Sveriges Kommuner och Landsting

Konsensusmetoder inom hälso- och sjukvård

En kunskapsöversikt

Förord

Denna kunskapsöversikt har tagits fram på uppdrag av Sveriges Kommuner och Landsting inom ramen för projektet Nationella medicinska indikationer. Finansieringen skedde via SKL:s FoU-råd som årligen beviljar medel till kunskapsöversikter som bedöms vara intressanta och viktiga för huvudmännen.

Allt arbete inom hälso- och sjukvården ska vara grundat på vetenskap och beprövad erfarenhet. Detta gäller vid varje beslut om åtgärd, exempelvis att skicka remiss, utfärda recept, besluta om röntgenundersökning eller operation. Under senare år har det funnits ett stort intresse för att sammanställa resultat från klinisk forskning rörande olika åtgärder och evidensbaserad vård har allt mer kommit i fokus. En betydande andel av de åtgärder som utförs inom vården har dock aldrig varit föremål för någon vetenskaplig utvärdering. I avsaknad av sådana utvärderingar har man endast att förlita sig på beprövad erfarenhet, det vill säga kunskap som delas av många men som inte verifierats i vetenskapliga utvärderingar. En svårighet är då att på ett systematiskt sätt kunna fånga det som kan anses utgöra den beprövade erfarenheten. För detta krävs någon form av formell konsensusmetodik.

Syftet med denna kunskapsöversikt är att ge en beskrivning av litteraturen om olika typer av konsensusmetodik som använts inom hälso- och sjukvårdsområdet samt bedöma metodernas användbarhet för att fånga och mäta beprövad erfarenhet. Kunskapsöversikten ska även tjäna som utgångspunkt för att vidareutveckla en lämplig metodik för tillämpning i det fortsatta arbetet med nationella medicinska indikationer.

Kunskapsöversikten har genomförts och författats av docent Ingemar Bohlin, Avdelningen för teknik- och vetenskapsstudier vid Sociologiska institutionen, Göteborgs Universitet.

Sveriges Kommuner och Landsting, oktober 2009

Göran Stiernstedt
Avdelningschef
Avdelningen för vård och omsorg

Sveriges Kommuner och Landsting
118 82 Stockholm, *Besök* Hornsgatan 20
Tfn 08-452 70 00, *Fax* 08-452 70 50
info@skl.se, www.skl.se

Grafisk form och produktion SKL FS Grafisk Produktion

ISBN 978-91-7164-486-2

Innehåll

Inledning

Erfarenhetsbaserad kunskap kontra vetenskap

Diametralt olika hållningar intas inom modern medicin till omdömen som bygger på erfarenhet av kliniskt arbete. Klinisk erfarenhet och experters åsikter brukar placeras i botten av så kallade evidenshierarkier. "Consensus opinion of experts" har till exempel placerats på nivå tio på en elvagradig skala där endast anekdotiska vittnesberättelser placeras lägre.[1] Samma hållning uttrycks tydligt av Edward Huth, tidigare redaktör för Annals of Internal Medicine, i en formulering om "medicine's long road from 'experience and expertise' to 'evidence' as the justification for particular medical treatments."[2] Denna hållning till erfarenhetsbaserad kunskap utgör kanske det mest radikala draget i evidensbaserad medicin (EBM), och ger ständigt upphov till debatt. Förespråkare av en motsatt hållning framhåller värdet av kliniskt omdöme när extern evidens, det vill säga resultaten av vetenskapliga studier, ska tillämpas på individuella patienter.[3] Än större, hävdas det ofta, är behovet av klinisk erfarenhet på områden som är bristfälligt täckta i litteraturen, och dessutom tyder en hel del på att även steget från stora och välgjorda studier till rekommendationer kräver omdömesbaserade avvägningar.

