Available online at www.sciencedirect.com
ScienceDirect
ZEFQ
www.elsevier.de/zefq
Z. Evid. Fortbild. Qual. Gesundh. wesen (ZEFQ) 104 (2010) 209–226

Schwerpunkt Ib

Zur Notwendigkeit indikationsübergreifender Nutzenmaße☆

Friedrich Breyer*

Universität Konstanz und DIW Berlin

Zusammenfassung

Der Methoden-Katalog des IQWiG hat in der Fachwelt erhebliche Diskussionen ausgelöst, und einer der wesentlichen Streitpunkte ist die Entscheidung der Autoren dieses Katalogs, bei der Nutzenmessung allein indikationsbezogene Outputmaße einzusetzen. Dagegen verlangen zahlreiche Kritiker des Entwurfs, dass der Nutzen auf einer einheitlichen Skala gemessen werden soll, die für alle Indikationen gültig und interpretierbar ist, wie etwa dem „qualitätsbereinigten Lebensjahr" (QALY). Dieser Beitrag setzt sich kritisch mit der Argumentation des IQWiG auseinander und betont die Nachteile einer rein indikationsbezogenen Nutzenmessung. Anschließend werden die rechtlichen Möglichkeiten einer indikationsübergreifenden Messung ausgelotet und Vorschläge für die weitere Vorgehensweise der Evaluation in Deutschland unterbreitet.

Schlüsselwörter: Evaluation im Gesundheitswesen, Nutzenmessung, QALY, Effizienzgrenze

(Wie vom Gastherausgeber eingereicht)

On the necessity of benefit assessments across all indications

Summary

IQWiG's General Methods catalogue has raised a controversial debate among experts, and one of the major issues is the catalogue's authors' decision to exclusively use indication-specific output measures for benefit assessments. In contrast, numerous critics of this approach demand that benefit be measured using a uniform scale which is valid and interpretable for all indications, such as the "quality-adjusted life-year" (QALY). The present article will take a critical look at the arguments put forward by IQWiG and point out the disadvantages of purely indication-specific benefit assessments. We will then explore the legal possibilities of benefit assessments independent of the type of disease, and make some suggestions for the future approach to healthcare evaluation in Germany.

Key words: healthcare evaluation, benefit assessment, QALY, efficiency frontier

(As supplied by publisher)

☆Vortrag auf dem Herbstsymposium des IQWiG am 28.11.2009 in Köln.

*Korrespondenzadresse. FB Wirtschaftswissenschaften, Universität Konstanz, Fach 135, 78457 Konstanz. Tel.: +07531 88 2568.
E-Mail: Friedrich.Breyer@uni-konstanz.de

Einleitung

In jedem Land, das ein kollektiv finanziertes Gesundheitswesen unterhält – sei es in Form eines Nationalen Gesundheitsdienstes oder einer sozialen Krankenversicherung – muss entschieden werden, mit welchem Leistungskatalog und welchen sonstigen Regeln, z.B. im Hinblick auf die Finanzierung sie ausgestattet wird. In den vergangenen Jahrzehnten sind in mehreren Ländern Initiativen ergriffen worden, die Bestimmung des Leistungskatalogs nicht mehr zufälligen Einzelentscheidungen zu überlassen, sondern zu systematisieren und rationalisieren. Leuchtende Beispiele für solche Prozesse sind etwa der Oregon Health Plan [1] und die Gründung des National Institute for Health and Clinical Excellence (NICE) in England. Letzteres ist zwar nicht entscheidungsbefugt, aber beauftragt, Richtlinien für die Aufnahme von Leistungen in den Leistungskatalog des NHS zu entwerfen.

Die Zielsetzung dieser Richtlinien ist es, mit einem begrenzten Mitteleinsatz für das kollektiv finanzierte Gesundheitswesen ein Maximum an gesundheitlichen Erfolgen im Sinne einer Verlängerung der Lebensdauer und Verbesserung der Lebensqualität für die Bürger zu erreichen. Grundlage für diese Richtlinien ist daher eine möglichst vollständige und unverzerrte Abwägung aller Nutzen und Kosten, die mit einer neuen Therapie verbunden sind. Diese Abwägung, als Health Care Technology Assessment bezeichnet, hat sich in den vergangenen Jahrzehnten zu einer etablierten wissenschaftlichen Disziplin mit internationalen Standards entwickelt. Gesamtdarstellungen finden sich in Lehrbüchern [2,3]. Eine der Methoden, die Kosten-Nutzen-Analyse, hat ein solides ethisches Fundament in der sog. Wohlfahrtsökonomik [4].

In Deutschland ist die Kosten-Nutzen-Bewertung erst seit 2007 mit dem GKV-Wettbewerbsstärkungsgesetz vom Gesetzgeber in den Prozess der Bestimmung des Leistungskatalogs in der GKV eingeführt worden. Dieser hat das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) damit beauftragt, die Evaluationen durchzuführen und im ersten Schritt einen Katalog von Methoden zu entwickeln, die dabei zur Anwendung kommen sollen. Dieser Methoden-Katalog [5] hat in der Fachwelt erhebliche Diskussionen ausgelöst [6], und einer der wesentlichen Streitpunkte ist die Entscheidung der Autoren dieses Katalogs, bei der Nutzenmessung allein indikationsbezogene Outputmaße einzusetzen. Dagegen verlangen zahlreiche Kritiker des Entwurfs, dass der Nutzen auf einer einheitlichen Skala gemessen werden soll, die für alle Indikationen gültig und interpretierbar ist, wie etwa dem „qualitätsbereinigten Lebensjahr" (QALY).

Diese Kontroverse soll im Folgenden kritisch reflektiert werden. In Abschnitt 2 werden die rechtlichen Grundlagen der Kosten-Nutzen-Bewertung in Deutschland rekapituliert, Abschnitt 3 beschreibt das Konzept der Effizienzgrenze und gibt die Argumentation des IQWiG für die rein indikationsbezogene Nutzenmessung wieder. Abschnitt 4 diskutiert die Nachteile dieser Vorgehensweise, Abschnitt 5 lotet die rechtlichen Möglichkeiten hierzu aus und Abschnitt 6 enthält Vorschläge für die weitere Vorgehensweise der Evaluation in Deutschland.

Rechtliche Voraussetzungen für die Kosten-Nutzen-Bewertung in Deutschland

Zunächst müssen die Zielsetzungen und das Verfahren der Bewertung neuer Arzneimittel und Therapieformen unterschieden werden. Die Zielsetzungen gehen aus den Paragraphen 31 und 92 des Sozialgesetzbuchs V hervor. § 31 Abs. 2a bestimmt: „Für nicht festbetragsfähige Arzneimittel setzt der GKV-Spitzenverband einen Höchstbetrag fest, bis zu dem die Krankenkassen die Kosten tragen." Darüber hinaus sagt § 92 Abs. 1: „Der Gemeinsame Bundesausschuss kann die Erbringung und Verordnung von Leistungen einschließlich Arzneimitteln einschränken oder ausschließen, wenn ... der diagnostische oder therapeutische Nutzen ... oder die Wirtschaftlichkeit nicht nachgewiesen ist."

Die zuletzt genannte Bestimmung, der Leistungsausschluss, ist sicher das schärfere Schwert im Vergleich zur Setzung eines Höchstpreises für die Erstattung durch die Krankenkassen. Er wird interessanter Weise im Methodenpapier des IQWiG überhaupt nicht erwähnt. In der Tat ist es rechtlich umstritten (7, S.4), ob ein Medikament mit einem positiven, wenn auch geringen Zusatznutzen aus dem Leistungskatalog ausgeschlossen werden darf oder ob sich § 92 lediglich auf Medikament bezieht, die überhaupt keinen Zusatznutzen mit sich bringen. Diese würden nach dem Methodenpapier des IQWiG bereits an der ersten Hürde, der Nutzenbewertung, scheitern.

Zum Verfahren hat der Gesetzgeber in § 35b, Absatz 1 bestimmt: „Das IQWiG kann vom Gemeinsamen Bundesausschuss beauftragt werden, den Nutzen oder das Kosten-Nutzen-Verhältnis von Arzneimitteln zu bewerten ...Die

Bewertung erfolgt im Vergleich mit anderen Arzneimitteln und Behandlungsformen unter Berücksichtigung des therapeutischen Zusatznutzens für die Patienten im Verhältnis zu den Kosten. Beim Patientennutzen sollen insbesondere die Verbesserung des Gesundheitszustands . . . sowie eine *Verbesserung der Lebensqualität*, bei der wirtschaftlichen Bewertung soll *auch die Angemessenheit und Zumutbarkeit einer Kostenübernahme durch die Versichertengemeinschaft* berücksichtigt werden. . . . Das Institut bestimmt über die Methoden und Kriterien für die Erarbeitung von Bewertungen *auf der Grundlage der . . . anerkannten internationalen Standards der evidenzbasierten Medizin und der Gesundheitsökonomie*" (Hervorhebungen von mir). Im zweiten Absatz heißt es weiter: „Die Bewertungen nach Absatz 1 werden dem Gemeinsamen Bundesausschuss als Empfehlung zur Beschlussfassung nach § 92 Absatz 1 Satz 2 Nr.6 zugeleitet."

Die rechtlichen Vorgaben ergeben im Hinblick auf die Frage, ob die Nutzenmessung indikationsbezogen oder indikationsübergreifend erfolgen soll, bereits ein eindeutiges Bild: Zum einen werden „anerkannte internationale Standards der Gesundheitsökonomie" angesprochen, zu denen es ohne Zweifel gehört, den Nutzen von Therapien mit einem allgemein aussagekräftigen Maß wie dem QALY oder der Zahlungsbereitschaft in Geldeinheiten zu messen. Zum anderen wird der Patientennutzen dahingehend konkretisiert, dass es um die Lebensqualität gehen soll. Diese lässt sich jedoch nicht mit klinischen Outputmaßen erfassen, sondern erfordert eine subjektive Qualitätskomponente, wie sie etwa beim QALY vorliegt.

