7.2 Gesundheitsbezogene Lebensqualität

Uwe Konerding

Der Zweck medizinischer Maßnahmen besteht nicht in erster Linie darin, Kosten zu sparen, sondern vielmehr darin, in irgendeiner Weise positiv auf die Gesundheit der Patienten zu wirken. Bei gesundheitsökonomischen Studien ist diese positive Wirkung den monetären Kosten entgegenzustellen. Wie erfasst man nun aber positive Wirkungen auf die Gesundheit, sodass sie sich monetären Kosten entgegenstellen lassen? Ein Aspekt, der in diesem Zusammenhang ohne Frage von großer Bedeutung ist, ist die Anzahl der Jahre, die der Patient durch die medizinische Maßnahme gewinnt. Mit dieser Größe allein lässt sich aber kaum das sinnvoll beschreiben, was sich die meisten von uns von medizinischen Maßnahmen erhoffen. Würde man nämlich die Wirkung einer medizinischen Maßnahme ausschließlich über die gewonnenen Lebensjahre definieren, würde es als Erfolg gelten, einen Menschen jahrelang bei unerträglichen Schmerzen und völliger Hilflosigkeit am Leben zu halten. Andererseits wäre es unter dieser Voraussetzung kein Erfolg, „nur" Schmerzen zu lindern, ohne dabei auch das Leben zu verlängern. Es scheint also noch etwas anderes zu geben, was bei der Beurteilung medizinischer Maßnahmen berücksichtigt werden sollte: die Auswirkungen auf die Lebensqualität.

Was genau ist aber nun gesundheitsbezogene Lebensqualität? Wie erfasst man sie? Wie stellt man sie quantitativ dar, so dass sie sich gegen monetäre Kosten aufrechnen lässt? Um diese Fragen geht es in diesem Kapitel. Als Erstes werden dazu einige Grundbegriffe zur Definition und Messung gesundheitsbezogener Lebensqualität eingeführt. Als Zweites werden Verfahren dargestellt, die darauf abzielen, Bewertungen von Gesundheitszuständen zu erfassen. Als Drittes werden einige der Fragebögen vorgestellt, die dazu dienen sollen, lebensqualitätsrelevante Dimensionen von Gesundheitszuständen zu erfassen. Schließlich werden die Ergebnisse dieses Kapitels noch einmal zusammenfassend diskutiert.

Grundbegriffe

Definition des Begriffs „Gesundheitsbezogene Lebensqualität"

Was ist Lebensqualität? Die Antworten auf diese Frage dürften sehr unterschiedlich ausfallen. Für manche bedeutet Lebensqualität berufliche Anerkennung, für andere familiäres Glück und für wieder andere Selbstverwirklichung. Bei manchen hängt das, was sie als Lebensqualität bezeichnen, vom Zustand der Umwelt ab, bei anderen vom kulturellen Angebot und bei wieder anderen vom Wetter. Stellt man die Frage nach dem Begriff der Lebensqualität so allgemein, müssen die Antworten auf diese Frage so unterschiedlich sein. Sie wird hier deshalb auch sofort eingegrenzt: Es soll speziell um gesundheitsbezogene Lebensqualität gehen. Auch auf die Frage, was unter dieser Bezeichnung zu verstehen ist, gibt es keine eindeutige, fraglos richtige Antwort. Die Antwort auf diese Frage hängt nämlich zunächst einmal von sozialen Festlegungen ab.

Das Organ, das in diesem Zusammenhang am ehesten aufgerufen ist, eine Festlegung zu treffen, ist die Weltgesundheitsorganisation (WHO). Sie hat sich zwar nicht um die Definition des Begriffs der gesundheitsbezogenen Lebensqualität, wohl aber um den der Gesundheit bemüht. Demgemäß ist Gesundheit der „Zustand des völligen körperlichen, psychischen und sozialen Wohlbefindens und nicht nur das Freisein von Krankheit und Gebrechen" (World Health Organization [WHO] 1948). Wie jeder Versuch einer normativen Setzung hat auch dieser heftige Widersprüche ausgelöst. Vor allem ist bemängelt worden, dass die Kriterien für Gesundheit zu hoch angesetzt sind. Unter Voraussetzung des Gesundheitsbegriffs der

WHO wären nahezu alle Menschen krank. Weithin akzeptiert wird aber die in diesem Begriff enthaltene Vorstellung, dass Gesundheit aus drei Komponenten besteht: einer physischen, einer psychischen und einer sozialen. Diese drei Komponenten werden deshalb auch gerne verwendet, um den Begriff der gesundheitsbezogenen Lebensqualität weiter einzugrenzen.

> Gesundheitsbezogene Lebensqualität zeigt sich dementsprechend darin, wie zufrieden Menschen mit ihrem physischen, psychischen und sozialen Gesundheitszustand sind oder – anders ausgedrückt – wie sie diese drei Aspekte ihrer Gesundheit bewerten.

Natürlich ist diese Begriffsdefinition noch sehr allgemein. Sie begründet noch lange kein quantitatives Maß, mit dem sich die Wirkung medizinischer Maßnahmen beurteilen lässt. Zu diesem Zweck muss der Begriff der gesundheitsbezogenen Lebensqualität wesentlich weiter konkretisiert werden. So muss zum einen genauer umrissen werden, in welchen Dimensionen physische, psychische und soziale Gesundheit bzw. Krankheit zu beschreiben ist; zum anderen muss festgestellt werden, wie Menschen die verschiedenen Formen physischer, psychischer und sozialer Gesundheit bzw. Krankheit bewerten. Dies wird sich natürlich nicht von heute auf morgen bewerkstelligen lassen. Ein langer Entwicklungsprozess aus theoretischen Diskussionen und empirischen Studien wird dazu nötig sein. Im weiteren Verlauf dieses Kapitels werden die wichtigsten Ergebnisse des bisher vonstatten gegangenen Entwicklungsprozesses vorgestellt.

Verfahren zur Erfassung gesundheitsbezogener Lebensqualität

Es gibt verschiedene Arten von Verfahren, die zur Erfassung von gesundheitsbezogener Lebensqualität eingesetzt werden. Sie haben unterschiedliche Charakteristika und sind entsprechend für unterschiedliche Zielsetzungen geeignet. Im Folgenden werden zunächst diese unterschiedlichen Verfahrenstypen eingeführt und dann ihre speziellen Verwendungsmöglichkeiten erörtert.

Unterscheidungen

Die erste Unterscheidung, die für Verfahren zur Erfassung gesundheitsbezogener Lebensqualität von Bedeutung ist, ist die zwischen **nutzentheoretischen** und **psychometrischen** Verfahren (vgl. Schöffski et al. 1998). Bei den nutzentheoretischen Verfahren handelt es sich um verschiedene Techniken, mit denen man Bewertungen erheben kann. Im einfachsten Fall bittet man Menschen, ihre Bewertung auf einer Skala mit zwei definierten Endpunkten abzugeben. Die nutzentheoretischen Verfahren dienen dazu, direkt die Bewertungen von Gesundheitszuständen zu erfassen. Bei den psychometrischen Verfahren handelt es sich dagegen um Fragebögen mit Fragen zu verschiedenen Aspekten von Krankheit bzw. Gesundheit, wie etwa Schmerzen oder Einschränkungen der Bewegungsfähigkeit. Diese Fragebögen dienen dazu, die Gesundheitszustände der Patienten in lebensqualitätsrelevanten Dimensionen zu erfassen. Durch die Auswahl der Fragen, die in einen derartigen Fragebogen aufgenommen werden, wird bei jedem Fragebogen der Begriff der Lebensqualität fragebogenspezifisch konkretisiert.

Die zweite bei der Messung von Lebensqualität wichtige Unterscheidung ist die zwischen **Profil-** und **Indexinstrumenten**. Diese Unterscheidung betrifft nur die psychometrischen Messverfahren. Bei einem Profilinstrument werden Gesundheitszustände durch Abstufungen auf verschiedenen Dimensionen beschrieben. Diese Abstufungen bilden das jeweilige Krankheitsprofil. Ein solches Profil könnte etwa aus den Werten für die zur Zeit der Befragung empfundenen Schmerzen, Ängstlichkeit, Einschränkungen bei der Fähigkeit zur Fortbe-

wegung und Einschränkungen bei der Fähigkeit zur Körperpflege bestehen. Bei einem Indexinstrument werden die Gesundheitszustände ebenfalls durch Abstufungen auf verschiedenen Dimensionen beschrieben. Zusätzlich dazu wird aber noch auf der Grundlage dieser Abstufungen ein einziger Wert, ein sog. Index, berechnet, der die Gesamtlebensqualität bei dem auf diese Weise beschriebenen Gesundheitszustand widerspiegeln soll. Hier würden also beispielsweise Bewertungen von Schmerzen, Ängstlichkeit, Einschränkungen bei der Fähigkeit zur Fortbewegung und Einschränkungen bei der Fähigkeit zur Körperpflege gegeneinander aufgerechnet. Die Unterscheidung zwischen Profil- und Indexinstrumenten ist damit keine Entweder-oder-Unterscheidung. Vielmehr liefert jedes Indexinstrument auch ein Profil, aber nicht jedes Profilinstrument einen Index.

Die dritte Unterscheidung ist die zwischen **krankheitsspezifischen** und **generischen Messinstrumenten**. Krankheitsspezifische Verfahren zielen darauf ab, die Beschwerden zu erfassen, die für bestimmte Krankheiten spezifisch sind. So gibt es mittlerweile krankheitsspezifische Instrumente für Krebs, Herz-Kreislauf-Erkrankungen, neurologische Beschwerden, Erkrankungen der Atemwege und rheumatische Beschwerden (vgl. Salek 1999). Generische Messinstrumente sind dagegen für den krankheitsübergreifenden Einsatz gedacht. Die Unterscheidung zwischen krankheitsspezifischen und generischen Verfahren ist vor allem für psychometrische Verfahren von Bedeutung. Diese Unterscheidung ist – zumindest rein analytisch – unabhängig von der Unterscheidung zwischen Profil- und Indexinstrumenten. Es kann also sowohl krankheitsspezifische Profilinstrumente als auch krankheitsspezifische Indexinstrumente geben. Ebenso kann es sowohl generische Profilinstrumente als auch generische Indexinstrumente geben.

Die vierte und letzte Unterscheidung ist die zwischen **Selbst-** und **Fremdeinschätzung**. Auch diese Unterscheidung betrifft hauptsächlich die psychometrischen Verfahren. Selbsteinschätzung besteht darin, dass der Patient seinen Gesundheitszustand selbst einschätzt. Bei der Fremdeinschätzung liefert jemand anders diese Einschätzung. Vorzugsweise handelt es sich dabei um einen nahen Angehörigen, den behandelnden Arzt oder die betreuende Schwester. Die meisten psychometrischen Verfahren sind als Instrumente zur Selbsteinschätzung gedacht. Sie können oft aber auch zur Fremdeinschätzung benutzt werden.

Anwendungsbedingungen

Wann sind nutzentheoretische und wann psychometrische Verfahren angezeigt? Die Antwort auf diese Frage hängt zunächst einmal davon ab, ob ein Krankheitsprofil oder ein zusammenfassender Lebensqualitätswert benötigt wird. Krankheitsprofile lassen sich nur mithilfe psychometrischer Verfahren erfassen. Zusammenfassende Lebensqualitätswerte können dagegen – im Prinzip – sowohl mithilfe nutzentheoretischer Verfahren als auch mithilfe psychometrischer Indexinstrumente bestimmt werden. Im ersten Fall würde man den Patienten seinen aktuellen Gesundheitszustand selbst bewerten lassen, im zweiten Fall würde man erst Fragen zu verschiedenen Aspekten des aktuellen Gesundheitszustandes stellen und dann aufgrund der Antworten zu diesen Fragen den Gesamtindex für die aktuelle Lebensqualität berechnen.

Beide Ansätze zur Bestimmung eines zusammenfassenden Lebensqualitätswertes haben spezielle Vor- und Nachteile. Bei der psychometrischen Indexmessung beruht der letztlich resultierende Lebensqualitätswert damit immer auf wesentlich mehr theoretischen Vorannahmen und empirischen Vorinformationen oder – schlimmer noch – auf mehr willkürlichen Setzungen als der Wert, der durch ein nutzentheoretisches Verfahren bestimmt wird. Die Vorannahmen und Vorinformationen können falsch und die Setzungen unsinnig sein. Dies spräche eher für die nutzentheoretischen Ver-

fahren. Andererseits sind bei der psychometrischen Indexmessung die Aspekte, auf denen die Gesamtbewertung beruht, klar festgelegt. Bittet man dagegen einen Patienten, seinen aktuellen Gesundheitszustand zu bewerten, ist weitgehend unklar, an welche Aspekte des Gesundheitszustandes der Patient dabei denkt. Damit ist fragwürdig, was der auf diese Weise gewonnene Messwert bedeutet.

