Medizinische Klinik
1998;93:571–6 (Nr. 9), © Urban & Vogel, München

Biometrische Grundlagen der Medizin

Markus Löffler[1]

In diesem Beitrag möchte ich Stellung und Bedeutung der Biometrie in einer rational begründeten Medizin erläutern und mit einigen aktuellen Beispielen belegen.

Bio-Metrie bezeichnet die Lehre vom Umgang mit Meß- und Zählbarem in den Biowissenschaften. Nach einer verbreiteten, in meinen Augen aber zu engen Auffassung beschäftigt sich Biometrie mit der Anwendung von statistischen Methoden auf biologische Phänomene. Biometrie sollte aber viel weiter greifen und als Methodik der rationalen Erkenntnisgewinnung in der Medizin mittels quantitativer Modellbildung verstanden werden.

Nachfolgend sollen zunächst drei Argumente belegen, weshalb eine Beschäftigung mit statistischen Verfahren im medizinischen Erkenntnisprozeß wichtig ist. Anschließend diskutiere ich weiterführende Beiträge der Biometrie zum medizinischen Erkenntnisprozeß.

Das Erkenntnisproblem

Eine wissenschaftlich begründete Medizin basiert auf dem in Abbildung 1 schematisch skizzierten Erkenntnisprozeß. Demnach imponiert dem Beobachter in seiner Rolle als Biowissenschaftler oder klinischem Forscher eine Phänomenologie. Im klinischen Kontext handelt es sich um eine Fülle von Symptomen und Befunden, die eingeordnet werden müssen. Es besteht die grundlegende Vermutung, daß dieser Phänomenologie eine Regelhaftigkeit, eine Ableitbarkeit zugrunde liegt. Das wissenschaftliche Erkenntnisinteresse besteht darin, solchen verborgenen Regelhaftigkeiten bestimmter Teilaspekte des biologischen Geschehens nachzuforschen und sie zu identifizieren. Im allgemeinen bedient man sich einer bewährten Taktik, die beobachtbare Phänomenologie in ausgewählten Aspekten quantitativ zu untersuchen. Anhand von geeigneten Stichproben und ausgewählten Observablen werden Zählungen und Messungen herbeigeführt. Aus diesen Daten wird der Inferenzversuch unternommen. Dieser besteht darin, mittels eines zunächst qualitativen und dann auch quantitativen Modells eine allgemeingültige Erklärung für die beobachtbare Phänomenologie zu geben. Im klinischen Kontext ist beispielsweise eine Diagnose ein solches Modell (genauer ein diagnostisches Klassifikationsmodell). Aber auch prognostische Aussagen oder Feststellungen über die Verhaltensweisen eines physiologischen Prozesses basieren oft auf Modellvorstellungen. In jedem Fall stellt sich die Frage nach der Adäquatheit zwischen Modell und wahrem biologischen Sachverhalt. Argumente für die Adäquatheit zu finden und sie zu untermauern ist Gegenstand des wissenschaftlichen Prozesses, in dem der beständige Zweifel an der Adäquatheit die Triebfeder darstellt.

Im klinischen Erkenntnisprozeß spielt die Untersuchung an Individuen eine wesentliche Rolle. Es ist bei allen Experimenten, insbesondere auch bei biologischen Versuchen, eine fundamentale Erfahrung, daß man auf ein und denselben Eingriff unterschiedliche Reaktionen erhält. Jeder Versuch fällt anders aus als seine Wiederholung oder ein Parallelversuch, selbst dann, wenn es gelingt, alle (bekannten) Einflußgrößen gleich zu halten. Damit wird deutlich, daß empirische Beobachtungen nur dann eine brauchbare Grundlage für eine Schlußfolgerung sein können, wenn sie durch wiederholte Versuche an einer Stichprobe von Individuen vorgenommen werden. Dabei muß eine Stichprobenauswahl bestimmte Kriterien erfüllen, die man an einen unselektionierten Zufallsprozeß stellt.

Damit ist das Grundproblem der statistischen Schlußweise skizziert. Offensichtlich beinhaltet der induktive Schluß von den Ergebnissen einer Stichprobenuntersuchung auf die gar nicht in Vollständigkeit beobachtete Gesamtheit aller Indiduen eine gewisse Unsicherheit. Um Aussagen über Art und Grad der Sicherheit bzw. Unsicherheit zu quantifizieren, ist es erforderlich, ein die untersuchten Zusammenhänge beschreibendes wahrscheinlichkeitstheoretisches Modell zu spezifizieren. Dies führt in das Fachgebiet der Statistik.