Die Effizienzgrenze und die Mehrdimensionalität des Nutzens

Kernstück des vom IQWiG vorgelegten Methodenvorschlags ist die sog. Effizienzgrenze. Dieses auf Markowitz [8] zurückgehende Konzept dient eigentlich der Bewertung von Wertpapieren innerhalb eines Portfolios nach den Kriterien von Rendite und Risiko. Ein Wertpapier heißt demnach effizient, wenn es weder von einem anderen Wertpapier noch von einer linearen Kombination zweier Papiere in dem Sinne dominiert wird, dass diese(s) eine höhere Rendite und gleichzeitig ein geringeres Risiko aufweist. Die Menge der effizienten Papiere sowie ihre Linearkombinationen bilden dann im zweidimensionalen Rendite-Risiko-Raum die Effizienzgrenze. Ganz abgesehen von der Frage, ob sich diese Grenze auf Kombinationen dieser beiden Kriterien extrapolieren lässt, die außerhalb des Beobachtungsraums der bislang bekannten Papiere liegen, steht und fällt die Anwendung dieses Konzepts damit, dass es außer diesen beiden Kriterien keine weiteren gibt. Auf den Fall der Bewertung von Arzneimitteln übertragen, lautet die Frage daher: Kann man den Nutzen aus einem Arzneimittel generell in einer einzigen Kenngröße ausdrücken (die dann mit den Kosten den zweidimensionalen Entscheidungsraum definiert)? Wir wissen, dass in der medizinischen Praxis fast immer mehrere Outcome-Dimensionen (mehrere positive Wirkungen, evtl. Nebenwirkungen) existieren, und es fragt sich daher, wie in diesem Fall vorzugehen ist. Dazu sind prinzipiell drei verschiedene Vorgehensweisen denkbar:

a) Man könnte sich auf eine einzige Outcome-Dimension beschränken, die man für besonders wichtig hält.[1] Diese Auswahl wäre aber willkürlich und höchstens dann zu rechtfertigen, wenn die Effektstärken in dieser Dimension diejenigen in den anderen Dimensionen deutlich übertreffen.
b) Man konstruiert für jede Outcome-Dimension eine gesonderte Effizienzgrenze: Dieses Verfahren wurde von IQWiG [5], S.31 vorgeschlagen: „Wird eine Kosten-Nutzen-Bewertung unter Verwendung verschiedener klinischer Maße durchgeführt, so wird für jedes klinische Maß eine Effizienzgrenze erstellt," und von Uwe Siebert in seiner Pilotstudie „Antivirale Therapie von chronischer Hepatitis C" praktiziert; es wirft aber die Frage auf, welche Schlüsse zu ziehen sind, wenn sich die Ergebnisse der verschiedenen Effizienzgrenzen widersprechen [9].
c) Man konstruiert aus den verschiedenen Outcome-Dimensionen einen Index. Dies ist der Vorschlag in IQWiG [5]. Dort wird ein „indikationsspezifisches integriertes Nutzenmaß" angeregt, ohne dass jedoch näher darauf eingegangen wird, wie man sich dieses vorzustellen hat.

Von diesen drei Antworten kann nur die dritte überzeugen. Auch sie lässt jedoch eine sich unmittelbar aufdrängende Frage offen: Wenn schon ein integriertes Nutzenmaß benötigt wird, warum soll dann nicht eines verwendet werden, das es schon gibt und das verbreitet angewendet wird, nämlich QALYs? Und was spricht dagegen, ein indikationsübergreifend interpretierbares Maß zu verwenden, wenn dieses nun einmal da ist? Anders ausgedrückt: Wel-

[1] Dies entspricht der Vorstellung des federführenden Mitglieds des Expertenpanels, Jaime Caro (mündliche Kommunikation auf der European Conference on Health Economics im Juli 2008 in Rom).

cher Schaden entsteht durch eine zusätzliche Funktion (nämlich indikationsübergreifend zu sein), wenn die hier zunächst erforderliche Funktion (mehrere Nutzendimensionen zu einem Index zusammenzufassen) erfüllt ist? Auf diese Fragen wird im folgenden Abschnitt eingegangen.

Für und Wider indikationsübergreifender Nutzenmaße

Die Haltung des IQWiG

Auf S.32 wiederholen die Autoren von IQWiG [5] ihren Vorschlag: „Eine weitere Möglichkeit der Darstellung des Nutzens auf der Nutzenachse der Effizienzgrenze ist die Aggregation verschiedener Nutzenparameter zu einem einzigen Maß und die anschließende Erstellung einer einzelnen Effizienzgrenze", um dann fortzufahren: „Da die Kosten-Nutzen-Bewertung in Deutschland nicht indikationsübergreifend, sondern innerhalb einer Indikation eingesetzt werden soll, kann mit indikationsspezifischen aggregierten Maßen gearbeitet werden. Es ist nicht notwendig, primär indikationsübergreifende aggregierte Maße zu verwenden. IQWiG lehnt indikationsübergreifende Outcome-Maße ab." Ähnlich argumentierte Jaime Caro bei seinem Vortrag am 26.7.2008 in Rom: „Consistency across therapeutic areas is not clearly defined and is not needed."[2] Begründet wird diese Auffassung mit der Behauptung, in Deutschland gebe es kein festes Budget für Gesundheitsausgaben.

Diese Argumentation kann nicht überzeugen, denn zumindest für die Gesetzliche Krankenversicherung gilt seit mehreren Jahrzehnten die politische Zielsetzung, die Beitragssätze möglichst stabil zu halten. Dies wird noch durch die regelmäßige Setzung von globalen Budgets in den einzelnen Leistungsbereichen unterstrichen. Damit muss den Entscheidungsträgern im Gesundheitswesen und gerade dem Gemeinsamen Bundesausschuss, der für den Leistungskatalog zuständig ist, klar sein, dass jede positive Vergütungsentscheidung Opportunitätskosten – auch innerhalb des Gesundheitssystems – hat.

Man kann das Vereinigte Königreich nur dazu beglückwünschen, dass die Erkenntnis der Opportunitätskosten einer jeden Ressourcenentscheidung dort schon weiter gediehen ist als in unserem Land. So liest man im British Medical Journal in einer Debatte über „orphan drugs": „In a system with finite resources that do not meet all needs, money spent on one service means that some other service cannot be provided (opportunity cost). ... Commissioning decisions should not be posed as isolated questions but need to take into account other priorities." [10].

Nachteile des Fehlens indikationsübergreifender Nutzenmaße

Wenn diese Opportunitätskosten nicht beachtet und – in Ermanglung eines transparenten indikationsübergreifenden Nutzenmaßes – die Entscheidungen im Spitzenverband der GKV über Höchstpreise immer nur fallweise getroffen werden, so drohen eine Reihe negativer Konsequenzen:

1. Die Entscheidungen, die das Gremium im Laufe der Zeit trifft, werden nicht miteinander konsistent sein: Beispielsweise wird man für den gleichen Gewinn an Lebensdauer bzw. -qualität starke Unterschiede in den von der Solidargemeinschaft zu tragenden Kosten beschließen. Nur die Orientierung an einem indikationsübergreifenden Nutzenmaß könnte dem Gremium selbst mehr Transparenz über die Konsequenzen seiner Entscheidungen geben und ihm damit helfen, mehr Konsistenz herzustellen.
2. In der Folge ergeben sich Ungerechtigkeiten zwischen Patientengruppen, die an verschiedenen Krankheiten leiden, wenn in einem Indikationsgebiet mehr Geld für den gleichen Zuwachs an Gesundheit ausgegeben wird als in einem anderen.
3. Wenn diese das erfahren, so könnte es zu Klagen von Betroffenen bei den Sozialgerichten kommen, wie Huster [11] befürchtet.

Rein indikationsbezogene Outcome-Maße haben zudem einen weiteren Nachteil: Denn die Steigung der Effizienzgrenze in einem Diagramm, das eine medizinische Ergebnisvariable zu Kosten in Beziehung setzt, kann nicht direkt mit einer Zahlungsbereitschaft verglichen werden, da sich letztere immer auf eine interpretierbare Nutzengröße bezieht. Dies liegt daran, dass Versicherte keine abstrakten klinischen Endpunkte nachfragen, sondern „Gesundheit" und Lebensqualität. Für Lebensqualität ist die Solidargemeinschaft folglich bereit, etwas zu zahlen. Ob sie für eine Blutdrucksenkung um eine bestimmte Zahl von Millimetern zu zahlen bereit ist, kann nicht beantwortet werden, ohne den Einfluss dieser Senkung z.B. auf das Risiko einer schweren Erkrankung (mit Verkürzung der Lebensdauer oder Beeinträchtigung der Lebensqualität) zu kennen.

Diesem Argument kann man auch nicht mit dem Einwand begegnen, Zahlungsbereitschaften seien für den Spitzenverband eine irrelevante Kategorie. Denn die Festlegung eines Höchstprei-

[2] Vortragsfolien und mündliche Kommunikation.

ses für ein neues Medikament ist ja nichts anderes als der Ausdruck einer Zahlungsbereitschaft, und diese wiederum sollte sich an der Zahlungsbereitschaft der Versicherten orientieren, die der Spitzenverband vertritt und in deren Auftrag er handelt. Genau dies dürfte der Gesetzgeber gemeint haben, wenn er in § 35b SGB V von der „Zumutbarkeit einer Kostenübernahme durch die Versichertengemeinschaft" gesprochen hat. Diese Zumutbarkeit ist nämlich dann gegeben, wenn der mit den Kosten erzielte *Wert* es rechtfertigt. Dieser kann jedoch nur subjektiv bestimmt werden und wird durch die Zahlungsbereitschaft der Versicherten für die Gesundheitsverbesserung ausgedrückt.