Die nutzentheoretischen Verfahren erfüllen allerdings noch eine weitere Funktion, die psychometrische Verfahren in dieser Form nicht erfüllen können. Sie werden bei der empirischen Forschung zur Entwicklung von Indexinstrumenten benötigt. Jeder Index entsteht dadurch, dass aufgrund von Werten aus so unterschiedlichen Dimensionen wie Schmerzen, Ängstlichkeit, Fähigkeit zur Fortbewegung und Fähigkeit zur Körperpflege ein Gesamtwert zur Lebensqualität bestimmt wird. Um hierfür sinnvolle Berechnungsvorschriften entwickeln zu können, braucht man Informationen darüber, wie Menschen die verschiedenen Abstufungen auf diesen Dimensionen oder – besser noch – die verschiedenen möglichen Kombinationen von Abstufungen auf diesen verschiedenen Dimensionen bewerten. Diese Informationen lassen sich am ehesten durch empirische Untersuchungen beschaffen, in denen Probanden hypothetische Gesundheitszustände, die durch Abstufungen auf diesen Dimensionen beschrieben sind, mithilfe nutzentheoretischer Verfahren bewerten.

Wann braucht man ein Profil und wann einen Index? Die Antwort auf diese Frage hängt von dem Zweck ab, dem die Untersuchung dienen soll. Wenn es darum geht, die Art der Wirkung einer medizinischen Maßnahme zu bestimmen, benötigt man ein Profil. Dieses Profil kann differenziert anzeigen, auf welche Aspekte der Lebensqualität sich die Maßnahme auswirkt und auf welche nicht. Diese Informationen sind insbesondere für die klinische Forschung von großer Bedeutung. Des Weiteren könnten diese Informationen auch wichtig sein, wenn Entscheidungen über ergänzende Maßnahmen zu treffen sind. Wenn es dagegen darum geht, die verschiedenen Wirkungen einer medizinischen Maßnahme gegeneinander abzuwägen, wird ein Index benötigt. Lebensqualitätsindizes sind damit speziell für die Beantwortung gesundheitsökonomischer Fragestellungen von Bedeutung.

Ob man eher krankheitsspezifische oder eher generische Messinstrumente einsetzen sollte, hängt im Wesentlichen davon ab, welche Art von medizinischen Maßnahmen miteinander verglichen werden sollen. Wenn es um verschiedene Behandlungsmethoden für die gleiche Krankheit geht, sind krankheitsspezifische Verfahren angezeigt. Diese Verfahren beziehen sich auf genau die Dimensionen der Lebensqualität, auf die die fragliche Krankheit einwirkt. Sie sind damit meist wesentlich besser als generische Messinstrumente geeignet, Änderungen bezüglich dieser Dimensionen festzustellen. Wenn dagegen Behandlungsmethoden für unterschiedliche Krankheiten miteinander verglichen werden sollen, wie es bei Entscheidungen über Mittelzuweisungen häufig der Fall ist, werden generische Verfahren benötigt.

Als Letztes bleibt die Frage, in welchen Fällen eine Selbsteinschätzung und in welchen Fällen eine Fremdeinschätzung angezeigt ist. Hier gibt es mindestens ein klares Kriterium, das gegen die Selbsteinschätzung spricht. Wenn der Patient nicht willens, nicht in der Lage oder weder willens noch in der Lage ist, über sich Auskunft zu geben, bleibt nur der Weg der Fremdeinschätzung. Bei Patienten, die über sich Auskunft geben können und wollen, hängt es wieder von den speziellen Zielsetzungen ab, welcher von beiden Verfahrensweisen der Vorzug zu geben ist. Wenn man ausdrücklich an dem subjektiven Befinden interessiert ist, ist die Selbsteinschätzung die Methode der Wahl. Wenn es dagegen darum geht, bestimmte äußere Bedingungen der Lebensqualität, wie etwa Einschränkungen der Beweglichkeit, möglichst unabhängig von subjektiven Wahrnehmungsverzerrungen zu erfassen, ist die Fremdbeobachtung günstiger. In diesem Fall wäre es auch besser, die Fremdbeobachtung nicht nahe ste-

henden Personen zu überlassen, sondern dafür eigens geschulte Beobachter einzusetzen.

Allgemeine Begriffe zur Beschreibung von Messinstrumenten

Die Begriffe, die hier zum Schluss eingeführt werden, sind allgemeinerer Art. Sie sind nicht nur im Zusammenhang mit der Erfassung gesundheitsbezogener Lebensqualität von Bedeutung, sondern betreffen nahezu alle Arten von Messungen. Zwei Arten von Begriffen, die sich auf zwei unterschiedliche Aspekte von Messungen beziehen, werden diskutiert: messtheoretische und testtheoretische Begriffe. Messtheoretische Begriffe betreffen die strukturellen Merkmale der Information, die durch die Messwerte widergespiegelt wird. Testtheoretische Begriffe betreffen die Güte der Messverfahren.

Messtheoretische Begriffe

Messen besteht darin, Objekten Zahlen zuzuordnen, sodass die numerischen Relationen zwischen den Zahlen empirische Relationen zwischen den Objekten widerspiegeln (vgl. u. a. Orth 1974). Der Begriff „Objekt" ist dabei sehr allgemein gemeint. Demgemäß sind nicht nur tote Gegenstände, sondern auch Menschen Objekte. Unter dem Begriff „empirische Relation" fallen auch sehr unterschiedliche Dinge. Eine empirische Relation zwischen zwei Menschen kann etwa darin bestehen, dass beide das gleiche Geschlecht haben. Eine weitere Relation könnte darin bestehen, dass die eine Person besser Tennis spielt als die andere. Eine Relation zwischen vier Menschen a, b, c und d könnte darin bestehen, dass der Intelligenzunterschied zwischen a und b kleiner ist als der zwischen c und d. Eine Relation zwischen drei Menschen e, f und g kann darin bestehen, dass e und f zusammengenommen genauso schwer sind wie g alleine. Wenn e und f außerdem noch gleich schwer sind, ergibt sich als weitere empirische Relation, dass g doppelt so schwer ist wie e und ebenfalls doppelt so schwer ist wie f.

Messungen können sich darin unterscheiden, in welchem Ausmaß sie empirische Relationen widerspiegeln. Das Ausmaß, in dem eine Messung dies tut, bezeichnet man als Skalenniveau. Je mehr empirische Relationen widergespiegelt werden, desto höher ist das Skalenniveau. Wenn die Messwerte lediglich Gleichheit oder Unterschiedlichkeit widerspiegeln, liegt Nominalskalenniveau vor. Dies wäre etwa gegeben, wenn man das Geschlecht durch zwei verschiedene Zahlen darstellt. Wenn zusätzlich durch die Zahlen Rangfolgebeziehungen widergespiegelt werden, liegt Ordinalskalenniveau vor. Ein Beispiel ist die Weltrangliste für Tennisspieler. Wenn zusätzlich dazu noch empirische Relationen widergespiegelt werden, die die gleiche Struktur aufweisen, wie sie auch bei Abständen besteht, liegt Intervallskalenniveau vor. Ein Beispiel sind hier wohl manche Intelligenztests. Wenn zusätzlich dazu noch empirische Relationen widergespiegelt werden, die die gleiche Struktur aufweisen, wie sie auch bei Verhältnissen zwischen Zahlen besteht, liegt Rationalskalenniveau vor. Beispiele hierfür sind die Gewichtsmessung in Gramm, Kilogramm oder Zentnern sowie die Längenmessung in Zentimetern, Metern oder Meilen. Intervall- und Rationalskalen werden oft auch zusammenfassend als Kardinalskalen bezeichnet.

Bei der Messung gesundheitsbezogener Lebensqualität, wie sie etwa mit einem Indexinstrument angestrebt wird, sind die Objekte Menschen. Die empirischen Relationen sind Relationen wie: „Die Person a fühlt sich bezüglich ihrer Gesundheit wohler als die Person b" oder „Die Personen a und b unterscheiden sich in ihrem auf die Gesundheit bezogenen Wohlbefinden weniger als die Personen c und d". Aus forschungslogischer Sicht sind diese Relationen äußerst problematisch. Sie können nämlich nicht ohne Messverfahren bestimmt werden, die bereits quantitative Werte liefern. Damit besteht die Gefahr, dass durch das Messverfahren etwas konstruiert wird, was es ohne das Messverfahren überhaupt nicht gäbe. Um dieser Gefahr zu entgegnen, werden Indexver-

fahren über Vorstudien begründet. Mithilfe nutzentheoretischer Verfahren werden Bewertungen für die verschiedenen Gesundheitszustände erhoben, die durch die zum Indexinstrument gehörenden Dimensionen beschrieben werden können. Die auf diese Weise bestimmten Bewertungen der Gesundheitszustände liefern letztlich den Index. Die empirischen Relationen, gegenüber denen Indexinstrumente zur Lebensqualität begründet werden, sind damit im Wesentlichen Bevorzugungsrelationen zwischen Beschreibungen von Krankheitszuständen.

Trotz dieses Kunstgriffs bestehen bei der Messung von Lebensqualität noch einige ungeklärte Probleme. So ist zurzeit noch offen, welche Aspekte von Gesundheitszuständen in welcher Weise zur Bestimmung von gesundheitsbezogener Lebensqualität zu berücksichtigen sind. Weiter gibt es verschiedene Auffassungen darüber, welche nutzentheoretischen Verfahren in welcher Weise die wirklich wichtige Information liefern. Mit anderen Worten: Es besteht noch in vielerlei Hinsicht Uneinigkeit über die empirische Basis, die durch Messverfahren zur Lebensqualität widergespiegelt werden soll. Selbst wenn hier Einigkeit erzielt werden sollte, wären damit nicht alle Probleme gelöst. So urteilen Menschen selten perfekt konsistent, und unterschiedliche Menschen urteilen unterschiedlich. Die empirischen Relationen, die durch das Messinstrument widergespiegelt werden sollen, sind damit deutlich weniger gut strukturiert als die Zahlen, die das Messinstrument liefert. Realistischerweise ist von einem Indexinstrument zur Lebensqualität lediglich zu erwarten, dass es mittlere Tendenzen in den Urteilen der befragten Personen widerspiegelt. Nur in diesem Sinne werden Messungen auf Ordinal- oder sogar Intervallskalenniveau möglich sein.

Für die gesundheitsökonomische Analyse ist die Frage nach dem Skalenniveau, auf dem gesundheitsbezogene Lebensqualität erfasst werden kann, von entscheidender Bedeutung. Spätestens dann, wenn darüber entschieden werden soll, ob die durch eine bestimmte medizinische Maßnahme gewonnene Lebensqualität die zusätzlichen Kosten rechtfertigt, ist es unerlässlich, dass die Lebensqualität in sinnvoller Weise auf Intervallskalenniveau gemessen wird. Als Grundlage für Entscheidungen dieser Art wird nämlich im Allgemeinen die Differenz zwischen zwei Lebensqualitätswerten mit der Differenz zwischen zwei monetären Werten ins Verhältnis gesetzt. Kenngrößen dieser Art haben nur dann eine stabile empirische Bedeutung, wenn sowohl die Größen im Zähler als auch die Größen im Nenner wenigstens auf Intervallskalenniveau gemessen werden. Anderenfalls spiegeln diese Kenngrößen im Wesentlichen willkürliche Setzungen der Forscher wider, nicht aber echte empirische Relationen, wie etwa Unterschiede in der gesundheitsbezogenen Lebensqualität.

Testtheoretische Begriffe

Messverfahren, die im Prinzip das gleiche Ausmaß an struktureller Information widerspiegeln, können dies in unterschiedlicher Güte tun. Dabei gibt es verschiedene Kriterien, von denen diese Güte abhängt. Diese Kriterien werden im Folgenden diskutiert. Es handelt sich dabei um:

- die Validität
- die Reliabilität
- die Objektivität
- die Praktikabilität

Die ersten drei Kriterien werden gelegentlich auch als die klassischen drei testtheoretischen Kriterien bezeichnet (Schöffski 1998). Sie werden wohl in jedem Lehrbuch zur Testtheorie diskutiert (s. u.a. Rost 1996). Für das letzte Kriterium gilt dies nicht unbedingt. Zudem werden für dieses Kriterium oft auch andere Bezeichnungen verwendet. Bei der Messung von Lebensqualität, insbesondere bei gesundheitsökonomischen Studien, spielt dieses Kriterium aber eine wichtige Rolle. Aus diesem Grund wird es hier zusammen mit den drei klassischen testtheoretischen Kriterien eingeführt.