Om erfarenhet och omdömesförmåga är av en sådan betydelse finns all anledning att samla in kunskap av detta slag. Situationen är densamma i många branscher, och den vanligaste åtgärden från beslutsfattares sida är ofta att inrätta expertkommittéer eller så kallade ad hoc-paneler med en rådgivande funktion. I Sverige har Forskningsberedningen tjänat ett sådant syfte; Genteknikнämnden och regeringen Bildts IT-kommission, som båda inrättades i mitten av 1990-talet, är andra

1 Rinchuse et al. 2008, 168.
2 Huth 2008.
3 Det evidensbaserade konceptets upphovsmän har gjort allvarliga försök att beakta denna kritik. Se Lambert 2006, 2636–37.

exempel. Även om ledamöterna i sådana kommittéer är djupt kunniga är det väl känt att arbetet kan präglas av grupprocesser som inte gynnar resultatens tillförlitlighet. Därför har sedan ett halvsekel så kallade formella konsensusmetoder utvecklats på olika håll, huvudsakligen i USA. Så här har den grundläggande logiken i sådana metoder formulerats:

> *Formal consensus development methods… are widely used because, unlike informal methods such as committees, they offer structured, transparent and replicable ways of synthesising individual judgments.*[4]

Med det genomslag evidensbaserade metoder har fått inom hälso- och sjukvården har farhågorna för att konsensusmetoder ger subjektiva, otillförlitliga resultat likväl ökat. Åtminstone vad gäller deras användning inom detta område är två tendenser därför tydliga: dels en tilltagande formalisering av metoderna, det vill säga en stark strävan efter rigorösa, standardiserade procedurer, dels ansträngningar för att på ett mer systematiskt sätt föra in evidens i processen.

Syftet med kunskapsöversikten

Syftet med denna rapport är att ge en översikt av konsensusmetoder som har använts inom hälso- och sjukvård, med fokus på en särskild typ av tillämpning: hur klinisk erfarenhet kan samlas in för att fastställa indikationer när vetenskapliga data saknas eller är otillräckliga. Den primära frågeställningen är alltså vilka metoder som står till buds för detta ändamål och hur de är beskaffade. Mer specifikt: Vilka syften är metoderna utformade för att tjäna, vilka är deras förtjänster och svagheter och vilka begränsningar bör i första hand beaktas när de tillämpas?

Fyra huvudsakliga metoder

Viktigast bland de formella konsensusmetoder som har använts inom hälso- och sjukvård är Delfimetoden, den nominella gruppens teknik, konsensuskonferenser och "the RAND Appropriateness Method". Beslutet att bland existerande konsensusmetoder välja ut dessa fyra har inte varit svårt. De bästa översikter av konsensusmetoder som har identifierats inom ramen för denna kunskapsöversikt skiljer sig

4 Raine et al. 2004, 429.

visserligen åt i betoning och perspektiv, men i frågan om vilka metoder som är centrala överlappar bedömningarna i hög grad.[5]

Samma metoder som tillämpas i utarbetandet av detaljerade indikationer för kliniska interventioner används i stor utsträckning även när riktlinjer av den typ som utfärdas av specialistsällskap och nationella organ, i Sverige exempelvis Socialstyrelsen, formuleras. Indikationer av det slag som kommer att behandlas i rapporten utgör helt enkelt en särskild form av kliniska riktlinjer, givet en bred definition av begreppet. Den definition som oftast citeras har en sådan bredd. Riktlinjer för klinisk praktik är, enligt denna definition,

> *systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances.*[6]

Relationen mellan riktlinjer och detaljerade indikationer, och mellan de respektive metoder som används för att utarbeta dem, faller i huvudsak utanför ämnet för föreliggande översikt. På ett par ställen kommer den ändå att beröras.

Tillvägagångssätt vid genomgången

Den framställning som följer är till stora delar kronologiskt uppbyggd. Presentationen av olika metoder följer en historisk utvecklingslinje, fram till avsnittet om "the appropriateness method", som delvis är tematiskt. Denna metod ges betydligt större utrymme än övriga, eftersom den förefaller lämpligast för den uppgift som uppdraget har gällt. Det historiska angreppssättet gör logiken i introducerandet av olika metoder tydligare, och ger en bakgrund till nutida diskussioner som kan vara värdefull, i och med att många frågeställningar som idag är centrala har formulerats på ett tidigt stadium. En avsevärd metodologisk utveckling har ägt rum sedan de tidigaste försöken att utveckla formella konsensusmetoder gjordes, men flera problem är så besvärliga att de gjort sig gällande under lång tid; inga "quick fixes" har varit möjliga, utan man är tvungen att hantera svårigheterna på sätt som varken är invändningsfria eller slutgiltiga.

5 De bästa översikterna är Fink et al 1984; Jones & Hunter 1995; Murphy et al. 1998; Wortman 2004; och Black 2006.
6 Field & Lohr 1992, 27.