Rechtliche Bewertung

Welche Vorgaben der Gesetzgeber über das zu verwendende Nutzenmaß gemacht hat, lässt sich aus einem Rechtsgutachten von Huster [11] entnehmen: Dieser stellt zunächst fest, dass der Gesetzgeber keine bestimmte Methodik zwingend vorgeschrieben habe (ebenda, S.7).
Andererseits mahnt er an, dass nach dem verfassungsrechtlichen Gleichbehandlungsgebot (Art. 3 GG) eine „massive Ungleichbehandlung der Versicherten und der Pharmaunternehmen" wegen Abhängigkeit der Entscheidungen vom bestehenden Preisniveau nicht zulässig sei (S.12). Weiter führt er aus, dass § 35b SGB V eine indikationsübergreifende Bewertung weder erzwinge noch ausschließe, dass diese jedoch für den Spitzenverband unumgänglich sei: „Spätestens der SpiBu als das eigentliche Entscheidungsorgan, das öffentliche Gewalt ausübt, ist aus Gründen des verfassungsrechtlichen Gleichheitsgebotes verpflichtet, die Festsetzung des Erstattungshöchstbetrages auch im Verhältnis der Indikationen und Maßnahmen nicht willkürlich vorzunehmen" (S.17). Ferner müsse er eine Begründung geben, „warum in dem einen Fall dieser, im anderen Fall ein anderer Höchstbetrag für einen bestimmten Zusatznutzen festgesetzt worden ist. Da diese Begründung maßgeblich auf Art und Ausmaß des Zusatznutzens bezogen sein wird, ist irgendein indikationsübergreifender Vergleich unumgänglich" (ebenda).
Dies gilt ungeachtet der Tatsache, dass das bislang am breitesten verwendete indikationsübergreifende Nutzenmaß, das QALY, vielfältiger Kritik ausgesetzt ist ([12–14] sowie Weyma Lübbe in ihrem Beitrag zu diesem Symposium). In der Tat lässt sich die Verwendung von QALYs als Outputmaß weder rechtlich durchsetzen noch gibt es zwingende ökonomische Gründe, die für QALYs sprechen. Falls das IQWiG ein eigenes Maß entwickeln und anwenden wollte, das diesen Zweck erfüllt, so ließe sich das nicht kritisieren.

Wie könnte es weitergehen?

Aus ökonomischer Sicht und im Einklang mit der Rechtslage wäre das folgende Vorgehen wünschenswert:

1. Der Gemeinsame Bundesausschuss sollte dem IQWiG den Auftrag erteilen, den medizinischen Nutzen in einem indikationsübergreifenden Maß zu erfassen. Falls kein anderes Maß zur Verfügung steht, sollte vorläufig mit dem Maß der QALYs gearbeitet werden.
2. Wenn der Spitzenverband der Gesetzlichen Krankenkassen eine Höchstpreissetzung vornimmt, so sollte er gleichzeitig angeben, wie hoch die damit implizit verbundenen „Kosten je QALY" sind.
3. Um Konsistenz seiner Entscheidungen zu gewährleisten, sollte er einen „Normbereich" für akzeptierte Kosten je QALY definieren. Abweichungen von diesem Normbereich sollte er gesondert begründen.
4. Langfristig sollte angestrebt werden, die Präferenzen der Bevölkerung im Hinblick auf die Zahlungsbereitschaft für Gesundheitsverbesserungen (z.B. für ein QALY) zu ermitteln. In anderen Ländern gibt es schon umfangreiche Erfahrungen mit derartigen Erhebungen [15]. Auch in Deutschland sind Forschungsprojekte auf dem Weg, die dieses anstreben. Zu erwähnen ist hier die DFG-Forschergruppe FOR655, die es sich u.a. zum Ziel gesetzt hat, die Präferenzen der Bevölkerung hinsichtlich der Priorisierung von Gesundheitsleistungen zu ermitteln [16].

Literatur

[1] Garland MJ. Rationing in Public: Oregon's Priority-Setting Methodology. In: Fein A, et al., editors. Rationing America's Medical Care: the Oregon Plan and beyond. Washington, D.C.: Brookings Institution Press; 1992. p. 37–59.
[2] Drummond, M.F. (2005), *Methods for the economic evaluation of health care programmes*, 3rd ed., Oxford, New York.
[3] Schöffski, O. und von der Schulenburg, J.M. (2008), *Gesundheitsökonomische Evaluationen*, Berlin.
[4] Breyer, F., Zweifel, P. und Kifmann, M. (2005), *Gesundheitsökonomik*, 5. Aufl., Heidelberg u.a.
[5] IQWiG (2009a), *Entwurf einer Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung*, Version 2.0, Köln.
[6] IQWiG (2009b), *Dokumentation der Stellungnahmen zum ‚Entwurf einer Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung Version 2.0*, Version 1.0 vom 12.10.2009.

[7] Huster, S. (2009), *Die Methodik der Kosten-Nutzen-Bewertung in der Gesetzlichen Krankenversicherung. Analyse der rechtlichen Vorgaben, Vortrag auf dem 2*. Kölner Medizinrechtstag, 20.11.2009.
[8] Markowitz H. Portfolio Selection. Journal of Finance 1952;7:77–91.
[9] Rothgang, H. (2009), *Kommentare zur Pilotstudie ‚Antivirale Therapie von chronischer Hepatitis C" von Uwe Siebert, Workshop zur Vorstellung und Diskussion der Ergebnisse der Pilotprojekte zur Bewertung von Kosten-Nutzen Relationen mit der Methode der Effizienzgrenzen am 30*. Juni 2009 in Berlin.
[10] Burls A, Austin D, Moore D. Commissioning for Rare Diseases: View from the Frontline. British Medical Journal 2005;331:1019–21.
[11] Huster, S. (2008), *Die Methodik der Kosten-Nutzen-Bewertung in der Gesetzlichen Krankenversicherung - Analyse der rechtlichen Vorgaben - Gutachterliche Stellungnahme*, Juni.
[12] Ubel PA, Nord E, Gold M, Menzel P, Pinto Prades J-L, Richardson J. Improving Value Measurement in Cost-Effectiveness Analysis. Medical Care 2000;38:892–901.
[13] Dolan P, Shaw R, Tsuchiya A, Williams A. QALY Maximization and People's Preferences: A Methodological Review of the Literature. Health Economics 2005;14:197–208.
[14] Richardson J, McKie J. Empiricism, ethics and orthodox economic theory: what is the appropriate basis for decision-making in the health sector? Social Science and Medicine 2005;60:265–75.
[15] Telser H, Becker und K, Zweifel P. Validity and Reliability of Willingness-to- Pay Estimates: Evidence from Two Overlapping Discrete-Choice Experiments. The Patient 2008;1(4):283–98.
[16] Diederich, A., Schnoor, M. Winkelhage, J. und Schreier, M. (2009), *Präferenzen in der Bevölkerung hinsichtlich der Allokation medizinischer Leistungen – Entwicklung eines Fragebogens für eine repräsentative Bevölkerungsbefragung*, Diskussionspapier FOR 655 Nr. 21.

Schwerpunkt Ib

Measures of efficiency in healthcare: QALMs about QALYs?

Michael Schlander[a,b,c,*]

[a]Institute for Innovation & Valuation in Health Care (InnoVal[HC])
[b]Universität Heidelberg, Medizinische Fakultät Mannheim (Institut für Public Health)
[c]Hochschule für Wirtschaft Ludwigshafen

Summary

Comparative economic evaluations are concerned with the relative efficiency of alternative uses for scarce resources. Cost-benefit analysis (CBA) is grounded in economic welfare theory and attempts to identify alternatives with a net social benefit, measuring the created value in terms of individual willingness to pay (WTP). In applied health economics, cost-effectiveness evaluation (CEA) is more widely used than CBA, adopting a modified efficiency criterion, minimization of incremental costs per quality-adjusted life year (QALY) gained ("cost-utility analysis," CUA).
CBA has been greeted with skepticism in the health policy field, primarily owing to resistance to a monetary measure of benefit and owing to concerns that WTP may be unduly influenced by ability to pay. The move to CUA, however, has not been without problems. The framework deviates from economic theory in important aspects and rests on a set of highly restrictive assumptions, some of which must be considered as empirically falsified. Results of CUAs do not seem to be aligned with well-documented social preferences and the needs of healthcare policy makers acting on behalf of society. By implication, there is reason to assume that a context-independent value of a QALY does not exist, with potentially fatal consequences for any attempt to interpret CUAs in a normative way. Policy makers seem well advised to retain a pragmatic attitude towards the results of CUAs, while health economists should pay more attention to the further development of promising alternative evaluation paradigms as opposed to the application of algorithms grounded in poor theory.

Key words: efficiency, cost-benefit analysis, cost-effectiveness analysis, cost-utility analysis, willingness to pay. quality-adjusted life year (QALY)

*Korrespondenzadresse. Institute for Innovation & Valuation in Health Care–InnoVal[HC], An der Ringkirche 4, D-65197 Wiesbaden.
Tel.: +49 0 611 4080 789 10; fax: +49 0 611 4080 789 99.
E-Mail: michael.schlander@innoval-hc.com

Zusammenfassung

Vergleichende ökonomische Evaluationen gelten dem effizienten Einsatz knapper Ressourcen. Kosten-Nutzen-Bewertungen (KNBs) im engeren Sinn beruhen auf der wohlfahrtsökonomischen Theorie und nehmen die individuelle Zahlungsbereitschaft als Maß des Nutzens. In der angewandten Gesundheitsökonomie wird demgegenüber die Methode der Kosten-Effektivitäts-Analyse (CEA) häufiger eingesetzt. Das Effizienzkriterium in der Spielart von „Kosten-Nutzwert-Analysen" (CUAs) ist dann die Minimierung der inkrementalen Kosten je (zusätzlich) produziertem Qualitäts-adjustierten Lebensjahr (QALY).