Validität heißt übersetzt Gültigkeit. Ein Messverfahren ist in dem Maße valide, wie es das erfasst, was es erfassen soll (vgl. Rost 1996). Ein Messverfahren zur Lebensqualität ist dementsprechend dann valide, wenn es Lebensqualität erfasst, und nicht etwa Testängstlichkeit oder Intelligenz. Die Validität eines Messverfahrens lässt sich natürlich nur insoweit beurteilen, wie man theoretisch festgelegt hat, was unter der zu erfassenden Größe zu verstehen ist. Soweit diese Voraussetzung erfüllt ist, ergeben sich verschiedene Kriterien, anhand derer sich die Validität prüfen lässt.

Ein Kriterium für die Validität ist die Auswahl der einzelnen Fragen – oder allgemeiner Items –, aus denen das Messverfahren gebildet wird. Diese Items sollten möglichst repräsentativ den Bereich der empirischen Phänomene abdecken, in denen sich die zu erfassende Größe gemäß Definition zeigen müsste. So sollte ein Messverfahren zur Lebensqualität möglichst genau die Aspekte von Gesundheitszuständen umfassen, die gemäß Definition für die Lebensqualität entscheidend sind. Diese Form der Validität heißt Inhaltsvalidität. Weiter zeigt sich die Validität eines Messverfahrens darin, inwieweit die Messergebnisse mit den Größen zusammenhängen, mit denen sie gemäß theoretischer Definition zusammenhängen müssten. Dabei lassen sich wieder verschiedene Formen der Validität unterscheiden. Dies kann hier nicht vollständig ausgeführt werden. In der weiteren Diskussion sind aber insbesondere zwei Formen von Bedeutung: die konkurrente und die diskriminante Validität. Konkurrente Validität besteht, wenn das Messverfahren im Wesentlichen die gleichen Ergebnisse liefert wie andere Messverfahren, die dasselbe erfassen sollen. Diskrimante Validität besteht, wenn das Messverfahren dort Unterschiede aufdeckt, wo gemäß Definition des Konstruktes Unterschiede bestehen müssten. Die diskriminante Validität wird oft auch als Sensitivität bezeichnet.

Reliabilität heißt übersetzt Zuverlässigkeit. Ein Messverfahren ist in dem Maße reliabel, wie es unter gleichen Bedingungen die gleichen Ergebnisse liefert. Es gibt verschiedene Ansätze, die Reliabilität von Messverfahren zu bestimmen. Ein Ansatz besteht darin, das gleiche Messverfahren mit einem gewissen Abstand zweimal hintereinander anzuwenden und die Ergebnisse beider Messungen miteinander zu vergleichen. Die auf diese Weise bestimmte Reliabilität bezeichnet man als Test-Retest-Reliabilität. Ein zweiter Ansatz besteht darin, zwei möglichst gleichwertige Varianten des gleichen Verfahrens herzustellen, beide Varianten anzuwenden und die Ergebnisse miteinander zu vergleichen. Man redet dann von Test-Paralleltest-Reliabilität. Bei Messverfahren, die, wie die meisten psychologischen Tests, auf mehreren Einzeldaten beruhen, besteht ein dritter Ansatz darin, die Einzeldaten in zwei möglichst gleiche Hälften zu gliedern und die Ergebnisse beider Hälften miteinander zu vergleichen. Man redet dann von Testhälften-Reliabilität (split-half-reliability). Dieser Ansatz kann dahingehend verallgemeinert werden, dass man jedes einzelne Datum als unabhängige Messung derselben Größe betrachtet und aus der internen Struktur dieser Daten die Reliabilität bestimmt. Diese Form der Reliabilität bezeichnet man als interne Konsistenz (vgl. Rost 1996).

Der Begriff der **Objektivität** bedarf kaum einer Übersetzung. Ein Messverfahren ist in dem Maße objektiv, in dem es unabhängig von den Personen, die es anwenden, die gleichen Ergebnisse liefert. Drei Arten der Objektivität werden unterschieden: die Ausführungs-, die Auswertungs- und die Interpretationsobjektivität. Ein Verfahren ist ausführungsobjektiv, wenn die Personen, die das Verfahren anwenden, keinen Einfluss auf die Ergebnisse ausüben können. Dies dürfte in hohem Maße bei einer schriftlichen Befragung der Fall sein, bei der für die Befragten erkennbar Anonymität gewährleistet ist. Wenn dagegen die Befragung in Form eines Interviews durchgeführt wird, bei dem sich der Anwender des Verfahrens und der Befragte direkt gegenübersitzen, ist die Ausführungsobjektivität in hohem Maße gefährdet. Ein Verfahren ist auswertungsobjektiv, wenn unter-

schiedliche Personen bei der Kodierung der Daten zu den gleichen Ergebnissen kommen. Bei Fragebögen mit vorgegebenen Antwortkategorien dürfte dies in hohem Maße der Fall sein. Wenn dagegen frei formulierte Antworten kategorisiert werden müssen, ist die Auswertungsobjektivität wesentlich stärker gefährdet. Ein Verfahren ist interpretationsobjektiv, wenn unterschiedliche Personen angesichts der gleichen Ergebnisse zu den gleichen weiterführenden Interpretationen kommen.

Der Begriff der **Praktikabilität** bedarf ebenfalls keiner Übersetzung. Ein Messverfahren ist in dem Maße praktikabel, wie es unter den gegebenen Bedingungen auch angewendet werden kann. Dazu gehört erst einmal, dass die Ausführung möglichst wenig Zeit und Geld kostet. Bei Messverfahren, die bei Menschen angewendet werden, ist außerdem wichtig, dass diese Menschen nicht über Gebühr belastet werden. Mehr noch: Das Messverfahren muss so gestaltet werden, dass die Menschen freiwillig bereit sind, an der Messung teilzunehmen.

Nutzentheoretische Messverfahren

Im Folgenden wird eine spezielle Art von Verfahren betrachtet, die in der Gesundheitsökonomie immer wieder zur Erfassung gesundheitsbezogener Lebensqualität eingesetzt werden: die sog. nutzentheoretischen Messverfahren. Die theoretische Begründung dafür, diese Verfahren in diesem Zusammenhang zu verwenden, ist die oben vorgestellte Definition des Begriffs „gesundheitsbezogene Lebensqualität" (s. S. 160 f.). Gemäß dieser Definition ist die gesundheitsbezogene Lebensqualität eines Menschen dadurch gegeben, wie dieser Mensch seinen aktuellen Gesundheitszustand bewertet. Nutzentheoretische Messverfahren zielen darauf ab, derartige Bewertungen zu erfassen. In der Gesundheitsökonomie werden diese Verfahren benötigt, um zwei verschiedene Arten von Gesundheitszuständen bewerten zu lassen: zum einen aktuelle Gesundheitszustände (also die Gesundheitszustände, in denen sich die Befragten zum Zeitpunkt der Befragung befinden), und zum anderen fiktive Gesundheitszustände (also Beschreibungen bzw. Darstellungen von Gesundheitszuständen, in denen sich die Befragten irgendwann einmal befinden könnten).

In der Literatur werden unterschiedliche Auffassungen darüber vertreten, welche Arten von Verfahren dazu geeignet sein könnten, Bewertungen von Krankheitszuständen zu erfassen. Eine sehr grundlegende Frage ist dabei, ob und – wenn ja – in welcher Weise Bewertungen dadurch erfasst werden können, dass Menschen einen finanziellen Gegenwert für den bewerteten Gegenstand angeben. In der Ökonomie im Allgemeinen und der Gesundheitsökonomie im Besonderen gibt es entsprechend Verfahren, die darauf abzielen, einen solchen finanziellen Gegenwert zu ermitteln: die Verfahren der Zahlungsbereitschaft (willingness to pay) und der Annahmebereitschaft (willingness to accept). Manche Autoren (u.a. Konerding u. Schell 2001; Schöffski et al. 1998) zählen diese Verfahren zu den nutzentheoretischen Verfahren, andere (u.a. Green et al. 2000; Torrance 1986) nicht.

Da die Ergebnisse der Verfahren der Zahlungs- und der Annahmebereitschaft wesentlich von der finanziellen Situation der Befragten abhängen, ist es nicht sinnvoll, diese Verfahren zur Messung gesundheitsbezogener Lebensqualität einzusetzen. Aus diesem Grunde wird auf diese Verfahren im Folgenden nicht weiter eingegangen. Stattdessen werden nur solche Verfahren betrachtet, bei denen nicht nach einem finanziellen Gegenwert gefragt wird. Dabei gibt es drei derartige Verfahren, die in der Gesundheitsökonomie immer wieder verwendet werden:

- das Urteilsskalen-Verfahren (rating-scale-procedure)
- das Standardspiel (standard-gamble)
- das Zeitausgleichs-Verfahren (time-trade-off)

Diese drei Verfahren werden im Folgenden zunächst dargestellt und dann im Vergleich zueinander kritisch gewürdigt.

Darstellung

Das Urteilsskalen-Verfahren

Beim Urteilsskalen-Verfahren werden Bewertungen mithilfe einer Urteilsskala erfasst. Eine solche Urteilsskala kann sehr verschiedene Formen haben. Sie kann in einer Folge nebeneinander geordneter Kästchen, in der Folge der natürlichen Zahlen von 0 bis 100 oder in einer beidseitig begrenzten grafischen Linie bestehen. Die verschiedenen Skalenformen können auch miteinander kombiniert werden. Die letztgenannte Art von Urteilsskala wird meist auch als visuelle Analog-Skala bezeichnet. Üblicherweise werden den beiden Enden einer Urteilsskala bereits in der Darstellung bestimmte Zustände zugewiesen. Bei der Erfassung gesundheitsbezogener Lebensqualität sind dies oft die Zustände „schlechtestmöglicher Gesundheitszustand" und „bestmöglicher Gesundheitszustand". Die Befragten werden gebeten, die zu beurteilenden Krankheitszuständen so auf dieser Skala einzuordnen, dass die Abstände zwischen den Platzierungen der Krankheitszustände den Abständen zwischen den Bewertungen dieser Krankheitszustände entsprechen.

Bei der numerischen Darstellung der Angaben auf der Urteilsskala wird im Allgemeinen von der Annahme ausgegangen, dass die Befragten Abstände zwischen Bewertungen intern repräsentiert haben und die Bewertungen gemäß diesen intern repräsentierten Abständen unverzerrt auf die Urteilsskala projizieren. Entsprechend dieser Annahme werden im Allgemeinen die Werte für zwei Gesundheitszustände willkürlich gesetzt. In den meisten Fällen werden dazu – zumindest bei der ersten numerischen Kodierung – die beiden Gesundheitszustände ausgewählt, die mit den beiden Enden der Urteilsskala verbunden sind. Oft wird dann der Zustand am unteren Ende der Skala mit 0 und der am oberen Ende der Skala mit 100 kodiert. Die numerischen Bewertungswerte für alle anderen Gesundheitszustände werden dann so bestimmt, dass die Abstände zwischen den Zahlen den Abständen auf der Urteilsskala entsprechen. In diesem Sinne werden Urteilsskalen-Werte auch meist als intervallskaliert angesehen.

Wie alle hier vorgestellten Verfahren wird auch das Urteilsskalen-Verfahren in gesundheitsökonomischen Zusammenhängen meistens eingesetzt, um vergleichbare Werte für unterschiedliche Gesundheits- bzw. Krankheitszustände zu gewinnen. Dabei werden chronische und temporäre Krankheitszustände unterschieden. Chronische Krankheitszustände werden üblicherweise als Zustände beschrieben, die ohne Veränderungen vom Zeitpunkt des Krankheitsausbruchs bis zum Tod andauern. Zu einer eindeutigen Definition eines chronischen Krankheitszustandes gehören noch die Angabe des Alters, in dem die Krankheit ausbricht, sowie die Angabe des Alters, in dem der Tod eintritt. Temporäre Krankheitszustände werden üblicherweise als Zustände beschrieben, die vom Zeitpunkt des Krankheitsausbruchs eine Zeit lang ohne Veränderungen andauern und dann in den Zustand völliger Gesundheit übergehen. Zur vollständigen Definition eines temporären Krankheitszustandes gehört somit auch eine genaue Angabe der Länge des Zeitintervalls zwischen Krankheitsausbruch und Gesundung.