Die Statistik befaßt sich mit der Modellierung der Zufallsphänomene. Die statistischen Modelle beschreiben in Wahrscheinlichkeitsbegriffen, zu welchem Anteil Variationen von Meßgrößen auf bekannte Einflußgrößen zurückgeführt werden können. Die vielleicht häufigsten Fälle statistischer Modellierungen in diesem Sinne sind Regressionsmodelle. Die Ausprägung einer zu erklärenden Zielgröße wird vermöge einer mathematischen Funktion, deren Gestalt man aufgrund anderer Argumente begründen muß, auf „kausale" Einflußgrößen sowie bekannte Kovariablen und Zufallseffekte zurückgeführt (zum Beispiel: y = f [Therapie, Kovariable, Zufall]). Bei dieser Sichtweise wird deutlich, daß eine zentrale Bedingung für die Modellierung von Zufallsphänomenen eine reduktionistische Sichtweise ist. Sie setzt die Annahme voraus, daß Individuen partiell vergleichbar sind, das heißt, hinsichtlich der ausgewählten Merkmale Ähnlichkeiten aufweisen. Auf dieser Annahme basierend, ist die schließende Statistik eine Sammlung von Methoden, die es erlaubt, in einem mathematischen Sin-

[1] Institut für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig.

Nach einem Vortrag auf dem Symposium „Zu den Grundlagen der Medizin" zu Ehren von Prof. Dr. med. Dr. h. c. Rudolf Gross anläßlich seines 80. Geburtstags.

ne angebbare optimale Entscheidungen im Falle von Ungewißheit zu treffen und die Wahrscheinlichkeiten für Fehlentscheidungen zu minimieren. Zudem erlauben die statistischen Methoden, Art und Grad der Unsicherheit über eine Schlußweise zu quantifizieren, sofern ein zu überprüfendes Modell zugrunde liegt. Dabei sind die statistischen Modelle wahrscheinlichkeitstheoretische Modelle und die resultierenden Aussagen Wahrscheinlichkeitsaussagen.

Als Beispiel sei die Entscheidungsstrategie von Vergleichsfragestellungen im Paradigma der frequentistischen Statistik herausgegriffen. Sie ist eine der häufigsten Entscheidungsstrategien und tritt beispielsweise im Kontext vergleichender kontrollierender Therapiestudien auf. Es handelt sich dabei um ein weitgehend standardisiertes und kanonisch ablaufendes Entscheidungskonzept. Demnach wird angenommen, daß sich eine Meßgröße als Zufallsvariable verhält. Es wird ferner angenommen, daß man ein entsprechendes Zufallsexperiment grundsätzlich beliebig oft wiederholen kann. Sodann formuliert man eine quantitative Vermutung (Hypothese) über den wahren Wert dieser Meßgröße. Aufgrund der begrenzten Zufallsstichprobe kann die beobachtete Meßgröße eines Zufallsexperiments diesen vermuteten wahren Wert nicht genau einnehmen, sondern kann in einem bestimmten Intervall von ihm abweichen. Falls die Abweichung zu groß ist, wird die Entscheidung getroffen, daß die Hypothese falsch ist. Dabei wird der Grad der Unsicherheit dieser Aussage mit dem empirischen Signifikanzniveau (p-Wert) angegeben. Ist hingegen der Unterschied zwischen beobachteter Meßgröße und vermuteter wahrer Meßgröße nur gering, so tritt zunächst eine paradox erscheinende Situation ein: Eine geringe Abweichung läßt nicht die Entscheidung zu, daß die Vermutung zutreffend gewesen ist, denn es ist durchaus möglich, daß aufgrund möglicherweise vorliegender Zufallsschwankungen die beobachteten Werte nahe bei dem vermuteten Wert liegen, obwohl dieser die wahre Situation nicht zutreffend charakterisiert. Folglich läßt sich im Fall der nichtsignifikanten Testsituation keine konklusive Aussage machen.

Diese asymmetrische Entscheidungssituation erschien den Statistikern Neymann und Pearson unbefriedigend, und sie entwickelten eine Entscheidungsstrategie, die zwischen zwei Hypothesevermutungen (der Nullhypothese und der Alternativhypothese) in jedem Fall eine Entscheidung treffen sollte. Das Entscheidungsszenarium ist in Abbildung 2 widergegeben. Für den Fall eines statistischen Tests auf Unterschied wird in der Nullhypothese angenommen, daß es keinen Unterschied zwischen den Stichproben gibt, und in der Alternativhypothese wird davon ausgegangen, daß dieser ein bestimmtes festgelegtes Ausmaß erreicht. Aufgrund der Stichproben werden Entscheidungen für oder gegen die Nullhypothese getroffen. Dabei können Fehlentscheidungen eintreten.

Eine Fehlentscheidung liegt vor, wenn aufgrund der Stichprobe ein Unterschied gefolgert wird, obwohl er in Wirklichkeit nicht vorliegt. In diesem Fall hätte man ein statistisch falsch positives Ergebnis erhalten (Fehler erster Art). Ein falsch negatives Ergebnis liegt dann vor, wenn man aufgrund der Stichprobenergebnisse zu der Entscheidung kommt, daß die Daten mit der Nullhypothese verträglich sind, obwohl tatsächlich ein systematischer Unterschied zugrunde liegt. Aus theoretischen Gründen ist es unmöglich, diese beiden Fehler vollständig zu eliminieren. Sie lassen sich lediglich durch die Wahl sehr großer Stichproben verkleinern, und gegebenenfalls kann durch die Wahl geeigneter Entscheidungsstrategien einer der Fehler auf Kosten des anderen vermindert werden.