KNBs wurden im Gesundheitssektor mit Skepsis aufgenommen, primär wegen der Monetarisierung von Nutzen, aber auch wegen der befürchteten Abhängigkeit der Zahlungsbereitschaft von der Zahlungsfähigkeit. Ihr weitgehender Ersatz durch CUAs wirft zahlreiche Probleme auf. CUAs entsprechen in wesentlichen Punkten nicht der ökonomischen Theorie und beruhen auf äußerst restriktiven Annahmen, die teilweise als empirisch falsifiziert gelten müssen. Die Ergebnisse von CUAs stehen nicht im Einklang mit gut dokumentierten gesellschaftlichen Präferenzen. Als Folge muss davon ausgegangen werden, dass es eine kontextunabhängige Zahlungsbereitschaft für ein QALY nicht gibt, mit potenziell verheerenden Folgen für jeden Versuch einer normativen Interpretation der Ergebnisse von CUAs. Gesundheitspolitische Entscheidungsträger sollten deshalb eine pragmatische Einstellung gegenüber CUAs bewahren. Wirtschaftswissenschaftler sollten der Entwicklung viel versprechender alternativer Paradigmen für gesundheitsökonomische Evaluationen mehr Aufmerksamkeit widmen als der Anwendung von Algorithmen, die einer hinreichenden theoretischen Fundierung entbehren.

Schlüsselwörter: Effizienz, Kosten-Nutzen-Analyse, Kosten-Effektivitäts-Analyse, Kosten-Nutzwert-Analyse, Zahlungsbereitschaft, Qualitäts-adjustiertes Lebensjahr (QALY)

Economic evaluation of health care programs can take different forms. One group of analyses is purely descriptive in nature, such as burden of disease, cost of illness, and health care utilization studies. Such studies can adopt a variety of perspectives and may offer useful insights. However, they do not provide helpful information to health care policy makers seeking to increase the efficiency of health care delivery. This objective can be met only by comparative evaluation falling into the branch of normative health economics.

Then, economic evaluations are a tool for systematically weighing the benefits of a technology (which may be use of a product, of a procedure, or else) against the costs incurred by its adoption. As such, they attempt to assess the social desirability of one program compared to some alternative. Given scarcity of resources available (which does not require a fixed budget constraint), not everything is affordable that might produce at least some marginal benefit, and choices need to be made. In Ezra Mishan's (1969, p. 13) [1] words, "Theoretical welfare economics is ... that branch of study which endeavors to formulate propositions by which we may rank, on the scale better or worse, alternative economic situations open to society." Evidently, the terms "better" and "worse" are explicitly normative ones.

Cost benefit analysis

From the welfare economic perspective, **cost benefit analysis (CBA)** represents the standard procedure to achieve this objective. A majority of economists indeed seem to regard the results of CBAs as normative statements about what ought to be done [2-8], while a minority emphasize the importance of ("other") ethical aspects in economics [9,10]. Although having been subject of passionate controversy [11], CBA has been widely adopted in areas such as transportation, occupational risk and environmental protection. It is used to determine whether the amount of the benefits of a public program i, B_i, 'to whomsoever they accrue', exceed their estimated costs, C_i [6,8]. Accordingly the decision rule of CBA simply requires a positive net social benefit (NSB) of program i in order to recommend its adoption:

$$B_i > C_i \tag{1.1}$$

or for that matter

$$B_i/C_i > 1, \tag{1.2a}$$

or

$$\Delta B_i/\Delta C_i > 1, \tag{1.2b}$$

taking the principle of marginal evaluation into account, which is precisely the definition of efficiency;

$$NSB_i = B_i - C_i \tag{1.3}$$

$$NSB_i > 0, \tag{1.4}$$

or

$$NSB_i = \sum_{t=1}^{n} \frac{B_i(t) - C_i(t)}{(1+r)^{t-1}} \tag{1.5}$$

(when discounting with a rate r of future costs and benefits is included, in order to compare present values). This concept mirrors the process of net present value (NPV) calculation for private sector investment decisions. If a constraint (for example, a fixed budget) is added, then in the efficiency criterion (1.2) unity is replaced by a threshold reflecting the oppor-

tunity cost of the constrained resource.

CBA not only requires the benefits to be expressed in monetary terms, but also to measure costs incurred from a societal perspective, which corresponds to their interpretation as opportunity costs. Under the scarcity condition, opportunity costs are defined as the value that might have been created with the best alternative use of the resources committed to program *i*. As indicated, this approach is firmly grounded in economic welfare theory, with benefits being valued from the perspective of the *individuals* concerned, and their *maximum* willingness to pay (WTP) expressing their strength of preference for the program (including, but not necessarily restricted to, its outcomes), i.e., reflecting their expected utility gains [12].

Importantly, then social welfare W is assumed to be captured adequately by some aggregate of individual utility U, with U^1, U^2, ... U^m representing continuous individual utility functions, i.e., utility as assessed by the individuals themselves (who are considered the best judges of their own welfare),

$$W = W\ (U^1,\ U^2,\ \ldots,\ U^m), \tag{1.6}$$

which is equivalent to assuming that social welfare depends only on the welfare of the individuals.

Equation (1.6.) is a so called Bergson-Samuelson welfare function. In principle, this social welfare function (SWF) may take different forms depending on the degree of inequality aversion prevalent in society. Importantly, it does by no means stipulate a simple additive aggregation of individual utility, which – as a special case – would be represented by the act utilitarian SWF,

$$W = \sum_{j=1}^{m} U^j \tag{1.7}$$

In economics there has been some controversy, and possibly confusion, over utility measurability. At least in part, as Ng (2004, p. 15 [7]) noted, "this is due to the ambiguous use of the term utility both as a measure of subjective satisfaction and as an indicator of objective choice or preference." Originally, the utilitarian philosophers assumed that people ought to desire such things that will maximize their utility, with utility defined as the tendency to increase or decrease happiness, i.e., to bring either pleasure (positive utility) or pain (negative utility). Jeremy Bentham (1748-1832) and his classical followers hoped for the development of techniques that would enable direct measurement of utility [13]. Francis Edgeworth (1845-1926) for example proposed a "hedonimeter" [14]. In the meantime, however, they believed that the best approximation they had at hand was actual behavior in the marketplace. Actual choices made by people were believed to reflect the quantity of utility derived from these choices. In line with this approach, rational choice theory prescribes the most effective ways to achieve utility given desires [15]. Rational choice does require consistency of desires, but the theory does not put any further constraints on what people (should) want. It does not offer answers to questions like: is it rational for obese persons to overeat; for young people to undersave; for car drivers not to use seatbelts; and so on. Obviously preference-based utility may differ from welfare due to ignorance and imperfect foresight, i.e., there may be profound differences between *ex ante* expectations (or fears) and *ex post* welfare.

Furthermore, preferences of individuals may be influenced not only by concerns for their own well-being but also by their consideration of the welfare of others. There are compelling examples for "altruistic" behavior that don't lend themselves to reconstruction as a utility gain because of feeling better due to doing something good. Citizens may vote for a political party because they believe their country will be better off with that party in government, even though they know they won't be better off as individuals; parents are frequently prepared to sacrifice their own happiness for the welfare of the children; and so on. In fact, (at least outside the conventional welfare economic framework) it is widely accepted practice to use welfare instead of actual preferences (behavior) for normative purposes, for example by necessitating compulsory and sometimes heavily subsidized private pension plans to counter the irrationality of insufficient savings for old age.

Using the concept of marginal analysis, the additional pleasure (or pain) derived from one additional unit of a good was all that was needed for economic analysis [16]. As Hermann Gossen (1810-1858) put it, "Man maximizes his total life pleasure if he distributes his entire money income [...] among his various enjoyments [...] so that the last atom of money spent on each single pleasure yields the same amount of pleasure" (cited in [17], p. 244):

$$\frac{MU_1}{p_1} = \frac{MU_2}{p_2} = \ldots \frac{MU_i}{p_i} \text{ for all goods } i, \tag{1.8}$$

where MU_i is the marginal utility of good i, and p_i is its price.