Um vergleichbare Werte zu gewinnen, werden die Bewertungen für chronische Krankheitszustände auf die beiden Zustände „völlige Gesundheit" und „Tod" normiert. Die Bewertung für die völlige Gesundheit wird dabei gleich 1 und die für den Tod gleich 0 gesetzt. Die meisten Varianten des Urteilsskalen-Verfahrens liefern bei der ersten numerischen Kodierung keine Werte, die in dieser Form normiert sind. Zumindest für chronische Krankheitszustände lassen sich die ursprünglichen Werte aber durch eine einfache Umrechung in der hier gewünschten Form normieren. Es seien x(i) die ursprünglichen Urteilsskalen-Werte für die Krank-

heitszustände i, x(g) der ursprüngliche Urteilsskalen-Wert für den Zustand völliger Gesundheit und x(t) der ursprüngliche Urteilsskalen-Wert für den Zustand des Todes. Die normierten Skalenwerte y(i) ergeben sich dann durch folgende Gleichung (vgl. Torrance 1986, S. 19):

$$y(i) = [x(i) - x(t)]/[x(g) - x(t)] \quad (1)$$

Da nahezu jeder Mensch den Zustand der völligen Gesundheit als den bestmöglichen erachtet, hat die auf diese Weise normierte Skala mit 1 ein klar definiertes oberes Ende. Da aber nicht alle Menschen den Tod für den schlechtestmöglichen Zustand halten, ist das untere Ende nicht klar definiert. Es kann auf der normierten Skala auch negative Werte geben. Dabei hängt es von den Urteilen der Befragten ab, wie groß diese negativen Werte sind.

Temporäre Krankheitszustände sind etwas schwieriger zu normieren. Solange man nicht daran interessiert ist, temporäre mit chronischen Krankheiten zu vergleichen, kann man sich darauf beschränken, die Bewertungen der temporären Krankheiten auf sich selbst zu normieren. Man identifiziert dazu den Zustand völliger Gesundheit mit 1 und den schlimmstmöglichen temporären Krankheitszustand mit 0. Schwieriger wird es, wenn chronische mit temporären Krankheitszuständen verglichen werden sollen. Drummond et al. (1997) schlagen dazu vor, ergänzend einen chronischen Krankheitszustand mit der gleichen Krankheitsdauer und dem gleichen Krankheitsbild wie dem schlechtestmöglichen temporären Krankheitszustand bewerten zu lassen und den schlechtestmöglichen temporären Krankheitszustand mit diesem Wert gleichzusetzen. Die ursprünglichen Urteilsskalen-Werte seien wieder x(i). Der ursprüngliche Wert für den Zustand völliger Gesundheit sei x(g) und der ursprüngliche Wert für den schlechtestmöglichen temporären Zustand x(s). Der Wert, der dem entsprechenden chronischen Krankheitszustand auf einer bereits auf 0–1 normierten Skala zugewiesen worden ist, sei y(s). Die neue Skala y für die temporären Werte ergibt sich dann durch folgende Gleichung:

$$y(i) = [x(i) - x(i) \times y(s) + x(g) \times y(s) - x(s)]/[x(g) - x(s)] \quad (2)$$

Das Standardspiel

Beim Standardspiel müssen die Befragten zwischen zwei fiktiven Alternativen wählen. Eine dieser beiden Alternativen ist ein sicheres Ereignis. Wenn Gesundheits- bzw. Krankheitszustände bewertet werden sollen, wird meist der zu bewertende Krankheitszustand als sicheres Ereignis gewählt. Die andere Alternative ist dagegen eine Art Glückspiel mit zwei möglichen unsicheren Ausgängen, einem schlechten und einem guten. Wenn Krankheits- bzw. Gesundheitszustände bewertet werden sollen, wird dieses Glückspiel meistens als ein medizinischer Eingriff eingeführt, der entweder zu völliger Gesundheit oder zum schlechtestmöglichen Krankheitszustand führen kann. Es besteht also die Wahl zwischen den folgenden Alternativen:

- Alternative 1: Krankheitszustand i tritt sicher ein.
- Alternative 2: Der Zustand völliger Gesundheit g tritt mit der Wahrscheinlichkeit p ein und der schlimmstmögliche Krankheitszustand s mit der Wahrscheinlichkeit 1 – p.

Um die Bewertung für das sichere Ereignis zu bestimmen, werden die Wahrscheinlichkeiten für die beiden unsicheren Ausgänge bei der zweiten Alternative so lange verändert, bis die Befragten beide Alternativen für gleichwertig halten.

Bei der Übersetzung der Angaben der Befragten in numerische Werte für die Bewertungen von Krankheitszuständen wird davon ausgegangen, dass sich die Befragten gemäß der Theorie des subjektiv erwarteten Nutzens verhalten. Diese Theorie ist ursprünglich von Daniel Bernoulli formuliert worden (Bernoulli 1738). Im 20. Jahrhundert haben dann von Neumann und Morgenstern diese Theorie wieder aufgegriffen und axiomatisch dargestellt (von Neumann u. Morgenstern 1944). Wenn die Befragten die

beiden Alternativen beim Standardspiel für gleichwertig halten, müsste gemäß der Theorie des subjektiv erwarteten Nutzens gelten:

$$x(i) = p(i,g,s) \times x(g) + [1 - p(i,g,s)] \times x(s) \quad (3)$$

Dabei ist p(i,g,s) die Eintretenswahrscheinlichkeit für den Zustand völliger Gesundheit, bei der die befragte Person beide Alternativen für gleichwertig hält. Der Ausdruck x(i) steht wieder für die Bewertung des zu beurteilenden Krankheitszustandes, x(g) für die Bewertung völliger Gesundheit und x(s) für die Bewertung des schlimmstmöglichen Krankheitszustandes.

Setzt man die Bewertung für den Zustand völliger Gesundheit, also x(g) gleich 1, vereinfacht sich Gleichung 3 zu:

$$x(i) = p(i,g,s) + [1 - p(i,g,s)] \times x(s) \quad (4)$$

Wenn x(s) bekannt ist, lässt sich mit dieser Gleichung x(i) bestimmen. Wenn der Tod den schlimmstmöglichen Krankheitszustand bildet und damit x(s) gleich x(t) ist, muss x(s) für die normierte Skala ohnehin gleich 0 gesetzt werden. Die Gleichung 4 vereinfacht sich dann zu folgender:

$$x(i) = p(i,g,s) \quad (5)$$

Die Bewertung für den Krankheitszustand aus Alternative 1 ist also gleich der Eintretenswahrscheinlichkeit für völlige Gesundheit in Alternative 2.

Wenn der Tod nicht mit dem schlimmstmöglichen Krankheitszustand identisch ist, muss erst die Bewertung für den schlimmstmöglichen Krankheitszustand festgestellt werden. Wenn es sich bei dem schlimmstmöglichen Krankheitszustand um einen chronischen Zustand handelt, werden hierzu die folgenden beiden Alternativen vorgegeben:

- Alternative 1: Sofortiger Tod tritt sicher ein.
- Alternative 2: Der Zustand völliger Gesundheit g tritt mit der Wahrscheinlichkeit p ein und der schlimmstmögliche Krankheitszustand s mit Wahrscheinlichkeit 1– p.

Die Wahrscheinlichkeiten für Alternative 2 werden so lange verändert, bis die befragte Person beide Alternativen als gleichwertig empfindet. Die Wahrscheinlichkeit für den Zustand völliger Gesundheit, bei der die befragte Person Alternative 2 für genauso gut hält wie den sicheren Tod, sei p(t,g,s).

Unter Voraussetzung der Theorie des subjektiv erwarteten Nutzens gilt bei Unentschiedenheit zwischen den beiden Alternativen:

$$x(t) = p(t,g,s) \times x(g) + [1 - p(t,g,s)] \times x(s) \quad (6)$$

Setzt man wieder x(g) gleich 1 und x(t) gleich 0, vereinfacht sich die Gleichung zu:

$$0 = p(t,g,s) + [1 - p(t,g,s)] \times x(s) \quad (7)$$

Die Auflösung nach x(s) liefert:

$$x(s) = -p(t,g,s)/[1 - p(t,g,s)] \quad (8)$$

Das Einsetzen in die Gleichung 4 ergibt:

$$x(i) = p(i,g,s) - [1- p(i,g,s)] \times p(t,g,s)/[1 - p(t,g,s)] \quad (9)$$

Dies ist die Gleichung, mit der die Werte für alle anderen Krankheitszustände bestimmt werden können.

Wenn temporäre Krankheitszustände bewertet werden sollen, ist erst der Wert für den chronischen Krankheitszustand zu bestimmen, der dem schlimmsten temporären Krankheitszustand entspricht. Dazu können die Herangehensweisen verwendet werden, die eben beschrieben worden sind. Der normierte Wert für den entsprechenden chronischen Krankheitszustand kann dann in analoger Weise in die Bestimmungsgleichung für die x(i) eingesetzt werden. Dieser normierte Wert sei y(s). Es resultiert dann:

$$x(i) = p(i,g,s) + [1 - p(i,g,s)] \times y(s) \quad (10)$$

Viele Menschen können sich nicht viel unter Wahrscheinlichkeiten vorstellen. Bei der Anwendung des Standardspiels werden die Wahrscheinlichkeitsangaben deshalb oft über ein Glücksrad visualisiert. Auf diesem Glücksrad werden zwei Bereiche unterteilt. Einer entspricht p, der andere 1 – p. Weiter ist bei der Anwendung des Standardspiels zu beachten, dass die Ergebnisse dadurch beeinflusst werden, in welcher Reihenfolge die verschiedenen Varianten der Alternative 2 gebildet werden. Schöffski (1998) schlägt deshalb vor, mit den beiden Extremfällen p = 1 und p = 0 zu beginnen und dann die Wahrscheinlichkeiten alter-

nierend von beiden Seiten in Schritten von 0,1 zu verändern.

Das Zeitausgleichs-Verfahren

Auch beim Zeitausgleichs-Verfahren müssen die Befragten zwischen zwei Alternativen wählen. Diesmal bestehen die Alternativen in Kombinationen aus Krankheitszuständen mit Zeitdauern. Die genaue Beschreibung dieser Alternativen hängt davon ab, ob chronische oder temporäre Krankheitszustände beurteilt werden sollen. Des Weiteren ist bei den chronischen Krankheitszuständen eine spezielle Modifikation notwendig, wenn Krankheitszustände, die schlimmer als der Tod sind, bewertet werden sollen.

Bei chronischen Krankheitszuständen, die nicht schlimmer als der Tod sind, lauten die beiden Alternativen:

- **Alternative 1:** Der Krankheitszustand i dauert r Jahre und wird dann durch den Tod beendet.
- **Alternative 2:** Der Zustand völliger Gesundheit g dauert z Jahre ($z < r$) und wird dann durch den Tod beendet.

Der Zeitwert in Alternative 2 wird so lange verändert, bis die befragte Person beide Alternativen für gleichwertig hält.

Bei der Übersetzung der Angaben der befragten Personen in numerische Wert für Bewertungen wird – meist eher implizit – die Annahme vorausgesetzt, dass sich die Bewertung einer Folge von Gesundheitszuständen als Summe der Produkte aus den Zeitdauern der Zustände mit deren Bewertungen ergibt (vgl. Konerding 2003). Es sei z(i,g,r) die Zeitdauer, bei der die befragte Person beide Alternativen für gleichwertig hält. Dann müsste entsprechend der eben formulierten Annahme folgende Gleichung gelten:

$$x(i) \times r = x(g) \times z(i,g,r) \quad (11)$$

Um zu der normierten Skala zu kommen, wird x(g) wieder gleich 1 gesetzt. Die Gleichung 11 vereinfacht sich dann zu:

$$x(i) \times r = z(i,g,r) \quad (12)$$

Die Auflösung nach x(i) liefert:

$$x(i) = z(i,g,r)/r \quad (13)$$

Die Bewertung für den Krankheitszustand i ergibt sich hier also als Quotient der beiden Zeitdauern.

Wenn chronische Krankheitszustände, die schlimmer als der Tod sind, bewertet werden sollen, müssen die Alternativen völlig anders konstruiert werden. Sie lauten dann:

- **Alternative 1:** Zuerst gibt es eine Zeit im Krankheitszustand i. Diese Zeit dauert r Jahre. Danach folgt eine Zeit in völliger Gesundheit (g). Diese Zeit dauert z Jahre.
- **Alternative 2:** Sofortiger Tod (t).

Diesmal wird der Wert z in Alternative 1 so lange verändert, bis die befragte Person beide Alternativen für gleichwertig hält.