Kognitive Grenzen menschlicher Wahrscheinlichkeitsurteile

Als zweites Argument für eine Beschäftigung mit der Statistik in der Medizin seien die Probleme angeführt, die menschlichen Wahrscheinlichkeitsurteilen zugrunde liegen.

So haben Kahnemann et al. [8] eindrucksvolle Untersuchungen darüber vorgelegt, wie unsere menschliche Kognition Schlußfolgerungen aus Stichproben zieht und zu heuristischen Wahrscheinlichkeitsurteilen kommt. Die wichtigste Heuristik ist die der Repräsentativität. Sie weist folgende Merkmale auf:

1. Wir erwarten, daß die Zusammensetzung von Stichproben der Zusammensetzung der Grundgesamtheit stark ähnelt;

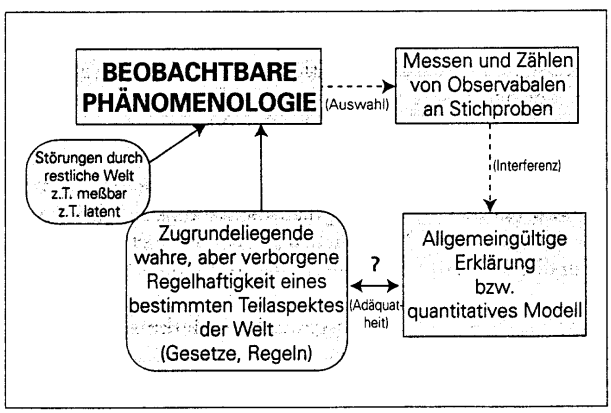

Abbildung 1. Grundlegende Erkenntnisabsicht.

NULLHYPOTHESE		Unbekannte wahre Situation	
Es gibt keinen systematischen Unterschied		Kein Unterschied	Systematischer Unterschied Δ
Entscheidung aufgrund der beobachteten Daten in der Stichprobe	Die Daten sind mit der Nullhypothese verträglich	Richtig	Fehler 2. Art falsch negativ
	Ein systematischer Unterschied ist anzunehmen	Fehler 1. Art falsch positiv	Richtig

Abbildung 2. Fehlentscheidungen bei statistischen Tests.

2. Wir erwarten, daß der Zufallsprozeß in jeder Stichprobensequenz sichtbar ist (folglich halten wir lange gleichbleibende Sequenzen, zum Beispiel ständig „Rouge" beim Roulette, für unvereinbar mit einem Zufallsprozeß).
3. Die Abhängigkeit der Variabilität vom Stichprobenumfang wird von uns nicht wahrgenommen.
4. Darüber hinaus nehmen wir Vorinformationen über den Zufallsprozeß nicht wahr und gründen Vorhersagen jeweils nur auf die zuletzt beobachteten Stichproben.

Diese Heuristik führt, wie Kahnemann et al. nachgewiesen haben, zu einer extremen Überbewertung kleiner Stichproben. Zugespitzt kann man sagen, daß unsere menschliche Kognition dem Gesetz der kleinen Zahl gehorcht. Wir neigen dazu, auf der Basis von sehr kleinen Fallzahlen weitreichende allgemeingültige Schlüsse zu ziehen. Hieraus resultiert eine erhebliche Gefahr von Fehlurteilen, die auch im klinischen Alltag durchaus von Bedeutung sein kann. So werden in der Regel die Erfahrungen an den letzten fünf Patienten überbewertet und Schlußfolgerungen über therapeutische und diagnostische Verfahrensweisen aufgrund solcher Heuristiken voreilig getroffen. Einen Schutz vor solchen Vor-Urteilen bieten daher die mathematisch begründeten wahrscheinlichkeitstheoretischen Modelle der Zufallsprozesse.