The concept of marginal (instead of absolute) utility, while still in keeping with the idea of cardinal measurability, was eroded by the analytical problem of utility dependence, i.e., the fact that utilities of different goods are not independent from each other: the marginal utility from a gallon of

petrol, for instance, will depend on the type of vehicle owned. Edgeworth proposed that the utility of a bundle of goods x_1, x_2, ..., x_n should be conceptualized as a multidimensional construct,

$$U = U(x_1, x_2, \ldots, x_n), \quad (1.9)$$

with each good representing one dimension,
and that bundles having the same [''multiattribute''] utility value could be linked by an indifference curve. From there it was a quick leap to the ordinal revolution, with Vilfredo Pareto (1848-1923) rejecting the idea that utility needed to be quantified. Mapping preferences on Edgeworth's indifference curves was sufficient to enable economic analysis, which simply required pairwise comparisons between different bundles. It is also possible to conclude that a bundle of goods, because it was chosen by a consumer, must represent a point on the highest indifference curve. This leads directly to the concept of ''revealed preferences'' [18].
Of note, ordinal utility (restricted to a ranking of states) no longer has a relationship with any absolute degrees of happiness, and numbers assigned to states cannot be combined across people. As a consequence, hypothetical compensation tests are required if one state does not represent an absolute Pareto improvement over another, i.e., whether the gains of winners are great enough that they might effectively compensate losers. It is immediately evident that in major parts of clinical medicine, particularly those that should be an essential part of a basic ''health benefit basket'' covered by a collectively financed health scheme [19], compensation of losers for health care foregone seems a rather theoretical proposition [20-22].
An important practical advantage of adopting equation (1.7) instead of (1.6) is that social welfare is not incomparable if some W^i (=U^i!) increase and some decrease. The technical difficulty, of course, is the need to find a common unit of utility that can be measured cardinally; in other words, to allow summing up, the utility functions must be ''unit comparable'' [23]. Solving this technical issue satisfactorily would still leave policy-makers with the ethical implications of interpersonal prioritization of services on grounds of the efficiency criterion (1.2).
As indicated above, Bentham and his early followers had been primarily concerned with the maximization of welfare and reverted to its approximation, however rough, by utility defined as strength of preference (which in turn was equated with [maximum] individual willingness to pay [12]). Over the past two decades, a new empirical approach to utilitarianism has emerged. This has been driven by Daniel Kahneman's [24] project to explore experienced (*ex post*) utility as opposed to decision (expected, *ex ante*) utility, even though Kahneman attempted to distance himself from ''Bentham's view of pleasure and pain as sovereign masters of human action'' (Kahneman et al., 1997, p. 377 [24]). Here the interested reader can only be referred to Kahneman's thoughtful analysis of the underlying reasons why (*ex ante*) decision utility systematically differs from (*ex post*) experienced utility. Kahneman and colleagues [24] concluded, ''Admitting experienced utility as a measure of outcomes turns utility maximization into an empirical [note added by MS: i.e., falsifiable] proposition, which will probably be found to provide a good approximation to truth in many situations and to fail severely in others. The scientific merit of economic analyses that assume utility maximization will vary accordingly'' (ibid., p. 397). Note that socitey may not, of course, wish to *maximize* happiness. We may for example give resources to a badly injured soldier even though they will never be capable of true happiness again.
Nevertheless, the willingness-to-pay approach represents a powerful concept that is flexible to accommodate any dimension of benefit deemed relevant. It can for example incorporate process-related utility, instead of consequences only. Of note, compassionate externalities (such as benefits people obtain from caring) can also be incorporated into the framework. (Then at least two issues arise, (a) the revealed preference definition of utility reduces to a tautology, since after all, by definition, any observed actions reflect utility; and (b), should that possibly imply that if others obtain utility if a given patient lives, that patient should live – and *vice versa*?) In practice, the latter approach has not been operationalized in health economics [25]. As measurements of the (''selfish'') benefit of such externalities have never been done, this theoretical argument may well appear as a defense to rationalize the purely individual focus of existing practice.

Cost effectiveness analysis

In any event, health care policy makers have not enthusiastically welcomed CBA of clinical interventions. Apart from sensibilities against the monetary measurement of health benefits *per se*, there have been widespread concerns that the individual WTP measure may ''inherently favor the wealthy over the poor'' [26]. i.e., that it may be contaminated by differences in ability to pay and therefore lead to recommendations skewed in favor of the rich. The consequences for a collectively financed health scheme would indeed seem paradox, as they implied that – with benefits

defined by WTP and thus in part determined by income – members of the health scheme would have to support the wealthy more than the poor. The most obvious response is to adjust WTP to take account of the distribution of income [27], but this leads to a metric that is difficult to interpret and neither in line with welfare theory nor with the "extrawelfarist" framework, which has one of its roots in the rejection of individual willingness-to-pay [28-30]. A key feature of applied extrawelfarism (cost utility analysis) is the exclusive focus on health-related outcomes and, corresponding to its "decision-making perspective" [31] and despite intense scholarly debate in this respect (cf. Table 2) [26,32-36], costs from the perspective of the health scheme in question ("payers perspective") [33-35]. It has been claimed [30] that the extrawelfarist approach can be traced back to Amartya Sen's theory of capabilities and functionings [36-39], a proposition that will have to be discussed briefly (see "*Some limitations and critique*," below) [36-40].

The resulting move to **cost effectiveness analysis (CEA)** can be stylized by separating the effects of an intervention from their valuation (pricing) [6]:

$$B_i = P_i \, x \, E_i \tag{2.1}$$

with (1.1) resulting in

$$P_1 \cdot E_1 > C_1 \tag{2.2}$$

$$\frac{P_1 \cdot E_1}{C_1} > 1 \tag{2.3}$$

It appears noteworthy that, with this departure from CBA, a number of restrictive assumptions are introduced simultaneously, including an explicit exclusive analytic focus on some set of defined *consequences* (health-related outcomes, "effects" E) and a change of perspective, since a hypothetical health care decision maker is thought of wishing to maximize either specific health outcomes (the "effects" of CEA) or total aggregate health gains for the community (in cost utility analysis, see below) under a given budget constraint.

This approach leads to a ranking of interventions according to:

$$\frac{P_1 \cdot E_1}{C_1} > \frac{P_2 \cdot E_2}{C_2} \tag{2.4}$$

The monetary value, or "price", P_i, attached to an effect, E_i, is often interpreted as the (marginal) willingness-to-pay for a(n additional) unit of this effect, and this WTP can be looked at from an individual perspective (in this regard consistent with economic welfare theory) or, alternatively, from a policy-maker's or "social" perspective. Accounting for the economic principle of comparing alternatives at the margin, this can be rewritten for CEAs (having specified both indication and target population, the incremental costs and effects of an intervention versus a defined alternative will be of interest) as:

$$\frac{P_1 \cdot \Delta E_1}{\Delta C_1} > \frac{P_2 \cdot \Delta E_2}{\Delta C_2} \tag{2.5}$$

If the prices P_1 and P_2 of the respective effects E_1 and E_2 are considered the same, which of course is valid only if a common unit of effect is being compared (which needs to be measurable on an interval or "cardinal" scale, with equal differences along the scale implicitly being considered of equal value, irrespective of where on the scale they occur; cf. (3.4) and Fig. 4, below), this reduces to

$$\frac{\Delta E_1}{\Delta C_1} > \frac{\Delta E_2}{\Delta C_2} \tag{2.6}$$

which represents the efficiency criterion of CEA,

or

$$\frac{\Delta C_1}{\Delta E_1} < \frac{\Delta C_2}{\Delta E_2} \tag{2.7}$$

In the CEA model, analyses are thus confined to a comparison of the incremental costs and effects of two (or more, cf. Fig. 5, below) alternatives. This can be visualized by means of a widely used diagram, the "cost effectiveness plane" (Fig. 1) [41].

In CEA, integrated measures can be used if a bundle of effects is of interest, instead of a single outcome. This metric *may* be one out of the group of health-adjusted life years (HALYs), which have in common that they integrate two dimensions, quality and length of life, into one measure [42], and it may be determined by means of a multiattribute utility (MAU) model. Importantly, in the context of simple CEA the (individual or social) willingness-to-pay for a HALY unit will be situation-specific; this type of analysis is primarily concerned with issues of technical efficiency [43,44].

As different programs produce different types of outcomes, health-adjusted life years – in practice, most commonly the QALY variant, see below – promised to solve the "apple and pies" problem and provide for a tool enabling to determine allocative efficiency across all sorts of health-related goods and services. This is the objective of so called "cost utility analysis" [21,26,43,44].

Cost utility analysis

In response to the desire for a common denominator E (i.e. an effect measure that is both universally applicable and comprehensively capturing all effects of interest), which according to standard theory should ideally reflect (under conventional assumptions: individual) preferences for health states – i.e., some kind of a *value* function that is based on actual choices – , many analysts using CEA prefer the quality-adjusted life year (QALY) as an outcome metric [26,45]. The QALY combines, by means of multiplication,

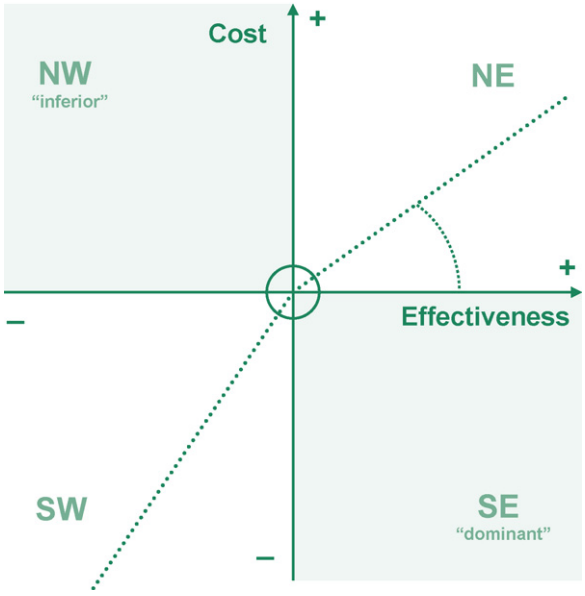

Fig. 1 The cost effectiveness (CE) plane.
A simple graphical representation of the relationship between incremental costs and effects of an intervention compared to its alternative (usually standard treatment or a competing program), O. If the intervention of interest is more costly and less effective than its comparator, O, it will be located in the NW quadrant of the CE plane and be considered "inferior", whereas in the opposite case it is said to "dominate" its alternative if it is located in the SE quadrant (i.e., more effective an less costly). In the NE and SW quadrants the choice will depend on the maximum cost effectiveness ratio (ICER, Fig. 4) the policy maker is willing to accept [41].

