alpha = 0,01$) ausgewertet werden, um so größer wird die Wahrscheinlichkeit, daß man ein irrtümlich signifikantes Ergebnis erhält. In diesem Beispiel mit 100 Tests beträgt die Wahrscheinlichkeit, irrtümlich mindestens ein signifikantes Ergebnis zufällig zu erhalten, immerhin schon $1 - 0,99^{100} = 63$ Prozent. Wer also 100 Parameter multipel testet, erzielt mit einer Wahrscheinlichkeit von 63 Prozent ein – möglicherweise zufälliges und irrtümliches – signifikantes Ergebnis.

Trugschluß 2: Es wird aus verschiedenen Datenbeständen (zum Beispiel 100 Stationen) derselbe Para-

meter (zum Beispiel Blutverlust im Vergleich zweier OP-Techniken) getestet. Die Wahrscheinlichkeit, daß mindestens ein Wissenschaftler irrtümlich ein signifikantes Ergebnis veröffentlicht wird, liegt ebenfalls bei $1 - 0,99^{100} = 63$ Prozent.

Trugschluß 3: Der positive prädiktive Wert (ppW), also die Wahrscheinlichkeit, daß ein Testpositiver auch wirklich krank ist, liege bei einem isolierten diagnostischen Eingangstest (zum Beispiel Diagnostik mit MRT-Bild, mit dem eine OP-Indikation festgestellt werden soll) bei $0,995 = 99,5$ Prozent. Der Test also „funktioniere mit 99,54prozentiger Sicherheit“. Die Wahrscheinlichkeit, daß bei vielen untersuchten Patienten (zum Beispiel 300 Fällen), deren OP-Indikation ausschließlich zur Vereinfachung aufgrund des Bildbefundes gestellt werden soll, mindestens ein Patient auf dem OP-Tisch liegen wird, obwohl er gesund ist, liegt bei $1 - 0,995^{300} = 0,78 = 78$ Prozent.

Entscheidungen „online“

Qualitätssicherung als aktuelles Thema im Gesundheitswesen ist ohne Biometrie nicht realisierbar, wenn man die konkreten medizinischen Ergebnisse bewerten möchte. Sichern gegenwärtige Qualitätssicherungsprojekte nur die organisatorischen Rahmenbedingungen, die eine qualitativ hochwertige Arbeit überhaupt erst ermöglichen, so geht die Biometrie weiter und bei medizinisch-inhaltlichen Bewertungen in medias res. Damit ist sie eine der konkreten Ausbaustufen, die ein systematisch aufgebautes Qualitätssicherungssystem bis ins medizinisch Detaillierte und Konkrete umzusetzen helfen. Die Biometrie kann dabei das Qualitätsmanagement des gesamten prä-, intra- und postoperativen Bereiches unterstützen und sogar rechnergestützt zu einer sofortigen medizinischen Entscheidungsfindung im Einzelfall herangezogen werden.

Der Einsatz von Rechenanlagen ermöglicht heutzutage – eine saubere Planung vorausgesetzt – durchaus „schnelle Ergebnisse“ sogar in Echtzeit. Echtzeit bedeutet, daß von der Anforderung bis zum Erhalt des Er-

gebnisses eine definierte geringe Antwortzeit nicht überschritten werden darf. Damit ist sogar die Unterstützung medizinischer Entscheidungsfindung online möglich. Telemedizin und hausinterne Informationssysteme bieten diesbezüglich Chancen, die nur unter Beachtung und weitgehender Ausschaltung ihrer Risiken für das Gesundheitswesen nutzbringend sein werden.

Eine intraoperative Vermessung von Bohrkanälen bei Kreuzbandplastiken durch Röntgenbildverarbeitung ist dafür ein aktuelles Beispiel. Durch ein derartiges Verfahren kann die Qualität der medizinischen Arbeit kontinuierlich durch biometrische Auswertung ausgewertet und durch darauf aufbauende medizinische Entscheidungsfindung gesichert werden. Unter Berücksichtigung der obigen Ausführungen ist für Knieoperationen zu folgern, daß es nicht hinreichend ist, Mittelwerte für die Qualität von Operationsergebnissen am Kniegelenk bundesweit anzugeben („Viele überflüssige Operationen“) (12), sondern es ist anschließend das medizinisch machbare Optimum (Machbarkeitsanalyse) biometrisch zu dokumentieren. Schlechtere Ergebnisse sind im Sinne einer Schwachstellenanalyse unbedingt zu isolieren – und dies ist mit heutigen Methoden sehr gut möglich. Studien wie die der Medizinischen Hochschule Hannover (12) sind ein professioneller Einstieg in die Planung detaillierterer Untersuchungen.

Insofern bleibt als Fazit: Die Biometrie ist – korrekt genutzt – die adäquate Methode, um die Qualität und die Unabhängigkeit ärztlicher Berufsausübung gewährleisten zu können.

Zitierweise dieses Beitrags:
Dt. Ärztebl 1996; 93: A-2206–2212
[Heft 36]

Die Zahlen in Klammern beziehen sich auf das Literaturverzeichnis im Sonderdruck, anzufordern über die Verfasser.

Anschrift für die Verfasser:

Heino Kuhlemann
Große Mantelgasse 5
69117 Heidelberg