Das Kausalitätsproblem in der klinischen Forschung

Das Ziel der Erkenntnisgewinnung im Bereich der Therapiestudienforschung ist oftmals der Vergleich verschiedener Therapiemodalitäten miteinander, um entweder Wirksamkeitsunterschiede oder bei gleicher Wirksamkeit Unterschiede in Nebenwirkungen, Lebensqualität oder Kosteneffizienz zu ermitteln. Bei solchen vergleichenden Fragestellungen wäre optimale Erkenntnisgewinnung dann zu erwarten, wenn man den Vergleich der Therapiemodalitäten möglichst direkt unter Ausschaltung von Störgrößen und Minderung von Zufallsschwankungen führen könnte. Dieses Ziel wäre in idealer Weise erfüllt, wenn es möglich wäre, an demselben Individuum unter identischen Bedingungen gleichzeitig die Effekte zweier Therapiemodalitäten unabhängig voneinander zu beobachten. Die Unmöglichkeit dieser experimentellen Bedingung wird auch als das Fundamentalproblem kausaler Schlußfolgerung bezeichnet. Da diese Bedingungen nicht herstellbar sind, besteht lediglich die Möglichkeit, mit Versuchswiederholung zu operieren und gleichzeitig die Erfüllung von vereinfachenden, nicht verifizierbaren Annahmen zu unterstellen. So fordert man, wie in jedem naturwissenschaftlichen Experiment, daß zeitlich konstante Versuchsbedingungen, die Homogenität der Beobachtungseinheiten und die kausale Transienz gelten. Diese Bedingungen sollen durch Versuchskonstruktion hergestellt werden. Sie sind jedoch selbst nicht oder nur unvollständig verifizierbar und haben daher den Status von vereinfachenden Annahmen. Übertragen auf die Situation von vergleichenden Studien, führt die Annahme des konstanten Effektes zu einer Annahme über die Konstanz der Wirkungsdifferenz beider Therapien, die bei allen Probanden gleichartig und zeitlich invariant ist. Unter der Voraussetzung der Effekthomogenität ist es mathematisch äquivalent, statt des Mittelwertes der intraindividuellen Wirkungsunterschiede, der ja wegen des Fundamentalproblems der Kausalität nicht bestimmbar ist, die Differenz der Mittelwerte zweier unterschiedlich behandelter Vergleichsgruppen zu bestimmen. Somit ersetzt man die Bestimmung des mittleren Wirkungsunterschieds durch die Bestimmung des Unterschieds der mittleren Wirkungen an Gruppen von Individuen. Dies gelingt allerdings nur, wenn die Zuordnung der Probanden zu den beiden Behandlungsmodalitäten so geschieht, daß die Kovariablen der Probanden oder Therapieergebnisse anderer Probanden keinen Einfluß auf die Therapiezuordnung haben. Diese Annahme der unabhängigen Therapiezuordnung entspricht der Voraussetzung der Homogenität der Beobachtungseinheiten im Hinblick auf die relevante Wirkungsdifferenz.

Als Konsequenz aus diesen Überlegungen zur Umgehung des Fundamentalproblems kausaler Schlüsse zeigt sich, weshalb Populationsbetrachtungen in der Medizin unumgänglich sind. Diese führen unvermeidlich zur statistischen Sichtweise. Darüber hinaus wird die Erfüllung der evidenzsichernden Annahmen am ehesten im Rahmen kontrollierter Studien garantiert, bei denen neben der unabhängigen Therapiezuordnung auf Vergleichsgruppen besondere Sorgfalt auf die Konstanz der Therapiebedingungen und die Kontrolle der Kovariablen mit möglichem Einfluß auf die Effektdifferenz zu legen ist. Ein Großteil des Beitrags der Biometrie zur medizinischen Erkenntnisgewinnung liegt heute darin, zu präzisieren, wie diese Voraussetzungen geschaffen und kontrolliert werden können.

Ein wichtiges Instrument zur Herstellung dieser Voraussetzungen ist die Zufallszuordnung, auch Randomisation genannt. Sie ist definierbar als zufällige Zuordnung von Versuchspersonen zu zwei oder mehreren Maßnahmen, über deren Wirksamkeit vergleichbare Unsicherheit besteht. Der wesentliche Vorteil der Randomisation ist, daß man in einer prospektiven Studie die Einflüsse von unbekannten systematischen Störgrößen und Selektionsprozessen gleichartig auf die Behandlungsgruppen verteilt. Hierdurch werden sie strukturgleich, und die interessierende Behandlung ist die einzige unterschiedliche Einflußgröße. Die Randomisation ist somit ein wesentliches Hilfsmittel für die Durchführung kontrollierter Ursachen-Wirkungs-Forschung im naturwissenschaftlichen Sinne. Allerdings ist die Randomisation allein nicht ausreichend, um die Vergleichbarkeit von Therapiemodalitäten herzustellen. Sie ist lediglich eine Verfahrensweise, die Strukturgleichheit der Populationen herzustellen. Weitere zum Teil aufwendige Maßnahmen müssen ergriffen werden, um so weit wie möglich Gleichheit bei den Durchführungsbedingungen der Therapien, bei der Beobachtung und Messung der Ergebnisse sowie bei der Auswertung und Bewertung der Ergebnisse herzustellen. Eine große Vielzahl von Verzerrungsmöglichkeiten ist bekannt und gefürchtet, und folglich sind die Anstrengungen erheblich, wenn diese Abweichungen von einem idealen Versuchsplan minimiert werden sollen. Detaillierte Anforderungen im Bereich der Arzneimittelzulassungsstudien sind in gesetzlichen Bestimmungen und in Verwaltungsrichtlinien niedergelegt. Es seien hier nur die „Guidelines for Good Clinical

Practice" (GCP) genannt. Auch Verfahrensweisen, wie die Blindung von Patient und behandelndem Arzt im sogenannten Doppelblindversuch, dienen dazu, solche Vergleichbarkeiten herzustellen, sofern diese vertretbar und machbar sind.

Die Bedeutung randomisierter kontrollierter prospektiver Studien liegt vor allem darin, daß sie helfen, übertriebene Erwartungen an neue therapeutische Modalitäten zu kontrollieren und falsch positive Resultate zu vermeiden. Damit stellen sie ein wesentliches Instrument zur Sicherung von Evidenz dar. Dies sei an zwei Beispielen erläutert.