Bei der Bestimmung der Bewertung des Krankheitszustandes i wird auch hier wieder vorausgesetzt, dass sich die Gesamtbewertungen für beide Alternativen als Summen der Produkte aus Zeitdauern und Bewertungen für die Krankheitszustände zusammensetzen. Wenn z(i,t,r) die Zeitdauer ist, bei der die befragte Person beide Alternativen für gleichwertig hält, müsste entsprechend gelten:

$$x(t) = x(g) \times z(i,t,r) + x(i) \times r \quad (14)$$

Für die normierte Skala sind wieder x(t) gleich 0 und x(g) gleich 1 zu setzen. Die Gleichung 14 vereinfacht sich dann zu folgender:

$$0 = z(i,t,r) + x(i) \times r \quad (15)$$

Die Auflösung nach x(i) liefert:

$$x(i) = -z(i,t,r)/r \quad (16)$$

Dieser Wert ist also immer kleiner als 0.

Bei temporären Krankheitszuständen muss erst der schlimmstmögliche temporäre Krankheitszustand bestimmt werden. Alle anderen temporären Krankheitszustände werden dann im Vergleich zu diesem schlimmstmöglichen Zustand beurteilt. Die beiden Alternativen lauten dann:

- **Alternative 1:** Der schlimmstmögliche Krankheitszustand s dauert z Wochen. Danach tritt völlige Gesundheit ein.

- **Alternative 2:** Der Krankheitszustand i dauert r (r > z) Wochen. Danach tritt völlige Gesundheit ein.

Auch hier wird der Wert z in Alternative 1 so lange verändert, bis die befragte Person beide Alternativen für gleichwertig hält.
Die Zeitdauer, bei der die befragte Person beide Alternativen für gleichwertig hält, sei z(i,s,r). Unter den bisher getroffenen Annahmen müsste dann gelten:

$$r \times x(i) = z(i,s,r) \times x(s) + [r - z(i,s,r)] \times x(g) \quad (17)$$

Setzt man x(g) wieder gleich 1, löst nach x(i) auf und formt etwas um, ergibt sich:

$$x(i) = 1 - z(i,s,r) \times [1 - x(s)]/r \quad (18)$$

Will man lediglich temporäre Krankheitszustände untereinander vergleichen, ist es am einfachsten, x(s) gleich 0 zu setzen. Die Bestimmungsformel vereinfacht sich dann zu:

$$x(i) = 1 - z(i,s,r)/r \quad (19)$$

Will man dagegen die temporären Krankheitszustände mit chronischen vergleichen, ist zunächst der Wert für den chronischen Krankheitszustand zu bestimmen, der dem schlimmstmöglichen temporären Krankheitszustand entspricht. Hierzu können die bisher beschriebenen Vorgehensweisen verwendet werden. Der resultierende Wert kann dann in die Bestimmungsgleichungen eingesetzt werden.

Kritische Würdigung

Jedes der drei hier vorgestellten Verfahren beruht auf einer etwas anderen Form menschlicher Urteile. Die Prinzipien, nach denen diese Urteile in numerische Werte für Bewertungen übersetzt werden, beruhen auf bestimmten Annahmen darüber, wie Menschen diese Urteile treffen. Bei der Interpretation der dabei gewonnenen Werte stellen sich damit zwei Fragen:

- Inwieweit sind die Annahmen, auf denen die numerische Kodierung dieser Verfahren beruht, empirisch gültig?
- Inwieweit spiegeln diese drei verschiedenen Verfahren dieselbe Größe wider?

Beiden Fragen wird jetzt nacheinander nachgegangen.

Die Gültigkeit der vorausgesetzten Annahmen

In der Form, in der das Urteilsskalen-Verfahren üblicherweise praktiziert wird, beruht es auf zwei oft nur implizit vorausgesetzten Annahmen:

- Die Befragten können bezüglich der Größe, nach der sie gefragt werden, konsistent Abstände einschätzen.
- Die Befragten bilden die Größe, nach der sie gefragt werden, entsprechend den von ihnen wahrgenommenen Abständen unverzerrt auf der Urteilsskala ab.

Die erste dieser beiden Annahmen ist für Bewertungen von Krankheitszuständen bisher noch nie ernsthaft kritisch geprüft worden. Es gibt allerdings einige wenige Untersuchungen, in denen diese Annahme bezüglich anderer Bewertungen geprüft worden ist. Diese Untersuchungen stammen alle aus den 80er Jahren (Konerding 1989; Orth 1982; Orth u. Wegener 1983; Westermann 1984; 1985; Westermann u. Hager 1983; 1985). Als Ergebnis zeigte sich vor allem, dass große individuelle Unterschiede bestehen. Manche Personen scheinen konsistent Abstände zwischen Bewertungen beurteilen zu können, andere nicht.
Die zweite Annahme lässt sich vollständig nur prüfen, wenn man weiß, welche Abstände die Befragten tatsächlich intern repräsentiert haben. Das ist im Allgemeinen nicht der Fall. Es gibt aber Folgerungen aus der zweiten Annahme, die ohne Kenntnis der wahren internen Urteile empirisch geprüft werden können. So müssten unter Gültigkeit der zweiten Annahme Urteilsskalen-Werte, die sich auf die gleichen Gegenstände beziehen, aber unter verschiedenen Randbedingungen erhoben worden sind, in einer linearen Beziehung miteinander stehen. Empirisch zeigt sich aber, dass sich die Urteilsskalen-Werte in nichtlinearer Weise in Abhängigkeit davon verändern, welche Gegenstände

zusammen beurteilt werden (Parducci 1965; Parducci u. Weddell 1986; Robinson et al. 2001). Wenn beispielsweise mehrere Krankheitszustände mit mittleren Werte zusammen mit nur wenigen Krankheitszuständen mit extremen Werten beurteilt werden, werden die Abstände zwischen den Krankheitszuständen in der Mitte bei der Abbildung auf die Urteilsskala vergrößert und die Abstände zu den extremen Krankheitszuständen verkleinert.

Das Standardspiel beruht auf der Annahme, dass sich Menschen bei Entscheidungen gemäß dem Modell des subjektiv erwarteten Nutzens verhalten. Im Zusammenhang mit Entscheidungen zwischen Krankheitszuständen ist diese Annahme bisher kaum kritisch geprüft worden, wohl aber sehr ausführlich im Zusammenhang mit Wahlen zwischen monetären Lotterien. Hier zeigt sich u. a., dass Entscheidungen beim Standardspiel wesentlich davon abhängen, wie die Frage formuliert ist. Bei inhaltlich gleichen, aber unterschiedlich formulierten Alternativenpaaren entscheiden sich Menschen völlig unterschiedlich (Hershey et al. 1988). Des Weiteren verhalten sich Menschen bei Entscheidungen zwischen Lotterien, bei denen die möglichen Ausgänge mit sehr kleinen oder sehr großen Wahrscheinlichkeiten verbunden sind, kaum gemäß dem Modell des subjektiv erwarteten Nutzens. Wenn die möglichen Ausgänge mit mittleren Wahrscheinlichkeiten verbunden sind, ist dies dagegen eher der Fall. Menschen scheinen also große Schwierigkeiten beim rationalen Umgang mit extrem kleinen und extrem großen Wahrscheinlichkeiten zu haben (Kahneman u. Tversky 1979; 1982). Andererseits kam Hertwig (1997) nach einer umfassenden Literaturdurchsicht zu dem Schluss, dass das menschliche Entscheidungsverhalten im Großen und Ganzen schon dem Modell des subjektiv erwarteten Nutzens entspricht.

Das Zeitausgleichs-Verfahren beruht auf der Annahme, dass sich die Gesamtbewertung einer Folge von Gesundheitszuständen als Summe der Produkte aus den Bewertungen der einzelnen Gesundheitszustände mit den dazugehörigen Zeitdauern ergibt. Auch diese Annahme ist bisher nur wenig empirisch geprüft worden. Es gibt hier allerdings eine wichtige Untersuchung von Dolan (1996). In dieser Untersuchung wurden die gleichen Gesundheitszustände einmal mit der Dauer von einem Monat, das zweite Mal mit der Dauer von einem Jahr und ein drittes Mal mit der Dauer von zehn Jahren mithilfe des Urteilsskalen-Verfahrens bewertet. Unter Voraussetzung der hier vorgestellten Annahmen für das Urteilsskalen- und das Zeitausgleichs-Verfahren müssten die Verhältnisse zwischen den Abständen der Bewertungen der verschiedenen Zeitdauern innerhalb der drei Zeitdauergruppen gleich sein. Dolan fand aber, dass bei längerer Dauer die schwereren Krankheitszustände im Vergleich zu den leichteren Krankheitszuständen schlechter bewertet werden als bei kürzerer Dauer. Darüber hinaus zeigt sich in anderen empirischen Untersuchungen, dass die Bevorzugungen zwischen Folgen von Zuständen von der Reihenfolge und der zeitlichen Distanz der Zustände abzuhängen (vgl. Jungermann et al. 1998). Auch dies steht im Widerspruch zu der Annahme, auf der das Zeitausgleichs-Verfahren beruht.

Die Beziehungen zwischen den drei Verfahren

Die drei hier betrachteten nutzentheoretischen Verfahren beziehen sich auf unterschiedliche Formen menschlicher Urteile und beruhen außerdem auf unterschiedlichen theoretischen Konzeptionen. Der wichtigste Unterschied besteht in dem erkenntnistheoretischen Prinzip, nach dem der Bewertungsbegriff festgelegt wird. Beim Urteilsskalen-Verfahren geschieht dies durch Rückgriff auf das Sprachverständnis der Befragten. Sie werden direkt nach Bewertungen gefragt. Bewertung ist damit das, was die Befragten unter der Bezeichnung „Bewertung" verstehen. Was *das* aber ist, bleibt bis zu einem gewissen Grade offen. In diesem Sinne ist auch offen, was die mit dem Urteilsskalen-Verfahren gewonnenen Werte inhaltlich bedeuten. Beim Standardspiel und beim Zeitaus-

gleichs-Verfahren wird dagegen überhaupt nicht direkt nach Bewertungen gefragt, sondern lediglich nach Bevorzugungen zwischen zwei Alternativen. Der Begriff der Bewertung wird indirekt durch die Modellannahmen festgelegt, nach denen diese Bevorzugungen interpretiert werden. Da die Modellannahmen klar beschrieben werden können, ist damit auch der dazugehörige Bewertungsbegriff klar festgelegt.

Der Unterschied zwischen dem Urteilsskalen-Verfahren auf der einen Seite und dem Standardspiel und dem Zeitausgleichs-Verfahren auf der anderen ist auch aus ökonomischer Sicht mehrfach problematisiert worden (s. u. a. Green et al. 2000). Aus dieser Sicht äußert sich der Wert eines Gutes vor allem darin, inwieweit man bereit ist, für dieses Gut etwas einzutauschen. Das Standardspiel und das Zeitausgleichs-Verfahrens beruhen darauf, dass die Befragten eine Wahlmöglichkeit für eine andere Wahlmöglichkeit eintauschen. Beim Urteilsskalen-Verfahren findet ein solcher Tausch nicht statt. Damit wäre eher vom Standardspiel und vom Zeitausgleichs-Verfahren zu erwarten, dass sie Bewertungen in dem Sinne erfassen, wie sie in der Ökonomie verstanden werden.

Im Vergleich zum Urteilsskalen-Verfahren haben das Standardspiel und das Zeitausgleichs-Verfahren zwar einige wichtige Gemeinsamkeiten, es bestehen aber auch Unterschiede. Beim Standardspiel geht es um Entscheidungen zwischen Wahlmöglichkeiten mit unsicheren Ausgängen, beim Zeitausgleichs-Verfahren um Wahlmöglichkeiten zwischen zeitlichen Folgen sicherer Ereignisse. Das erstgenannte Verfahren beruht also auf Entscheidungen unter Unsicherheit, das zweite auf Entscheidungen unter Sicherheit. Manche Ökonomen (u. a. Drummond et al. 1997; Gold et al. 1996; Mehrez u. Gafni 1991) weisen diesem Unterschied eine große Bedeutung zu. Aus ihrer Sicht bilden Entscheidungen unter Unsicherheit deshalb eine validere Grundlage für die Messung von Bewertungen, weil unsere ganze Welt voller Unsicherheit sei. Bewertungen, die über Entscheidungen unter Unsicherheit bestimmt worden sind, werden dort oft auch als „utilities" bezeichnet; Bewertungen, die auf Urteilen unter Sicherheit beruhen, dagegen als „values".