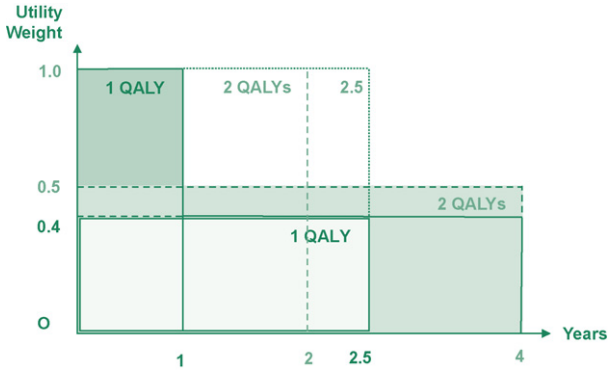

Fig. 2 The quality-adjusted life year (QALY) concept.
Length (horizontal axis) and quality (vertical axis) of life determine the number of quality-adjusted life years (QALYs). The quality (or utility) weights should be based on actual preferences and measured on a cardinal scale to enable a meaningful computation of sums and differences (cf. Figs. 3 and 4). For example, four life years spent in a health state with a utility of 0.5, such as blindness according to some studies, give 0.5 x 4 = 2 QALYs, equivalent to 2 years spent in full health. For a sequence of health states, the area under the curve (AUC) is the number of QALYs corresponding to this trajectory [26].

length of life with health-related quality of life in one single metric.[1] Quality of life is represented by an index, which is assumed to represent the expected utility of a given health state and can vary between 1 for "perfect health" and 0 for "dead" (Fig. 2).

[1] There is a host of practical / methodological considerations that are relevant to HALY measurements and their results, discussion of which is beyond the scope of the present paper. Interested readers may wish to consult the excellent guide by John Brazier et al (2007) [45].

This variant of CEA is sometimes referred to as **cost utility analysis (CUA)**, and (2.7) can then be written as follows:

$$\frac{\Delta C_1}{\Delta QALY_1} < \frac{\Delta C_2}{\Delta QALY_2} \qquad (3.1)$$

with the algorithm (3.2) below for computing the number of QALYs:

$$QALYs = \sum_{h=1}^{n} w_h \times t_h \qquad (3.2)$$

where w_h = quality weight (utility index), a preference-based measure reflecting the utility of health-related quality of life in a given health state h, and t_h = time (expressed as number of years) spent in that health state. It can be seen that the underlying assumptions (or, as has been argued from a theoretical perspective, *implications* [8,46]) include "additive separability" (i.e., the requirement that the utility of a given health state is unaffected by states that precede it or follow it) and a "constant proportional trade-off" (i.e., the proportion of remaining life that one would trade-off for a given quality improvement is independent of the amount of remaining life). – Perhaps unsurprisingly, empirical research indicates that both conditions may be violated [46].

When QALYs are calculated on the basis of average utility per year (in fact, any time period could be used) and discounting of future effects is factored in, this is equivalent to:

$$QALYs = \sum_{t=1}^{n} \frac{w_t}{(1+r)^{t-1}} \qquad (3.3)$$

with t = year, w_t = average health state utility during year t, and r = discount rate (cf. Fig. 3).

Rankings of interventions on the basis of their incremental cost per QALY gained (3.1), putatively reflecting an increasing social desirability with decreasing incremental cost effectiveness ratios (ICERs), are often referred to as cost effectiveness league tables

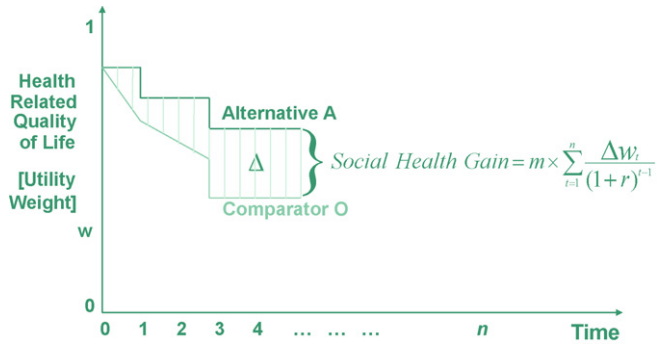

Fig. 3 The QALY aggregation algorithm.
Assuming discrete time intervals spent in any given health state, the average utility weight (reflecting the valuation of health-related quality of life experienced in that state) can be multiplied with the duration of the interval, expressed in years. Health states in the distant future will be valued less due to constant rate temporal discounting. Then, the QALY gain from replacing standard treatment, O, with an alternative, A, can be calculated by simple additive aggregation of the discounted gains in each interval. The social health gain from the decision to replace O by A will then be the product of the (average expected) individual health gain and the number, m, of individuals benefiting from the introduction of the new program, A [21,26,42,43].

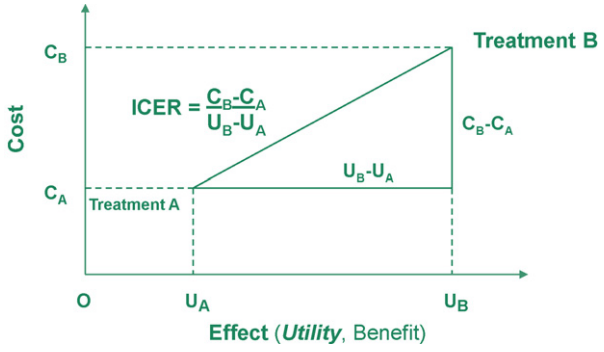

Fig. 4 Incremental cost effectiveness ratios (ICERs).
Results of cost effectiveness analyses (CEAs) and of cost utility analyses (CUAs) are usually reported as incremental cost effectiveness ratios, ICERs. This is intuitively appealing as efficiency can be interpreted as the ratio of inputs to outputs [48]. As a ratio of two absolute differences, the ICER possesses weird statistical properties, complicating probabilistic sensitivity analyses (capturing parameter uncertainty) and the computation of ICER confidence intervals. It also does not provide any information about the size of its numerator and its denominator and, therefore, the budgetary impact of adopting an intervention [6,26,43,54].

[43]. Note that this interpretation requires aggregate individual utilities (if QALY weights are determined from an individual perspective, which is most often the case, for example by applying the standard gamble or time trade-off techniques [45]) to map into social utility [47]. Further, then the validity of such rankings is directly linked to the implicit assumption of a constant, context-independent societal (or, on behalf of society, decision maker's) willingness to pay for a QALY, lambda (λ) [32]. (Alternatively, WTP for a QALY or λ can be interpreted as the shadow price of a QALY in a given health care system with a budget constraint [43].) According to the logic of cost effectiveness, the decision rule of whether or not to accept a new program reduces to:

$$ICER = \frac{C_A - C_B}{E_A - E_B} = \frac{\Delta Costs}{\Delta Effects} = \frac{\Delta Costs}{\Delta QALYs} < \lambda \qquad (3.4)$$

where ICER is the incremental cost effectiveness ratio, and λ represents the slope of the cut-off line on the cost effectiveness plane (Figs. 1 and 4).
It will be immediately evident to most readers that this decision rule depends upon the validity of the hypothesis that "a QALY is a QALY is a QALY – regardless of who gains and who loses it" [32,48-51], a position that in the literature sometimes has been labeled "QALY egalitarianism." This, of course, is directly linked to the existence of a (quasi-utilitarian) additive QALY aggregation function. It has indeed been claimed that the principal objective of a collectively financed health care scheme – in that particular case, of the National Health Service (NHS) in the United Kingdom – "ought to be to maximize the aggregate improvement in the health status of the whole community" [52], or to maximize "social health gain" – i.e., the number of QALYs produced – given a budget constraint:

$$Social\ Health\ Gain = m \times \sum_{t=1}^{n} \frac{\Delta w_t}{(1+r)^{t-1}} \quad \text{(for any given program)}, \qquad (3.5)$$

which can be written for multiple programs in different therapeutic areas:

$$Social\ Health\ Gain = \sum_{j=1}^{k} \left(m_j \times \sum_{t=1}^{n} \frac{\Delta w_{j,t}}{(1+r)^{t-1}} \right) \qquad (3.6)$$

with m_j = the number of beneficiaries of intervention j.

Cost minimization analysis

Finally, a fourth type of comparative health economic evaluation is **cost minimization analysis (CMA)**. In CMA, consequences other than costs play no part in the evaluation. Therefore, both effects and prices (valuation) disappear from equation (2.4), which thus becomes:

$$\frac{1}{C_1} > \frac{1}{C_2} \qquad (4.1)$$

which is equivalent to

$$C_1 < C_2 \qquad (4.2)$$

With very few exceptions, assuming (or establishing) the equivalence of outcomes is fraught with conceptual and empirical problems [53]. Thus this method is rarely used in practice, and reports of CMAs should be interpreted with particular caution only (cf. Table 1).

Some limitations and critique

Although currently representing the dominant paradigm for health economic evaluations, the logic of cost effectiveness is not uncontroversial. Imposing a constant WTP for a QALY marks a crucial departure from economic theory. Also the central role of the ICER has been seriously challenged by economists, not least for the ratio failing to provide decision makers with any information about the size of its numerator and denominator, and hence any useful information about the opportunity cost from the health scheme's perspective of adopting a new health care intervention [54]. From a policy makers' perspective, this implies a linearity assumption, with social utility being strictly proportional to the number of persons benefiting (as is the case in the standard utilitarian model, cf. equations (1.7) and (3.6) above).

Next, QALYs cannot be interpreted as an economic measure of health-related utility, unless one is prepared to impose a linear utility function instead of diminishing marginal utility over time. This is because the QALY represents a simple additive aggregation of utility-adjusted time intervals (cf. Fig. 3), which provides for analytical convenience at the expense of generalizability. An interesting proposal to improve how HALYs reflect actual individual preferences was made by Amiram Gafni and Abraham Mehrez (1989) [55]. Unlike QALYs, healthy year equivalents (HYEs) do not impose specific conditions on individual preference between length of life and quality of life and can be measured using the standard gamble technique [42,56]. HYEs (and subsequent extensions [57]) have gained little practical relevance, partly due to theoretical controversy surrounding their properties, partly due to more pragmatic concerns regarding measurement difficulties [42,43,58,59].