In den 70er Jahren wurde in die Behandlung des hochmalignen Non-Hodgkin-Lymphoms eine multizyklische zytotoxische Chemotherapie nach den CHOP-Schema eingeführt. Sie erwies sich als wirksam und gestattete das Erreichen von anhaltenden Dauerremissionen bei einem größeren Anteil von Patienten. In der Folgezeit wurde das Prinzip der Polychemotherapie weiter ausgedehnt. Neue Schemata mit bis zu zehn unterschiedlichen Substanzen wurden eingeführt. Zahlreiche Studiengruppen berichteten über nicht vergleichende Beobachtungsstudien an ausgewählten Patientenpopulationen. Es entstand der Eindruck, zumindest in der publizierten Literatur, daß diese neuen Schemata dem CHOP-Schema überlegen wären. Erst Ende der 80er Jahre traten Zweifel an diesen unkontrollierten Studien auf und führten zur Konzeption einer prospektiv randomisierten Studie in der amerikanischen SWOG-Studiengruppe [5]. Dabei wurde an ca. 900 Patienten das klassische CHOP-Schema mit drei modernen Schemata verglichen. Die Ergebnisse waren ernüchternd, denn alle Schemata zeigten identische Wirksamkeiten, jedoch waren die modernen Schemata deutlich nebenwirkungsreicher als das CHOP-Schema. Die SWOG-Studie zeigte somit, daß CHOP weiterhin als Referenzstandard für alle Therapievergleiche zu gelten hat und daß solche Vergleiche unbedingt erforderlich sind, um Patienten vor unnötigen Risiken zu schützen.

Als zweites Beispiel sei die HD4-Studie der Deutschen Hodgkin-Studiengruppe unter Leitung von Herrn Prof. Diehl angeführt. Bei dieser Studie sollten Patienten in limitierten Stadien (PS I, II ohne Risikofaktoren) mit einer Großfeldbestrahlung behandelt werden. Im randomisierten Vergleich sollten zwei verschiedene Strahlendosen (30 versus 40 Gy) verglichen werden. Die Patienten wurden in den Jahren 1988 bis 1993 eingeschlossen. Bei einer ersten Zwischenauswertung im Jahr 1994 zeigte sich erstaunlicherweise eine schlechtere Tumorkontrolle in der Population, die eine höhere Strahlendosis erhalten hatte [2]. Der Unterschied war statistisch signifikant. Dieses unplausible Ergebnis führte zu der Frage, ob die statistische Signifikanz auch eine klinische Relevanz hatte oder ob wir bei dieser Zwischenanalyse lediglich einen zwar seltenen, aber möglichen statistischen Zufallseffekt sahen. Immerhin bestand die Möglichkeit, daß die behandelnden Ärzte die Patienten, für die eine höhere Therapiedosis vorgesehen waren, in ihren Feldbemessungen anders (enger) behandelt hatten als in der Vergleichsgruppe. Dies retrospektiv herauszufinden wäre zweifellos kaum möglich gewesen. Jedoch hatte die Studiengruppe schon zu Beginn der Studie festgelegt, daß von allen Patienten alle Röntgen- und CT-Bilder des Primärbefalls sowie die Kontrollaufnahmen und Dosimetrieberechnungen an ein unabhängiges Gutachtergremium von Strahlentherapeuten eingesandt werden mußten, die in Unkenntnis der Therapieergebnisse beurteilten, ob die Feld- und Dosiswahl korrekt durchgeführt worden ist. Es bestand somit eine etablierte Reviewbeurteilung der Therapiemaßnahmen. Reviewbeurteilungen sind Beurteilungen von nicht oder nur schwer objektivierbaren diagnostischen oder therapeutischen Merkmalen nach einem festgelegten Verfahren durch unabhängige Beobachter, die hinsichtlich aller anderen Merkmale in Unkenntnis gelassen werden. Damit entsteht eine Blindung der Beurteiler und in deren Folge eine Gleichheit der Beurteilungen für alle Probanden. Die Durchführung von Reviewbeurteilungen stellt sich somit als ein wesentliches Element zur Herstellung der Beobachtungsgleichheit im Rahmen kontrollierter Studien heraus. Im Rahmen der HD4-Studie erwies sich nun, daß die Protokollabweichungen in beiden Therapiearmen gleich häufig und von gleicher Art waren, so daß sie die Unterschiede in dem therapeutischen Ergebnis nicht erklären konnten. Damit war klar, daß der statistisch signifikante Therapieeffekt nicht bedeutsam war, da insbesondere keine Verzerrungen in der Studiendurchführung aufgetreten waren und man das Ergebnis als falsch positiven Effekt werten konnte. Als klinische Schlußfolgerung aus diesem Ergebnis war jedoch ebenfalls eindeutig, daß mit einer niedriger dosierten Strahlentherapie keine therapeutische Unterlegenheit gegenüber der Standarddosierung resultierte.