Die Beziehungen zwischen den drei hier betrachteten nutzentheoretischen Verfahren sind auch schon von verschiedenen Perspektiven her empirisch betrachtet worden. Green et al. (2000) fanden bei einer sehr umfassenden Literaturanalyse, dass Standardspiel- und Zeitausgleichs-Werte meistens recht hoch miteinander korrelieren, während die Korrelationen mit Urteilsskalen-Werten vergleichsweise geringer sind. Aufgrund dieser Ergebnisse schlossen die Autoren, dass Urteilsskalen-Werte einen anderen Aspekt gesundheitsbezogener Lebensqualität widerspiegeln als Werte, die mit dem Standardspiel oder dem Zeitausgleichs-Verfahren bestimmt worden sind. Dies entspräche auch den hier herausgearbeiteten theoretischen Unterschieden zwischen dem Urteilsskalen-Verfahren auf der einen und dem Standardspiel und dem Zeitausgleichs-Verfahren auf der anderen Seite.

Verschiedene Autoren haben versucht, eine Funktion zu finden, mit der sich Urteilsskalen-Werte in Standardspiel-Werte übersetzen lassen. Ein kurzer Überblick hierzu findet sich bei Torrance et al. (2001). Gemäß diesem Überblick haben sich zwei Funktionen zwischen den Urteilsskalen-Werten x und den Standardspiel-Werten y empirisch besonders bewährt. Bei beiden Funktionen wird vorausgesetzt, dass die Werte auf 0 für den Tod und 1 für völlige Gesundheit normiert sind. Die erste Funktion lautet:

$$y = 1 - (1 - x)^b \text{ mit } 1{,}6 \leq b \leq 2{,}9 \qquad (20)$$

Die zweite Funktion lautet:

$$y = x^b \text{ mit } 0{,}46 \leq b \leq 0{,}56 \qquad (21)$$

Beide Funktionen sind konkav. An den Punkten 0 und 1 sind Urteilsskalen- und Standardspiel-Werte identisch. Dazwischen sind die Standardspiel-Werte höher als die zum gleichen Gesundheitszustand gehörenden Urteilsskalen-Werte.

Cook et al. (2001) haben Urteilsskalen-Werte, Standardspiel-Werte und Zeitausgleichs-Ver-

fahrens-Werte in einer gemeinsamen Analyse betrachtet. Die Ergebnisse dieser Analyse sprechen sehr dagegen, dass auch nur zwei dieser drei Skalen in systematischer Weise zusammenhängen. Des Weiteren schließen die Autoren aufgrund ihrer Ergebnisse, dass keine dieser Skalen gleichabständig Bewertungen widerspiegelt. Da die Autoren bei ihrer Analyse aber auf keine Bewertungswerte zurückgreifen konnten, die unabhängig von den drei betrachteten Verfahren bestimmt worden sind, ist diese Schlussfolgerung eher fragwürdig.

Psychometrische Messverfahren

Psychometrische Verfahren zur Erfassung von Lebensqualität sind Fragebögen, die auf solche Aspekte des Gesundheitszustandes abzielen, von denen man meint, dass sie für die Lebensqualität von Bedeutung sind. Wie weiter oben ausgeführt, unterscheidet man generische und krankheitsspezifische Verfahren. Ebenso unterscheidet man zwischen Index- und Profilverfahren. Im Prinzip sind beide Unterscheidungen unabhängig voneinander. Im Folgenden wird ein kurzer Überblick über das derzeit vorliegende Angebot an psychometrischen Verfahren gegeben. Außerdem werden einige Verfahren beispielhaft dargestellt. Dabei werden als Erstes die generischen und als Zweites die krankheitsspezifischen Fragebögen diskutiert.

Generische Lebensqualitätsfragebögen

Es lässt sich nicht genau sagen, wie viele generische Lebensqualitätsfragebögen zurzeit weltweit zur Verfügung stehen. Im Jahr 2000 waren es ungefähr 25 (Konerding u. Schell 2001). Seitdem dürften es kaum weniger geworden sein. Allerdings haben nicht alle die gleiche Beachtung gefunden. Die generischen Verfahren, die am häufigsten verwendet werden, sind folgende (vgl. Coons et al. 2000):

- die 36-Item Short-Form des Medical Outcomes Study Health Survey (SF-36)
- das Nottingham Health Profile (NHP)
- das Sickness Impact Profile (SIP)
- die Dartmouth Primary Care Cooperative Information Project Charts (COOP)
- die Quality of Well-Being Scale (QWB)
- der Health Utilities Index (HUI)
- der European Quality of Life Questionnaire (EQ-5D)

Die vier erstgenannten Fragebögen sind reine Profilinstrumente, die drei letztgenannten Indexinstrumente. Dabei ist der SF-36 wohl das am meisten verwendete generische Profilinstrument und der EQ-5D das am meisten verwendete generische Indexinstrument. Für beide Verfahren gibt es auch deutschsprachige Versionen. Diese werden im Folgenden nacheinander dargestellt.

Der SF-36 als Beispiel für ein generisches Profilinstrument

Der SF-36 (Short Form 36) ist mit Unterstützung des RAND-Instituts, einer amerikanischen Organisation für Forschung und Entwicklung (Research and Development), als standardisiertes Instrument zur Erfassung gesundheitsbezogener Lebensqualität entwickelt worden. Die zurzeit gebräuchliche Form dieses Fragebogens wurde 1992 von Ware und Sherbourne vorgestellt (Ware u. Sherbourne 1992). Sie liegt mittlerweile in verschiedenen Sprachen vor und umfasst 36 Fragen (Items). Aus diesem Grund heißt dieser Fragebogen auch SF-36. Von den 36 Fragen werden 35 acht verschiedenen Dimensionen (Subskalen) zugeordnet. Diese acht Dimensionen sind:

- körperliche Funktionsfähigkeit (10 Items)
- körperliche Rollenfunktion (4 Items)
- emotionale Rollenfunktion (3 Items)
- soziale Funktionsfähigkeit (2 Items)
- psychisches Wohlbefinden (5 Items)
- körperliche Schmerzen (2 Items)
- Vitalität (4 Items)

- allgemeine Gesundheitswahrnehmung (5 Items)

Die letzte Frage bezieht sich auf den Vergleich des aktuellen Gesundheitszustandes mit dem vor einem Jahr. Bei einem Teil der Fragen des SF-36 werden lediglich die beiden Antwortmöglichkeiten „ja" und „nein" vorgegeben, bei dem anderen Teil sechsstufige Antwortskalen. Der Fragebogen kann sowohl zur Selbst- als auch zur Fremdbeurteilung eingesetzt werden. Des Weiteren eignet er sich sowohl zum Selbstausfüllen als auch zur Erfassung der Lebensqualität über persönliche oder telefonische Interviews durch entsprechend trainierte Interviewer. Die Bearbeitungszeit beträgt ca. zehn Minuten.

Zur Auswertung des Fragebogens werden die Antworten auf die Fragen numerisch so kodiert, dass die Abstände zwischen den Zahlen möglichst den Abständen zwischen den Antwortmöglichkeiten entsprechen. Die Kodierungsregel findet man u.a. bei Bullinger (1995). Im zweiten Schritt werden die Werte für jede Dimension addiert und linear auf Werte zwischen 0 und 100 transformiert (s. Ware 2002). Es wird hier also vorausgesetzt, dass die Fragen, die derselben Dimension zugeordnet werden, im gleichen Maß auf genau diese Dimension abzielen. Darüber hinaus gibt es die Möglichkeit, einen psychischen und einen körperlichen Summenwert zu errechnen: Für diese Berechnung werden alle Items unabhängig von den acht ursprünglichen Skalen, aber für den psychischen und körperlichen Summenwert getrennt, gewichtet und addiert. Die beiden resultierenden Summenwerte werden dann wieder in Skalenwerte von 0 bis 100 transformiert. Hays et al. (1993) diskutierten noch einige alternative Herangehensweisen, die Antworten zum SF-36 numerisch zu kodieren.

Mehrere Untersuchungen zur Güte der deutschsprachigen Version des SF-36 liegen vor. Wichtig ist hier vor allem eine an 2914 Personen durchgeführte, bevölkerungsrepräsentative Studie (vgl. Bullinger 1998). Die Reliabilität der acht Teilskalen wurde dort über interne Konsistenz geprüft. Abgesehen von den Teilskalen „soziale Funktionsfähigkeit" und „allgemeine Gesundheitswahrnehmung" weisen bei dieser Untersuchung alle Teilskalen hinreichend hohe Werte auf. Bullinger (1995) präsentierte auch eine Studie zur Validität der deutschen Version des SF-36. Als Ergebnis dieser Studie zeigte sich, dass die Teilskalen des SF-36 hoch mit den inhaltlich entsprechenden Teilskalen des Nottingham Health Profile (NHP) korrelieren. Das Nottingham Health Profile ist, wie bereits gesagt, ein anderes Profilinstrument zur Messung gesundheitsbezogener Lebensqualität. Weiter zeigte sich als Ergebnis, dass der SF-36 sehr gut zwischen Personen mit stark ausgeprägten und Personen mit gering ausgeprägten Beschwerden unterscheiden kann. Bullinger wertete dies als Beleg für die Validität des Verfahrens. Ware (2002) berichtete von einer Vielzahl von Studien, in denen anderssprachige Versionen des SF-36 nach ähnlichen Prinzipien und mit ähnlichen Ergebnissen untersucht worden sind.

Der EQ-5D als Beispiel für ein generisches Indexinstrument

Der EQ-5D ist von einer 1987 gegründeten Forschergruppe mit Mitgliedern aus verschiedenen Fachdisziplinen und verschiedenen europäischen Ländern entwickelt worden. Diese Gruppe bezeichnet sich selbst als die EuroQol-Gruppe. Die Zielsetzung dieser Gruppe bestand darin, ein einfach anzuwendendes Instrument zur standardisierten Beschreibung und Bewertung unterschiedlichster Krankheitszustände bereitzustellen. Die erste Version dieses Fragebogens wurde im Jahre 1990 veröffentlicht (EuroQol Group 1990). Im Oktober 1991 wurde der Fragebogen weiter überarbeitet (Brooks et al. 1996). Diese jetzt immer noch aktuelle Version wird von der EuroQol-Gruppe selbst als EQ-5D (EuroQol-5 Dimensionen) bezeichnet. Von ihr gibt es mittlerweile von der EuroQol-Gruppe offiziell anerkannte Übersetzungen in Afrikaans, Bulgarisch, Katalanisch,

Kroatisch, Tschechisch, Dänisch, Niederländisch, Englisch, Finnisch, Französisch, Deutsch, Griechisch, Ungarisch, Italienisch, Japanisch, Norwegisch, Polnisch, Portugiesisch, Spanisch, Schwedisch und Türkisch (EuroQol-Group 2002). Vorausgesetzt, es handelt sich nicht um Forschung mit kommerzieller Zielsetzung, z. B. durch die pharmazeutische Industrie, kann der EQ-5D ohne Entgelt benutzt werden. Allerdings erwartet die EuroQol-Gruppe, dass eine offizielle Version angewendet wird und der Benutzer sein Projekt registrieren lässt.

Der EQ-5D ist in vier Teile gegliedert: Der erste Teil umfasst fünf Fragen zu verschiedenen Dimensionen der Lebensqualität. Im Einzelnen handelt es sich dabei um folgende Dimensionen:

- Mobilität
- Körperpflege
- allgemeine Tätigkeiten
- Schmerzen/körperliche Beschwerden
- Ängstlichkeit/Niedergeschlagenheit

Jede dieser fünf Dimensionen umfasst drei mögliche Stufen:

- Stufe 1: keine Probleme
- Stufe 2: einige Probleme
- Stufe 3: extreme Probleme

Es gibt somit 243 mögliche Gesundheitszustände, die durch den EQ-5D unterschieden werden können. Außerdem werden die Probanden bzw. Patienten im ersten Teil um einen Vergleich ihres aktuellen Gesundheitszustandes mit dem während der vergangenen zwölf Monate gebeten. Der zweite Teil beinhaltet eine Beurteilungsskala von 0 (schlechtestdenkbarer Gesundheitszustand) bis 100 (bestdenkbarer Gesundheitszustand) zur Selbsteinschätzung des aktuellen Gesundheitszustandes. Der dritte Teil dient zur Ermittlung von Daten, die zur Begleitforschung zum EQ-5D herangezogen werden können. Der vierte Teil beinhaltet Fragen zu Hintergrunddaten des jeweiligen Patienten, z. B. Alter, Bildung usw. Lediglich die beiden erstgenannten Teile werden in klinischen und gesundheitsökonomischen Studien eingesetzt.