The potential impact of the analytical shortcut in the computa-

Table 1. Comparative economic evaluations: a typology.

Type of analysis	Measurement and valuation of costs	Measurement of consequences (effects)	Valuation of consequences (effects)	Theoretical foundation (standard)
CMA: cost minimization analysis	Monetary units (usually from a "decision maker's perspective")	None	None	Costing theory
CEA: cost effectiveness analysis	Monetary units (usually from a "decision maker's perspective")	Single effect measure of interest, common to alternatives evaluated, but achieved to different degrees	Natural units (e.g., life years gained, response rates, etc.)	Decision analysis and operations research; goal: technical efficiency
CUA: cost utility analysis	Monetary units (in theory, often recommended to be determined from a "societal perspective"; in practice, often from a "health care policy maker's perspective"	Single or multiple effects, not necessarily common to alternatives evaluated	Health-adjusted life years (usually QALYs)	"Extrawelfarism" – maximizing total health gains under a resource constraint; goal can be technical or allocative efficiency (usually applying a cost/QALY benchmark)
CBA: cost benefit analysis	Monetary units (from a "societal perspective", i.e., ignoring transfer payments)	Single or multiple effects, not necessarily common to alternatives evaluated	Monetary units (usually WTP)	Economic welfare theory – maximizing the impact of health care on overall well-being; goal: allocative efficiency

Similarities and differences of commonly used techniques for the comparative economic evaluation of health care programs [6,8,26,43,44,46].

tion of QALYs is illustrated by studies using shorter intervals such as quality-adjusted life months (QALMs) or even quality-adjusted life days, which have been converted to QALYs by simple multiplication with a factor of 12 (or 365, respectively) [60-62]. A practical example of the potential implications is the evaluation of an acute pain service [63], that (according to the logic of cost effectiveness, applying a cost per QALY cost effectiveness benchmark of £30,000; as adopted by NICE, cf. below) would have to be considered inefficient if its (total) marginal cost exceeded £164, even if it (hypothetically) completely eliminated postoperative pain "as severe as dead" over two full days (with a health state utility weight $w_0 = 0.0$ compared to $w_1 = 1.0$)[2] – irrespective of the fact that only a small number of patients would be affected, with the implication of small to moderate budgetary impact from a payers' perspective. As this and other examples [63-67] show, that reasoning has been be extended to quality adjusted life days – and by implication, might even be used to calculate quality adjusted life minutes (again, "QALMs"). This kind of arithmetic might well leave policy makers with *QALMs about* (the use of) *QALYs* (by extrawelfarists) [68].

Importantly, a growing number of empirical studies reveal a broad range of "contextual" factors impacting on WTP for QALY gains. Many of these factors are related to the distinction between preferences people have about their own lives ("self-regarding" or "personal" preferences, which are the focus of traditional welfare economics) and those about other people's lives ("other-regarding" or "external" or "social" preferences [69]). Collectively these studies cast doubt on the validity of the assumption that health gain maximization (or, for that matter, maximization of the utility derived hereof) is indeed the overriding social objective [21,22,70,71], thus undermining the purely efficiency-focused perspective of health economic evaluations [22]. (The extrawelfarist approach also cannot solve the ethical challenge of how losers might be compensated, which was briefly indicated earlier in the context of the Kaldorian hypothetical compensation tests of welfare economics.) Well-documented "contextual" factors include (but are not limited to) severity of the initial health state (which is of course not identical with the improvement achieved by an intervention), the patient's potential to benefit from an intervention (i.e., no discrimination against people in double-jeopardy, such as the permanently disabled and the chronically ill), the number of patients afflicted with a given disorder (i.e., the number potentially sharing the benefit), parent and/or caregiver status, the "rule of rescue" (i.e., the imperative to help [visible?] people in urgent need of intervention), and even the very role of costs [72] (for review and discussion, see Dolan et al., 2005 [71], and Richardson and McKie, 2007 [22]).

Analysts soon realized that, in the absence of some standardization, the inevitable variety of evaluation approaches would greatly decrease the policy value of economic analyses. Hence health economists [26,33] as well as policy makers [33,34] attempted to develop consensus statements on methods [33,35] (cf. Table 2). These conventions include the use of preference-based analysis, although it is known that people often underpredict their potential to adapt to poor health states [73]. Thus in many cases survey results from community samples give lower utility weights compared to patient self-reports for chronic health problems. Accordingly, the old debate about whose values should count, and in particular the question of whether *ex ante* decision utility or *ex post* experienced utility should enter meaningful economic evaluation, finds an echo in extrawelfarism, with potentially far-reaching normative implications [24,74-76].

Although the methodological and technical issues associated with practical cost utility analyses are not subject of the present paper, one salient methodological choice should be mentioned. Health care expenditures are highly concentrated among a relatively small number of patients, in whom coexistence of multiple disorders is rather the rule than the exception [77-80]. Comorbidity will inevitably reduce the potential to benefit from an intervention, i.e., the maximum quality of life (utility) weight that can be achieved; or conversely, multiple conditions require multiple simultaneous treatments to attain an improved health status, with implications for resource utilization and (again) incremental cost utility ratios. For example, Melissa Brown and colleagues (2005, pp. 163ff.) [81] defend the *ad hoc* convention that comorbidities should be ignored in CUAs on three grounds, (1) to avoid discrimination of the disabled (cf. above), (2) because of violation of the Americans with Disabilities Act of 1990, and (3) because "an almost infinite number of cost-utility analyses [would be] required for just one intervention." It is probably for the third reason given (again: analytical convenience or, rather, feasibility) that this convention reflects prevalent practice. Since actual clinical decision-making (like any mea-

[2] Because an incremental cost of £165 for two days of complete pain relief (in an otherwise healthy patient) would translate into an ICER of £165/[(1.0-0.0) × (2/365) QALY] = £30,112.50/QALY gained. If other conditions than pain after surgery were taken into account, too, this ICER would necessarily deteriorate further.

Table 2. International guidelines for economic evaluation.

Issue	Washington Panel reference case	NICE reference case	Methodological guidelines
Problem definition	The Panel's framing recommendations are kept separate from its reference case definition	Scope from NICE	Usually expected to define indication, patient (sub)groups, comparator, and perspective
Comparator(s)	Existing practice; if not cost-effective, consider a (a) best available, (b) viable low cost, or (c) "do nothing" alternative	Alternative therapies routinely used within the NHS; will be defined in the scope developed by NICE and will require definition and justification	Usually common practice ("f"); however, somewhat vague ("existing practice", "common practice")
Evidence on outcomes	Data should be selected from the best designed (and least biased) sources that are relevant to the question and population under study	Systematic review, with a preference for quantitative meta-analysis of randomized clinical trials data	Usually (long-term) effectiveness, not efficacy; with a broadly prevailing preference for data from randomized clinical trials
Economic evaluation	Cost-effectiveness analysis (CEA)	Cost-effectiveness analysis (CEA)	Usually cost-effectiveness analysis (CEA); sometimes more flexible (including cost-minimization and cost-benefit analysis, CBA)
Perspective on outcomes	All health effects, encompassing the range of groups of people affected, over a time horizon long enough to capture all relevant future effects	All direct health effects on individuals, whether patients or others (principally caregivers); time horizon should be sufficiently long to reflect any differences between the technologies being compared	Usually all relevant health outcomes
Perspective on costs	Societal perspective, long-term using opportunity cost; excluding indirect (productivity) costs; perspective should be explicitly identified	National Health Service (NHS) and personal social services (PSS)	Heterogeneous; direct health care costs only or direct and indirect (productivity) costs ("f"); societal perspective requested more often in informal guidelines ("i")
Discount rate	A real, riskless discount rate of 3.0% should be used, complemented by sensitivity analysis (drawn from 0% to 7%, including 5%)	An annual rate of 3.5% p.a. on both costs and health effects	Often 5% discount rate ("f"); heterogeneous recommendations from 2.5% to 10% in informal guidelines ("i")
Addressing uncertainty	Univariate sensitivity analysis as a minimum; multivariate sensitivity analyses recommended	Probabilistic sensitivity analysis mandatory (or, where appropriate, stochastic analysis of patient-level data)	Sensitivity analysis
Measure of health benefits	Quality-adjusted life years (QALYs)	Quality-adjusted life years (QALYs)	Usually including QALYs, with more flexibility as to other measures ("f", "i"), especially physical units; sometimes willingness to pay
Source of preference data for calculation of utility weights	Community preferences; if unavailable, patient preferences may be used as an approximation	Representative sample of the public (UK)	If QALYs are used, usually community preferences
Health state valuation method	Quality weights must be preference-based and interval-scaled	Choice-based method (for example, time trade-off or standard gamble; not rating scale)	If QALYs are used, usually choice-based method; often standard gamble and time-trade off; sometimes rating scales (!)
Description of health states for calculating QALYs	A generic classification scheme, or one that is capable of being compared to a generic system	Using a standardized and validated generic instrument	Heterogeneous; sometimes disease-specific instruments allowed ("f")

Table 2 (Continued).