Das Problem der nichtsignifikanten Studien

Randomisierte prospektive Studien sind in den vergangenen Jahren ein zunehmend wichtigeres Instrument in der klinischen Forschung geworden. Sie haben wesentlich dazu beigetragen, übertriebene Erwartungen zu kontrollieren und neue therapeutische Verfahren mit etablierten Standards zu vergleichen. Ein wesentliches Problem heutzutage ist jedoch, daß eine große Zahl der durchgeführten und publizierten Studien unterdimensioniert sind und keine signifikanten Unterschiede in den Vergleichen aufdecken. Beispielsweise hat die European Organisation for the Research and Treatment of Cancer (EORTC) vor einigen Jahren bei 71 durchgeführten Studien lediglich in fünf Fällen signifikante Unterschied nachweisen können. Die medianen Fallzahlen in allen Studien lagen bei 240 Patienten. Mit diesen Fallzahlen ist es möglich, einen therapeutischen Unterschied im Überleben von etwa 20% mit einer Zuverlässigkeit von 80% aufzudecken. Therapeutische Fortschritte in dieser Größenordnung sind jedoch im Bereich der onkologischen Erkrankungen extrem selten. Die durchgeführten Studien müssen daher als unterdimensioniert angesehen werden, da ja nur kleinere Effekte realistisch gewesen wären, zu deren Aufdeckung man viel höhere Fallzahlen hätte anstreben müssen. Das Problem falsch negativer nichtsignifikanter Studien besteht darin, daß Studien mit zu kleinen Fallzahlen möglicherweise medizinisch therapeutische Effekte nicht aufzudecken und zu erkennen gestatten.

Unterdimensionierte Studien stellen insofern eine Gefahr für den medizini-

schen Fortschritt dar, als möglicherweise relevante, aber kleinere therapeutische Unterschiede übersehen und fälschlich als nicht vorhanden erklärt werden. Das Vorliegen auch kleiner, aber relevanter Vorteile wurde insbesondere mittels der Techniken der Metaanalyse im Bereich der Mammakarzinome durch die Gruppe um R. Peto in Oxford nachgewiesen (Early Breast Cancer Trialists' Collaborative Group [3, 4]). Unter Metaanalyse versteht man eine gemeinsame Auswertung vieler ähnlicher Studien mittels eines gemeinsamen statistischen Modells. Mit dieser Sicht gelingt es, auch kleinere Effekte aufzudecken und eine größere Repräsentativität zu erreichen. Andererseits sind Metaanalysen insofern angreifbar, als sie Studien unterschiedlicher Qualität und auch nicht genau identischen Designs miteinander in Beziehung setzen. Die Techniken der Metaanalysen tragen trotz aller Kritik in erheblichem Maß zur Sicherung der Evidenz im medizinischen Erkenntnisprozeß bei und gestatten es vor allem, auf kleinere relevante Effekte hinzuweisen, die bei unterdimensionierten Studien leicht übersehen werden. Dennoch ist unstrittig, daß die Durchführung ausreichend aussagekräftiger großer Studien, die auch auf die Aufdeckung moderater Effekte abzielen, in jedem Fall einer Metaanalyse überlegen ist, wenn die Durchführung der Studie alle oben genannten Qualitätsanforderungen einer kontrollierten Studie erfüllt.

Modellbasierte Studienplanung als neuer Beitrag der Biometrie zur Entwicklung therapeutischer Konzepte

Die Biometrie hat nach meiner Auffassung in Zukunft ein weiteres innovatives Arbeitsfeld vor sich. Sie muß und kann einen Beitrag zum Aufbau quantitativer Krankheitsmodelle leisten, auf deren Basis die Rationale für die Identifikation relevanter Studienfragestellungen zu gewinnen ist. Dabei sind ein detailliertes inhaltliches Verständnis der Krankheitsspezifika und die Auswahl eines adäquaten Modells der grundlegenden biologischen und therapeutischen Effekte erforderlich. Dies sei exemplarisch an einem Beispiel aus der onkologischen Chemotherapie illustriert.