Die fünf Dimensionen des EQ-5D liefern zunächst einmal ein Profil, mit dem verschiedene Krankheitszustände beschrieben werden können. Dieses Profil ist aber von vornherein als Grundlage für die Bestimmung eines eindimensionalen Indexes der Lebensqualität gedacht. Damit solch ein eindimensionaler Index mithilfe des EQ-5D bestimmt werden kann, wird außerdem noch eine Regel benötigt, nach der jedem Profil ein Bewertungswert zugeordnet wird. Da die Menschen in verschiedenen Ländern Krankheitszustände möglicherweise unterschiedlich bewerten, wird letztlich für jede sprach- und kulturspezifische Version eine eigene Übersetzungsregel benötigt. Ein wesentlicher Teil der Forschung zum EQ-5D zielt deshalb darauf ab, derartige Regeln zu bestimmen. Dabei wird üblicherweise ein Verfahren verwendet, das aus zwei Teilschritten besteht. Im ersten Teilschritt wird ein Teil der 243 Profile, die mit dem EQ-5D unterschieden werden können, von einer möglichst repräsentativen Stichprobe bewertet. Dabei werden üblicherweise eines oder mehrere der oben beschriebenen nutzentheoretischen Verfahren verwendet. Im zweiten Schritt wird auf der Grundlage der auf diese Weise empirisch bestimmten Werte mithilfe eines mathematischen Modells die Übersetzungsregel für alle 243 Profile bestimmt.

Auf die eben beschriebene Weise sind mittlerweile für verschiedene sprachspezifische Versionen des EQ-5D Übersetzungsregeln bestimmt worden. Eine Übersetzungsregel für die deutsche Version stammt von Claes et al. (2002). Zur Bestimmung dieser Regel haben diese Autoren zunächst eine für Norddeutschland repräsentative Stichprobe von 4000 Personen ausgewählt. Diese Personen wurden schriftlich um ihre Mitarbeit gebeten. Daraufhin erklärten sich 380 Personen bereit, an der Untersuchung teilzunehmen. Von diesen Personen wurden 339 im Zeitraum von Oktober 1997 bis März

1998 bei sich zu Hause von 18 trainierten studentischen Interviewern befragt. Als Erhebungsmethode wurde das Zeitausgleichs-Verfahren verwendet. Insgesamt wurden 36 verschiedene EQ-5D-Profile untersucht. Jeder einzelnen Person wurden aber nur maximal 15 Profile vorgegeben.

Zur Bestimmung der Regel, nach der den einzelnen Profilen Lebensqualitätsindizes zugeordnet werden, verwendeten Claes et al. zunächst folgendes Modell:

$$Y = \alpha + \beta_1 D1_1 + \beta_2 D1_2 + \beta_3 D1_3 + \beta_4 D1_4 + \beta_5 D1_5 + \beta_6 D2_1 + \beta_7 D2_2 + \beta_8 D2_3 + \beta_9 D2_4 + \beta_{10} D2_5 + \beta_{11} N3 \quad (22)$$

Dabei ist Y der zu bestimmende Lebensqualitätsindex, α und β_i sind Parameter, die aus den Daten geschätzt werden müssen, und $D1_i$, $D2_i$ und N3 sind Dummy-Variablen, für die Folgendes gilt:

- $D1_i = 0$, wenn Dimension i auf Stufe 1 (keine Probleme)
- $D1_i = 1$, wenn Dimension i auf Stufe 2 (einige Probleme)
- $D1_i = 2$, wenn Dimension i auf Stufe 3 (extreme Probleme)
- $D2_i = 1$, wenn Dimension i auf Stufe 3 (extreme Probleme)
- $D2_i = 0$, wenn Dimension i nicht auf Stufe 3 liegt
- N3 = 1, wenn mindestens eine Dimension auf Stufe 3 liegt
- N3 = 0, wenn keine Dimension auf Stufe 3 liegt

Für die Dimensionsindizes gilt dabei:

- i = 1: Mobilität
- i = 2: Körperpflege
- i = 3: allgemeine Tätigkeiten

Tab. 7.2-1 Parameter zur Bestimmung der Lebensqualitätsindizes

Parameter	Variable	erstes Modell		zweites Modell	
α		0,999	$p < 0,01$	0,999	$p < 0,01$
β_1	$D1_1$	−0,100	$p < 0,01$	−0,099	$p < 0,01$
β_2	$D1_2$	−0,067	$p < 0,01$	−0,087	$p < 0,01$
β_3	$D1_3$	−0,014	n. s.		
β_4	$D1_4$	−0,114	$p < 0,01$	−0,112	$p < 0,01$
β_5	$D1_5$	−0,006	n. s.		
β_6	$D2_1$	−0,130	$p < 0,01$	−0,129	$p < 0,01$
β_7	$D2_2$	−0,040	n. s.		
β_8	$D2_3$	0,038	n. s.		
β_9	$D2_4$	−0,084	$p < 0,01$	−0,091	$p < 0,01$
β_{10}	$D2_5$	−0,060	$p < 0,01$	−0,065	$p < 0,01$
β_{11}	N3	−0,318	$p < 0,01$	−0,323	$p < 0,01$

n.s. = nicht signifikant

- i = 4: Schmerzen/körperliche Beschwerden
- i = 5: Ängstlichkeit/Niedergeschlagenheit

Claes et al. (2002) stellten fest, dass sich bei Anwendung dieses Modells auf ihre Daten die Parameter β_3, β_5, β_7 und β_8 nicht signifikant von 0 unterschieden (s. Tab. 7.2-1). Daraufhin verwendeten sie folgendes Modell:

$$Y = \alpha + \beta_1 D1_1 + \beta_2 D1_2 + \beta_4 D1_4 + \beta_6 D2_1 + \beta_9 D2_4 + \beta_{10} D2_5 + \beta_{11} N3 \quad (23)$$

Die Parameter und Variablen in diesem Modell bedeuten dasselbe wie im erstgenannten Modell. Bei Anwendung dieses Modells auf die Daten unterschieden sich alle Parameter signifikant von 0 (s. Tab. 7.2-1). Die Autoren schlugen deshalb vor, die Lebensqualitätsindizes für die einzelnen EQ-5D-Profile mit dem zweiten Modell und den dazugehörigen, empirisch gewonnenen Parametern zu bestimmen.

Die Indizes für die verschiedenen Profile des EQ-5D ergeben sich gemäß dem Ansatz von Claes et al. (2002) also dadurch, dass die zu den Profilen gehörigen, in der Tabelle 7.2-1 aufgeführten Parameterwerte zum zweiten Modell in die Formel 23 eingesetzt werden. Man betrachte hierzu das Profil 21111, also einige Probleme bei der Dimension „Beweglichkeit/Mobilität" und keine Probleme bei den anderen vier Dimensionen. Bei diesem Profil ist $D1_1$ gleich 1, während alle anderen $D1_i$, alle $D2_i$ und N3 gleich 0 sind. Damit gilt:

$$\begin{aligned} Y(21111) &= \alpha + 1 \times \beta_1 + 0 \times \beta_2 + 0 \times \beta_4 + 0 \times \beta_6 + 0 \times \beta_9 + 0 \times \beta_{10} + 0 \times \beta_{11} \\ &= 0{,}999 - 1 \times 0{,}099 + 0 + 0 + 0 + 0 + 0 + 0 \\ &= 0{,}999 - 0{,}099 = 0{,}900 \end{aligned} \quad (24)$$

Beim Profil 23232 sind dagegen $D1_1$ gleich 1, $D1_2$ und $D1_4$ gleich 2, $D2_1$ und $D2_5$ gleich 0, $D2_4$ gleich 1 und N3 gleich 1. Damit gilt:

$$\begin{aligned} Y(23232) &= \alpha + 1 \times \beta_1 + 2 \times \beta_2 + 2 \times \beta_4 + 0 \times \beta_6 + 1 \times \beta_9 + 0 \times \beta_{10} + 1 \times \beta_{11} \\ &= 0{,}999 - 1 \times 0{,}099 - 2 \times 0{,}087 - 2 \times 0{,}112 + 0 - 1 \times 0{,}091 + 0 - 1 \times 0{,}323 \\ &= 0{,}999 - 0{,}099 - 0{,}174 - 0{,}224 - 0{,}091 - 0{,}323 = 0{,}088 \end{aligned} \quad (25)$$

Krankheitsspezifische Lebensqualitätsfragebögen

Das Angebot an krankheitsspezifischen Fragebögen zur Messung von Lebensqualität ist äußerst umfassend und damit kaum übersehbar. Salek (1999) führte insgesamt 200 verschiedene Instrumente zur Messung gesundheitsbezogener Lebensqualität auf. Der weitaus größte Teil dieser Instrumente ist krankheitsspezifisch. Im Jahre 2000 befanden sich in einer heute leider nicht mehr existierenden Internetpräsentation der Abteilung für Psychologie des Nationalen Instituts für Tumorforschung in Mailand Verweise auf etwa 800 verschiedene krankheitsspezifische Fragebögen zur Messung von Lebensqualität. Die Anzahl krankheitsspezifischer Fragebögen dürfte zwischenzeitlich kaum geringer geworden sein.

Obwohl es derartig viele krankheitsspezifische Fragebögen gibt, findet sich unter ihnen zurzeit kein Indexinstrument in dem hier verstandenen Sinne. Bei einigen dieser Fragebögen wird zwar die Bezeichnung „Index" im offiziellen Namen des Fragebogens verwendet, wie etwa beim Psoriasis Disability Index (Finlay u. Coles 1995), der zur Erfassung der Lebensqualität bei Schuppenflechte verwendet wird. Der Wert für diesen Index wird aber lediglich dadurch bestimmt, dass Fragen zu verschiedenen krankheitsrelevanten Symptomen, wie etwa „Sichtbarkeit der Schuppenflechte im Gesicht", gestellt und die Antworten zu den verschiedenen Fragen zu einem Wert zusammengefasst werden. Anders als beim EQ-5D beruht dieser zusammenfassende Wert nicht darauf, dass verschiedene Konstellationen von Symptomen explizit bewertet worden sind. Aus diesem Grund kann hier kein krankheitsspezifisches Indexinstrument vorgestellt werden.

Die Auswahl an krankheitsspezifischen Profilinstrumenten ist dafür erheblich. Als Beispiel für ein solches Instrument wird hier der Seattle Angina Questionnaire (SAQ) betrachtet. Dieser Fragebogen wurde von Spertus et al. (1995) zur Ermittlung der gesundheitsbezogenen Le-

bensqualität von Patienten entwickelt, die an einer koronaren Herzerkrankung erkrankt sind. Die koronare Herzerkrankung ist eine Manifestation der Arteriosklerose an den Koronararterien, d. h. den Blutgefäßen, die die Herzmuskulatur versorgen. Signifikante Einengungen (Stenosen) dieser Koronararterien führen zu einer Minderversorgung des Herzmuskels mit Blut und somit zu einer unzureichenden Sauerstoffversorgung. Diese wiederum kann Angina-pectoris-Anfälle hervorrufen, welche sich meist durch starke Schmerzen im Brustkorb bemerkbar machen, die ihrerseits häufig durch körperliche oder psychische Belastungen ausgelöst werden. Kommt es zu einem kompletten Verschluss eines Gefäßes, so resultiert daraus ein Herzinfarkt.

Der SAQ umfasst insgesamt 19 Items. Sie beziehen sich auf fünf Dimensionen der koronaren Herzerkrankung:

- körperliche Einschränkung (9 Items)
- Stabilität der Angina pectoris (1 Item)
- Frequenz der Angina pectoris (2 Items)
- Zufriedenheit mit der Behandlung (4 Items)
- Krankheitswahrnehmung (3 Items)

Der Patient hat jeweils fünf oder sechs vorgegebene Antwortmöglichkeiten zum Ankreuzen. Bei der Auswertung werden benachbarte Antwortmöglichkeiten durch gleichabständige Zahlen kodiert. Für jede der fünf Dimensionen werden diese Zahlen addiert und dann auf den Bereich zwischen 0 und 100 kodiert. Die Zahl 0 steht dabei für den schlechtestmöglichen und die Zahl 100 für den bestmöglichen Wert.