Issue	Washington Panel reference case	NICE reference case	Methodological guidelines
Equity position	Discussion of roles and limitations of CEA in Introductory Chapter (separate from reference case definition)	Each additional QALY has equal value	n.a.
Budget impact analysis	n.a.	Impact on NHS not part of the decision-making process; however, required to allow effective national and local financial planning	Usually n.a.; Ontario: products with high budget impact will need more rigorous documentation of cost-effectiveness

For comparison, methodological guidelines may be informal ("i"; usually academic) or formalized ("f"; issued by official bodies such as HTA or pricing and reimbursement agencies) [26,33-35].

ningful measure of social welfare) is concerned with patients (*not* programs) and the benefits conferred to (*not* by) them, this convention must detach the results of CUAs from reality. Many utility gains calculated on this basis will never accrue in real life, highlighting the technocratic nature of current practice [82] that can hardly be realigned with a consistent theory of heath economic evaluation [43,83]. This anomaly is firmly embedded in the rhetoric that decisions at the program level were morally different from bedside decisions [43]; hereby neglecting that, necessarily, any prioritization decision, irrespective of the "level" on which it was originally made, will ultimately arrive at the bedside, and here affect concrete patients. Highlighting this issue, German philosopher Weyma Luebbe [84] provocatively asked, "Have you ever sat at the deathbed of a statistical life?"

Historically, development of the cost-effectiveness framework in the 1970s was heavily influenced by decision analysts with operations research backgrounds, who were striving to transfer methods used to optimise the efficiency of manufacturing processes to the production of health [85]. While these analysts had initially been puzzled by the apparent absence of an objective function for the health care system, they quickly turned to the assumption that the objective of health care was to maximize the quantity and quality of life, using (expected) utility theory to measure the preferences for health outcomes [48,49,85,86]. As George Torrance (2006, p. 1071 [85]) noted, the underlying "axioms of decision sciences are certainly designed to be prescriptive. So, for them, the appropriate test would be whether or not decision makers wish their decisions to be consistent with the axioms." A serious test of that hypothesis was however never undertaken by these scholars; instead they simply *asserted* that their choices had been the right ones, i.e., that the societal value of health care should be proportional to the number of patients benefiting from a program and the absolute increase in utility they obtain, and (due to discounting) a little less than proportional to the duration of the utility gain, with the QALY as an appropriate metric to capture the gains (Fig. 3) [43,48,49,85,86].

This author (MS) emphasized elsewhere that this assumption marks a striking contrast to the historic social roots of medical care [20], stated (official) objectives of policy makers, payers, and providers of health care (virtually all appealing to some notion of need, solidarity, sharing, and caring for others), as well as empirical studies of social preferences discussed earlier (cf. above) [20]. Today, the fundamental premise of cost utility analysis, i.e., that societies – and policy makers acting on their behalf – expect health care to maximize QALYs, must be considered as empirically falsified [22,71], and by implication this eliminates the possibility of a context-independent value of a QALY [87]. Concordant with this conclusion, simplistic cost per QALY rankings have failed to pass tests of reflective equilibrium, as exemplified by the low cost per QALY (viz., high "efficiency," and derived from this, putative degree of social desirability [44,88]) of sildenafil treatment for men with erectile dysfunction [89].

An interesting *post hoc* rationalization of the extrawelfarist proposition was offered by Werner Brouwer and Anthony Culyer ([30], who claimed that an "especially influential seed" helping shape the framework had been Amartya Sen's seminal work on functioning and capability [37,39]. It was however not before the mid to late 1980s that specific reference was made by health economists to Sen's work [29,90], and the asserted link seems difficult to substantiate as

cost-utility analysis is narrowly concerned with an optimization of some health outcome, whereas Sen's concept is much broader in scope, focusing not only on functioning but especially on the extent to which a person is *able* to function in certain ways, irrespective of whether the person chooses to do so [39]. Also the list of ten "central human capabilities" proposed by Martha Nussbaum [91] exceeded far beyond bodily health and quality of life. While the use of a measure of health gain in cost utility analysis may be seen as related to Sen's approach, its use as the single maximand certainly represents an overly restrictive, reductionist interpretation of Sen's ideas [92].

Finally leaving the realm of empirical observations, it should be mentioned here that the (quasi) utilitarian perspective underlying health economic evaluations is controversial from a normative perspective as well [37,69,93,94].

Practical application

Still a sizable number of health economists, especially those who are heavily engaged in the evaluation of health care programs, seem to regard the logic of cost effectiveness (using cost per QALY benchmarks) as the best currently available tool for determining the "value for money" provided by medical interventions. However, given the brief discussion of some limitations and critique above, it can hardly surprise that actual implementation varies greatly across jurisdictions.

Health Technology Assessments (HTAs)

Agencies concerned with health technology assessments (HTAs) and/or pricing and reimbursement decisions in many countries, notably including Australia, Canada, and in Europe the Nordic countries, the Netherlands, and the United Kingdom (UK), have embraced cost effectiveness analysis and the use of QALYs [95,96]. The National Institute for Health and Clinical Excellence (NICE) probably represents the most prominent example for this. NICE was established in 1999 as a Special Health Authority within the UK NHS. Its mission includes the development of guidance on the use of new and existing medicines and treatments in the NHS of England and Wales, on grounds of their clinical and cost effectiveness. Unlike many other agencies, NICE in effect *insists* on the use of QALYs, hence CUA [95,97], and has adopted a benchmark in the range of £ 20,000 to £ 30,000 per QALY gained to determine acceptability of a technology in the NHS [50,97]. The cost effectiveness benchmark used by NICE has been subject to debate, and it has been proposed that NICE should be seeking (more) systematically for an appropriate threshold rather than merely assuming one [98-101]. In response to controversies surrounding certain negative cost effectiveness evaluations, NICE subsequently introduced some exceptions from this benchmark for "end of life QALYs" [102] and "ultra-orphan" treatments [103,104].

Yet some other agencies like the *Haute Autorité de Santé* (HAS) in France and the Institute for Quality and Efficiency in Health Care (*Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen*, IQWiG) in Germany – both having become effective as of 2004, both supporting the mission of promoting efficiency in health care, and both having established health economics departments – have adopted a very skeptical position towards the use

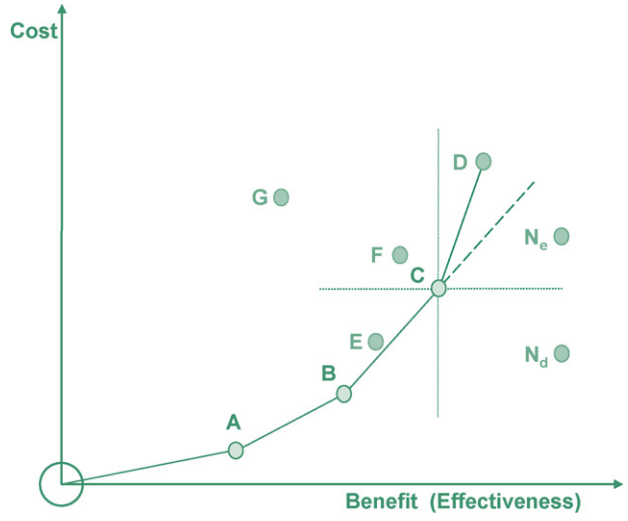

Fig. 5 Efficiency frontier analysis.
Germany's IQWiG intends to use efficiency frontier analysis to evaluate the cost effectiveness of new interventions, N [108]. Essentially, efficiency frontier analysis reflects the standard decision rules of cost effectiveness analysis. For example, new interventions F and G are dominated by alternatives C and B, respectively, since the latter provide more benefit for less cost. Intervention E is dominated in an extended sense, because a (linear) combination of interventions B and C would provide more benefit for less cost. A new intervention N_d would dominate C and, therefore, be considered cost effective.
The normative interpretation of an extrapolation beyond intervention C (i.e., into the NE quadrant of the CE plane, cf. dashed line and Figure 1), as suggested by IQWiG, has sparked controversial debate. IQWiG, insisting on within-indication comparisons only, proposes to accept a new intervention N_e because of extended dominance over C, whereas a new intervention D would be considered less efficient and hence be rejected.

of QALYs and have rejected the use of cost per QALY benchmarks [105-107]. In an attempt to escape from contentious interpersonal trade-offs, which are inevitably associated with comparisons across therapeutic areas that use a common yardstick to measure ''benefits'' (irrespective of the metric chosen, be it QALYs or WTP), IQWiG proposed the concept of ''efficiency frontier'' analysis (see Fig. 5) [108]. Clearly this approach is not at all without its own problems, as IQWiG's legal mandate is to make recommendations as to appropriate (maximum) pricing of pharmaceuticals (for reimbursement by the German Statutory Health Insurance), and this would have to be anticipated to frequently require extrapolation of the efficiency frontier (Fig. 5). The debate about the appropriate scope of comparative effectiveness research (CER) in the United States, and in particular the role of economic evaluation in CER, provides for another example of international heterogeneity in the formal use of health economic evaluations [109].

At last, the critique presented here is by no means meant to imply that current evaluations are worthless. The dimensions captured are certainly relevant ones. Rather, the critique is centered on the overly narrow focus of currently prevailing paradigms, in particular the highly restrictive approach called cost utility analysis, using cost per QALY benchmarks. Key underlying assumptions are probably wrong and may lead to potentially very misleading recommendations. The suggested implication for policy makers is to remain pragmatic and withstand demands of a thinly disguised normative nature for ''consistent'' or ''rational'' decision making as defined by the current framework. Conversely, health economists will need to pay relatively more attention to the further development of promising alternative paradigms [6,21,22,70,92,110-114], as opposed to the application of algorithms that are empirically questionable and grounded in poor theory.

Declaration of Interest

The author declares that no conflict of interest exists according to the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals* (effective October 2004).

Interessenkonflikte

Der Autor erklärt, daß kein Interessenkonflikt gemäß der *Uniform Requirements for Manuscripts Submitted to Biomedical Journals* (Stand Oktober 2004) besteht.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.zefq.2010.03.012.