Beim Morbus Hodgkin des Erwachsenen kommen durch die bisher übliche Polychemotherapie im fortgeschrittenen Stadium etwa 60% der Patienten in eine anhaltende Dauerremission. Die Chemotherapie ist somit im Grundsatz kurativ wirksam, jedoch wird ein Teil der Patienten nicht dauerhaft profitieren. Nachdem andere therapeutische Konzepte erfolglos blieben, kam Anfang der 90er Jahre der Gedanke auf, ob man mit den im Grundsatz wirksamen Chemotherapien dadurch eine Verbesserung erzielen könnte, daß man sie höher dosiert. Die wesentliche unbeantwortete Frage in diesem Kontext war jedoch, ob es bei der Tumorchemotherapie überhaupt eine relevante Dosis-Wirkungs-Beziehung bei Standardschemata gibt. Es war nicht bewiesen, daß mehr Dosis mehr Effekt verursacht. Darüber hinaus war unklar, welche Dosissteigerung überhaupt nötig sein würde, um einen klinischen Fortschritt zu erzielen, falls es eine positive Steigerung der Dosis-Wirkungs-Beziehung gibt. Aus klinischer Sicht ergaben sich im wesentlichen zwei mögliche Strategieoptionen. Die in den 90er Jahren bevorzugte Option bestand darin, eine Teilgruppe von prognostisch ungünstigen Patienten auszuwählen und diese mit massiven Dosissteigerungen und autologem Stammzelltransfer zu behandeln. Die Frage jedoch war, wie man solche Patienten auswählen soll und wie hoch diese massive Dosissteigerung ausfallen müßte. Hinzu kam das Risiko, daß einige Patienten mit konventioneller Therapie ausreichend behandelt wurden und durch eine Hochdosistherapie unnötigen Risiken ausgesetzt werden könnten. Eine alternative Strategieoption bestand darin, bei allen Patienten ohne Auswahl eine nur mäßige Dosissteigerung durchzuführen. Dies würde den Auswahlprozeß ersparen, aber die Frage nach dem Umfang der Dosissteigerung bestand auch bei dieser Option. Vorteilhaft erschien die letzte Option deshalb, weil eine Selektion von prognostisch ungünstigen Patienten aufgrund der damaligen Datenlage nicht möglich erschien. Somit bestand die Frage, ob die zweitgenannte Option überhaupt realistische Erfolgsaussicht haben könnte. Um einen Eindruck über diese Fragestellung zu erhalten, entwarfen wir ein statistisches Modell des Tumorwachstums und der Effekte der Chemotherapie, das als Populationsmodell aufgefaßt werden kann. Dieses Modell sollte einerseits die Heterogenität bezüglich der zytotoxischen Effekte in der Population und andererseits die Heterogenität der Tumorwachstumsgeschwindigkeiten während und nach der Therapie beschreiben. Es sollte zudem gestatten, Dosis-Wirkungs-Beziehungen zu inkorporieren und sich an vorhandene Populationsdaten von Remissionen anzupassen. Es ist uns gelungen, ein geeignetes parametrisches Modell mit begründeten Annahmen über die Verteilung der Chemosensitivität und der Tumorlatenzzeiten zu erstellen [7]. Dieses Modell wurde an Daten einer früheren Patientenpopulation angepaßt, in der zwar im Grundsatz die gleiche Chemotherapie eingesetzt wurde, die jedoch zwischen den Patienten in Dosierung und zeitlicher Applikation mäßig variierte. Die Modellanpassung zeigte eine signifikante Dosis-Wirkungs-Abhängigkeit. Auf der Basis dieser geschätzten Abhängigkeit ließ sich vorhersagen, daß eine mittlere Dosissteigerung der Chemotherapie um 30% einen therapeutischen Gewinn an Dauerremissionen von 10 bis 15% erwarten lassen würde. Die biometrische Modellbildung also hatte hiermit deutlich gemacht, daß die Strategie einer moderaten Dosissteigerung vielversprechend sein könnte, und die Fragestellung reduzierte sich nunmehr darauf, ob eine derartige moderate Dosissteigerung überhaupt klinisch machbar ist oder ob sie mit zu hohen Risiken aufgrund der Nebenwirkungen verbunden wäre. Die limitierende Toxizität bei diesen Chemotherapien besteht in der passageren Neutropenie. Die Verfügbarkeit des rekombinanten granulozytenstimulierenden Wachstumsfaktors (G-CSF) ließ jedoch die Hoffnung aufkommen, daß ein Teil der neutropenischen Effekte durch die Gabe von G-CSF abgemildert werden könnte, so daß bei einer höheren Dosis der Chemotherapie und mit dem gleichzeitigen Einsatz von G-CSF die resultierende Toxizität vergleichbar gestaltet werden könnte wie in den bisherigen Standardschemata. Von der Deutschen Hodgkin-Studiengruppe wurde mit dem BEACOPP-Schema ein zeitlich umgestelltes Chemotherapiesche-

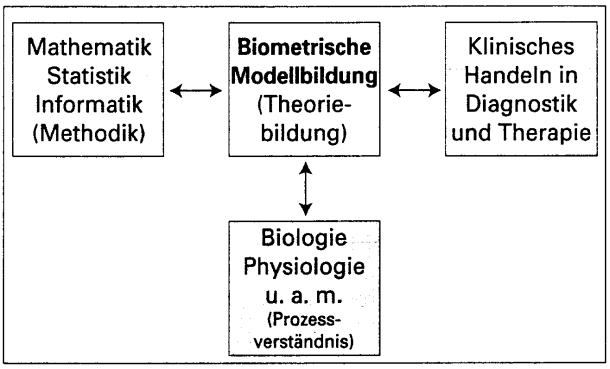

Abbildung 3. Stellung der Biometrie in der Medizin.