Spertus et al. (1995) evaluierten den SAQ in mehreren Studien hinsichtlich seiner Validität und Reliabilität. Sie prüften die Validität, indem sie die fünf Skalen des Fragebogens mit der Dauer von Laufbandübungen, Arztdiagnosen, Nitroglyzerinnachfüllungen und anderen validierten Instrumenten verglichen. Dabei ergab sich, dass alle fünf Skalen hoch mit den anderen Werten korrelieren. Weiter konnten Spertus et al. zeigen, dass sich die Ergebnisse des SAQ nach erfolgreichen medizinischen Eingriffen in der zu erwartenden Richtung verändern. Die Reliabilität des Fragebogens prüften die Autoren, indem sie den SAQ bei Patienten mit stabiler Angina pectoris in einem Abstand von drei Monaten zweimal anwendeten. Bei allen fünf Teilskalen unterschieden sich die Mittelwerte für beide Zeitpunkte nicht nennenswert. Außerdem korrelierten die Werte für die beiden Zeitpunkte bei vier der fünf Teilskalen recht hoch miteinander. Bei der Teilskala zur Stabilität war die Korrelation verhältnismäßig gering. Dies dürfte darauf zurückzuführen sein, dass diese Teilskala nur aus einem Item besteht. Insgesamt sprechen die Ergebnisse aber sowohl für die Validität als auch die Reliabilität des SAQ.

Zusammenfassende Diskussion

In diesem Kapitel sind die wichtigsten Ansätze vorgestellt, die zurzeit in der Gesundheitsökonomie zur Erfassung von Lebensqualität verwendet werden. Zwei sehr verschiedene Arten von Verfahren sind dabei betrachtet worden: nutzentheoretische und psychometrische Verfahren. Nutzentheoretische Verfahren zielen darauf ab, Bewertungen für Krankheitszustände zu bestimmen. Psychometrische Verfahren zielen dagegen zunächst einmal darauf ab, bestimmte lebensqualitätsrelevante Aspekte von Krankheitszuständen zu erfassen. Bei manchen dieser Verfahren, den sog. Indexmessverfahren, wird den auf diese Weise erfassten Krankheitszuständen noch ein Bewertungswert zugeordnet. Die nutzentheoretischen Verfahren sind damit eher allgemeine Herangehensweisen zur Erfassung von Lebensqualität. Die psychometrischen Verfahren sind dagegen immer Konkretisierungen einer ganz bestimmten Vorstellung darüber, was gesundheitsbezogene Lebensqualität ausmacht. Entsprechend gibt es mittlerweile eine sehr große, ständig wachsende Zahl psychometrischer Verfahren, aber nur sehr wenige nutzentheoretische Verfahren. Die wichtigsten nutzentheoretischen und einige prototypische psychometrische Verfahren sind in diesem Aufsatz vorgestellt worden.

Insgesamt steht die Forschung zu gesundheitsbezogener Lebensqualität noch in ihren Anfängen. Damit ist auch noch in vielerlei Hinsicht fraglich, welches Verfahren in welcher Hinsicht zur Messung gesundheitsbezogener Lebensqualität geeignet sein könnte. Bei den nutzentheoretischen Verfahren sind die wichtigsten Probleme, die mit diesen Verfahren verbunden sind, schon direkt bei der Darstellung der Verfahren diskutiert worden. Bei den psychometrischen Verfahren ist dies nicht so ohne weiteres möglich. Die Probleme liegen hier nämlich eher im Bereich der Methodologie, nach der diese Verfahren entwickelt und geprüft werden. So kann man durchaus unterschiedliche Auffassungen darüber haben, nach welchen Kriterien Validität und Reliabilität der Verfahren am besten zu prüfen sind. Ebenso sind die Prinzipien, nach denen die Fragebogenantworten in Zahlen übersetzt werden, noch durchaus diskussionsbedürftig. Insofern sind alle hier diskutierten Verfahren mit einer gewissen Vorsicht zu genießen. Andererseits haben sowohl die theoretische als auch die empirische Auseinandersetzung mit diesen Verfahren sicherlich sehr zur Präzisierung des Begriffs der gesundheitsbezogenen Lebensqualität beigetragen. Des Weiteren dürften auch die meisten der dabei entstandenen Verfahren trotz der mit ihnen verbundenen Probleme zumindest einen ungefähren Zugriff auf das erlauben, was mit gesundheitsbezogener Lebensqualität gemeint sein könnte.

Eine gewisse Vorsicht bei der Interpretation vorausgesetzt, lässt sich mithilfe dieser Verfahren sicherlich Einiges über den Nutzen medizinischer Maßnahmen aussagen. Insbesondere dann, wenn medizinische Maßnahmen miteinander verglichen werden sollen, die auf unterschiedliche medizinische Ergebnisparameter abzielen, ist es unerlässlich, mit diesen Verfahren die Auswirkungen auf die Lebensqualität zu erfassen. Beim gegenwärtigen Entwicklungsstand empfiehlt es sich dabei aber unbedingt, mehrere Verfahren parallel zu verwenden und deren Ergebnisse vergleichend zu interpretieren. Ebenso empfiehlt es sich bei der Planung und Auswertung von Untersuchungen zur Lebensqualität, Personen zu beteiligen, die mit den Problemen psychologischer Messungen und insbesondere der Messung von Lebensqualität vertraut sind. Begleitend dazu sollte die Diskussion darüber, was gesundheitsbezogene Lebensqualität ist und wie sie erfasst werden könnte, lebhaft fortgeführt werden. Ebenso ist zu diesem Thema nach wie vor eine umfassende empirische Forschung notwendig. Die Ergebnisse dieser Forschung könnten wesentlich dazu beitragen, die Grundlage gesundheitsökonomischer Entscheidungen zu verbessern.

Literatur

Bernoulli D. Specimen theoriae novae de mensura sortis. Comentarii Academiae Scientiarium Imperiales Petropolitanae 1738; 5: 175–92 (englische Übersetzung von L. Sommer in Econometrica 1954; 22: 23–36).

Brooks R & the EuroQol Group. EuroQol: the current state of play. Health Policy 1996; 37: 53–72.

Bullinger M. German translation and psychometric testing of the SF-36. Preliminary results from the IQOLA-Project. Soc Sci Med 1995; 41: 1359–66.

Bullinger M. Der SF-36 Health Survey als krankheitsübergreifendes Profilinstrument. In: Schöffski O, Glaser P, Schulenburg JM Graf v d (Hrsg). Gesundheitsökonomische Evaluationen: Grundlagen und Standortbestimmung. Berlin, Heidelberg, New York: Springer 1998; 177–87.

Claes C, Greiner W, Uber A. Der EQ-5D (EuroQol) als krankheitsübergreifendes Indexinstrument. In: Schöffski O, Glaser P, Schulenburg JM Graf v d (Hrsg). Gesundheitsökonomische Evaluationen: Grundlagen und Standortbestimmung. 2. Aufl. Berlin, Heidelberg, New York: Springer 2002; 351–65.

Cook KF, Ashton CM, Byrne MM, Brody B, Geraci J, Giesler RB, Hanita M, Souchek J, Wray NP. A psychometric analysis of the measurement level of the rating scale, time trade-off, and standard gamble. Soc Sci Med 2001; 53: 1275–85.

Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000; 17(1): 13–35.

Dolan P. Modeling valuations for health states: the effect of duration. Health Policy 1996; 38(3): 189–203.

Drummond MF, O'Brien B, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. 2nd ed. Oxford: Oxford Medical Publications 1997.

EuroQol Group. EuroQol – a new facility for the measurement of health-related quality of life. Health Policy 1990; 16: 199–208.

EuroQol-Group. Homepage of EQ-5D. http://www.Euroqol.org (September 2002).

Finlay AY, Coles EC. The effect of severe psoriasis on the quality of life of 369 patients. Br J Dermatol 1995; 132: 236–44.

Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-effectiveness in Health and Medicine. New York: Oxford University Press 1996.

Green C, Brazier J, Deverill M. Valuing health-related quality of life: A review of health state valuation techniques. Pharmacoeconomics 2000; 17: 151–65.

Hays RD, Sherbourne CD, Mazel RM. The RAND 36-item health survey 1.0. Health Econ 1993; 2: 217–27.

Hershey JC, Kunreuther HC, Schoemaker PJH. Sources of bias in assessment procedures for utility functions. In: Bell DE, Raiffa H, Tversky A (eds). Decision Making: descriptive, normative, and prescriptive interactions. Cambridge: Cambridge University Press 1988; 422–42.

Hertwig R. Sind die Gesetze des Denkens die Gesetze der Wahrscheinlichkeitstheorie und Logik? In: Mandl H (Hrsg). Bericht über den 40. Kongress der Deutschen Gesellschaft für Psychologie in München 1996. Göttingen: Hogrefe 1997; 102–13.

Jungermann H, Pfister HR, Fischer K. Die Psychologie der Entscheidung. Heidelberg, Berlin: Spektrum 1998.

Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. Econometrica 1979; 47: 263–91.

Kahneman D, Tversky A. The psychology of preferences. Sci Am 1982; 146: 160–73.

Kent G, Al-Abadie M. The psoriasis disability index – further analyses. Clin Exper Derm 1993; 18: 414–6.

Konerding U. Theorie und Messung subjektiver Einschätzungen: Entwurf einer axiomatisierten Urteilstheorie. TU Berlin: Unveröff. Dissertationsschrift 1989.

Konerding U. Psychometrische Probleme in der Gesundheitsökonomie. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 2003; 34: 125–40.

Konerding U, Schell H. Lebensqualität. In: Lauterbach K, Schrappe M (Hrsg). Gesundheitsökonomie, Qualitätsmanagement und Evidence-based Medicine. 1. Aufl. Stuttgart, New York: Schattauer 2001; 138–60.

Mehrez A, Gafni A. The healthy-years equivalents: How to measure them using the standard gamble approach. Med Decis Making 1991; 11: 140–6.

Neumann J v, Morgenstern O. Theory of Games and Economic Behaviour. Princeton, NJ: Princeton University Press 1944.

Orth B. Einführung in die Theorie des Messens. Stuttgart: Kohlhammer 1974.

Orth B. Zur Bestimmung der Skalenqualität bei „direkten" Skalierungsverfahren. Z Exp Angew Psychologie 1982; 24(1): 160–78.

Orth B, Wegener B. Scaling occupational prestige by magnitude estimation and category rating methods: A comparison with the sensory domain. Eur J Soc Psychology 1983; 13: 417–31.

Parducci A. Category judgement: a range-frequency model. Psychol Rev 1965; 72: 407–18.

Parducci A, Weddell DH. The category effect with rating scales: number of categories, number of stimuli, and method of presentation. J Exp Psychol Hum Percept Perform 1986; 12: 496–516.

Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. Med Decis Making 2001; 21(1): 17–27.

Rost J. Lehrbuch Testtheorie Testkonstruktion. Göttingen: Huber 1996.

Salek S. Compendium of Quality of Life Instruments. New York: Wiley 1999.

Schöffski O. Nutzentheoretische Lebensqualitätsmessung. In: Schöffski O, Glaser P, Schulenburg JM v d (Hrsg). Gesundheitsökonomische Evaluationen: Grundlagen und Standortbestimmung. Berlin, Heidelberg, New York: Springer 1998; 129–60.

Schöffski O, Glaser P, Schulenburg JM v d (Hrsg). Gesundheitsökonomische Evaluationen: Grundlagen und Standortbestimmung. Berlin, Heidelberg, New York: Springer 1998.

Spertus JA, Winder JA, Dewhurst TA, Deyo RA, Prodzinski J, McDonell M, Fihn SD. Development and evaluation of the Seattle Angina Questionnaire: a new functional status measure for coronary artery disease. J Am Coll Cardiol 1995; 25(2): 333–41.

Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ 1986; 1–30.

Torrance GW, Feeny D, Furlong W. Visual analog scales: Do they have a role in the measurement of preferences for health states? Med Decis Making 2001; 21(4): 329–34.

Ware JE. The SF-36 Health Survey. http://www.sf-36.com/general/sf36.html (September 2002).

Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36). Med Care 1992; 30(6): 473–83.

Westermann R. Zur empirischen Überprüfung des Skalenniveaus von individuellen Einschätzungen und Ratings. Z Psychol 1984; 192: 122–34.

Westermann R. Empirical tests of scale types for individual ratings. Appl Psychol Measurem 1985; 9: 265–74.

Westermann R, Hager W. Eine empirische Untersuchung zum Skalenniveau von Normwerten für die Bildhaftigkeit von Substantiven. Psychol Beiträge 1983; 25: 112–25.

Westermann R, Hager W. Zur Konstruktion metrischer Skalen für die Schwereeinschätzung von Delikten. Diagnostica 1985; 31: 153–63.

World Health Organization (WHO). Constitiution of the World Health Organization. In: World Health Organization (ed). Basic Documents. Genf: World Health Organization 1948.