ma eingeführt, das gestattete, dieselben Substanzen in festgelegter zeitlicher Applikation in steigenden Dosen zu geben. Im Rahmen einer Dosisfindungsstudie wurde nun ein vorgegebenes Maß an akzeptabler Toxizität fixiert, das geringfügig über dem bisherigen Niveau lag. Die Dosisfindungsstudie definierte dann denjenigen Dosisspiegel, der mit einer solchen Toxizität vereinbar ist. Es zeigte sich im Rahmen dieser Dosisfindungsstudie, daß eine 30- bis 40%ige mittlere Dosissteigerung des BEACOPP-Schemas möglich war. Somit waren die praktischen Voraussetzungen geschaffen, mit der Strategie der moderaten Dosissteigerung konventioneller Polychemotherapie die Frage zu klären, ob es eine Dosis-Wirkungs-Beziehung bei diesen Tumoren gibt. Diese Frage führte zur Definition der HD9-Studie der Deutschen Hodgkin-Studiengruppe, in der die gleiche Chemotherapie in Basisdosierung und in gesteigerter Dosierung appliziert wurde. Die erste Zwischenauswertung der HD9-Studie deutete an, daß sich tatsächlich ein erheblicher therapeutischer Gewinn abzeichnen wird, der sogar deutlich über dem vorhergesagten Zugewinn liegen dürfte. Obwohl die Beobachtungszeiten noch keine definitiven Schlüsse zulassen, deuten die vorliegenden Befunde darauf hin, daß die durch diese Modellüberlegungen gefundene Rationale für moderate Dosissteigerung für die Hodgkin-Lymphome einen erheblichen therapeutischen Vorteil verspricht. Dies ist bedeutsam, da die Therapiestrategien großenteils ohne Hospitalisierung durchgeführt werden können. Zugleich zeigte sich, daß die bisherigen therapeutischen Optionen nicht konsequent ausgenutzt wurden. Es ist zu vermuten, daß ähnliche therapeutische Gewinne auch bei anderen onkologischen Erkrankungen zu erzielen sind.

Stellung der Biometrie in der Medizin

Abbildung 3 zeigt die Stellung der Biometrie in der Medizin aus meiner Sicht. Demnach wird in Zukunft dem Paradigma der biometrischen Modellbildung und damit der Theoriebildung in der Medizin eine zunehmende Bedeutung zukommen. Die Biometrie wird dabei eine Mittlerrolle spielen zwischen den formalen Methodiken der Mathematik, Statistik und Informatik, dem Prozeßverständnis auf der Ebene der biologischen und physiologischen Vorgänge und dem pragmatischen klinischen Handeln in Diagnostik und Therapie. Damit wird sie einerseits weiterhin methodenwissenschaftlich tätig werden, aber andererseits auch mit einem tiefen Verständnis von Biologie und Medizin einhergehen müssen. Die Verbindung zwischen diesen Ansätzen liegt nach meiner Auffassung in einem modellbasierten Theoriebildungsprozeß, aus dem heraus dann Fragestellungen über die Verifikation dieser Modelle einerseits und über die Ableitung von Handlungsrationalen andererseits folgen. Medizinische Biometrie möchte ich daher als die Wissenschaft von der Modellentwicklung in der Medizin auffassen, die der Unterstützung und Objektivierung der Erkenntnisgewinnung dient. Sie befaßt sich somit mit der Lösung medizinischer Probleme durch Auswahl und Anwendung problemadäquater Methoden und Modelle und deren Interpretation.

Literatur

1. Diehl V, et al. Interim analysis of the HD9 study of the German Hodgkin study group (GHSG) – BEACOPP is more effective than COPP-ABVD in advanced stage Hodgkin's disease. JASH 1997;90:Suppl 1:339a/1512–31.
2. Duhmke E, Diehl V, Loeffler M, et al. Randomized trial with early-stage Hodgkin's disease testing 30 Gy vs. 40 Gy extended field radiotherapy alone. Int J Radiat Oncol Biol Phys 1996;36:305–10.
3. Early Breast Cancer Trialists' Collaborative Group: Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. N Engl J Med 1988;319:1681–92.
4. Early Breast Cancer Trialists' Collaborative Group: Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Lancet 1992;339:1–15.
5. Fisher RI, Gaynor ER, Dahlberg S, et al. Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma. N Engl J Med 1993;328:1002–6.
6. Gross R, Loeffler M. Prinzipien der Medizin. Eine Übersicht ihrer Grundlagen und Methoden. Berlin–Heidelberg–New York: Springer, 1997.
7. Hasenclever D, Loeffler M, Diehl V. Rationale for dose escalation of first line conventional chemotherapy in advanced Hodgkin's disease. German Hodgkin's Lymphoma Study Group. Ann Oncol 1996;7:Suppl 4:95–8.
8. Kahnemann D, Slovic P, Tversky A, eds. Judgement under uncertainty: heuristics and biases. Cambridge/Mass.: Cambridge Univ. Press, 1982.
9. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. Am J Med 1982;72:233–40.

Korrespondenzanschrift:
Prof. Dr. Markus Löffler,
Institut für Medizinische Informatik,
Statistik und Epidemiologie
der Universität,
Liebigstraße 27,
D-04103 Leipzig,

Telefon (+49/341) 9716-100,
Fax -109,
e-mail: Loeffler@imise.uni-leipzig.de