De metoder som kommer att behandlas är alltså formella, men har utformats i syfte att samla in kunskap av informell karaktär. En översikt av formella metoder kan i sin tur utföras med hjälp av metoder som inte är formaliserade. Författarna till den bredast anlagda översikt av konsensusmetoder som har identifierats i denna kunskapsöversikt, påpekar att de har valt en narrativ ansats, det vill säga att arbetet inte har utförts i enlighet med de procedurer som systematiska översikter följer.[7] Inte heller föreliggande rapport motsvarar kraven på en systematisk översikt. Förvisso har bibliografiska databaser av olika slag använts för att söka upp relevant litteratur, men det har inte funnits möjlighet att göra någon uttömmande sökning. Med den omfattande litteratur som finns inom området så skulle detta ha krävt genomgång av ett stort antal ytterligare referenser, vilket tidsramarna för uppdraget inte medgav. I arbetet med översikten har jag i stället litat till en kompetens som förvisso förutsätter vetenskaplig skolning men som inte kan formaliseras, och därför skulle kunna beskrivas som hantverksmässig.

Flera personer har bidragit med värdefulla synpunkter i inledningen av arbetet, och några har dessutom kommenterat manus. Jag vill därför framföra ett tack till Mats Eliasson, till mina uppdragsgivare Helene Törnqvist och Gunnar Moa och till Britt Nordlander, Nina Rehnqvist, Christina Kärvinge och Ingemar Månsson.

7 Murphy et al. 1998, iii.

Delfimetoden

Experters åsikter insamlade via enkäter

Delfimetoden kännetecknas framför allt av att de paneler av experter som väljs ut och ombeds medverka aldrig samlas till möten. De interagerar inte heller på annat vis inbördes, och får inte ens tillgång till varandras identitet. Panelens medlemmar besvarar enkäter som när metoden var ung sändes ut via post, men som nu brukar distribueras via epost eller nätbaserade plattformar. Proceduren upprepas alltid minst en gång, ibland i en serie omgångar. Inför varje ny insamling av enkätsvar får varje medlem av panelen tillgång till en sammanställning av de svar panelen tidigare givit, så att relationen mellan de genomsnittliga svarsvärdena och hans eller hennes egna framgår, och det förekommer att ledarna dessutom distribuerar annan information. Enkelt uttryckt avviker metoden från opinionsundersökningar genom sitt iterativa moment och från ad hoc-paneler, liksom möten och debatter av andra slag, i det att medlemmarna inte interagerar inbördes. Under processens gång konvergerar panelens svar i de allra flesta fall i riktning mot homogena värden – en konsensus. Begränsningen för hur långt detta konsensusförfarande kan drivas är av praktisk och social art; den avgörande frågan brukar vara hur många gånger man kan be en samling experter besvara samma enkät.

Historik

Metoden utvecklades under 1950-talet vid RAND Corporation i Kalifornien, en tankesmedja vars verksamhet till stor del finansierades med militära medel. En central fråga för dem som under denna fas av det kalla kriget fattade beslut av relevans för USA:s militära strategi var vilka mål Sovjetunionen skulle välja vid en eventuell kärnvapenattack mot USA, och hur många atombomber som skulle krävas för att slå ut dessa mål. Detta var ämnet för den första studie i vilken Delfimetoden tillämpades.[8] I början av 1950-talet upprättade Norman Dalkey och

8 Bakgrunden beskrivs till exempel i Linstone 1978 och Landeta 2006.

Olaf Helmer en panel bestående av sju militära experter, som fick besvara en serie av fem enkäter. Efter att under ett årtionde ha varit belagd med sekretess publicerades studien år 1963. Studiens empiriska resultat kan i detta sammanhang lämnas därhän, men det är värt att nämna att Dalkey och Helmer mellan enkätomgångarna, förutom information om vilka svar som lämnats i tidigare omgångar, till panelen distribuerade material som valts ut för att undanröja missförstånd beträffande de faktiska förhållanden på vilka experternas bedömningar byggde.[9]

Kontrollerad interaktion

I sin presentation av den metod de tillämpat diskuterar Dalkey och Helmer centrala metodologiska överväganden. Bland annat artikulerar de skälen till att en av Delfimetodens grundläggande principer är att kommunikation panelens medlemmar emellan undviks. Den form av kontrollerad interaktion metoden skapar, skriver de,

> *represents a deliberate attempt to avoid the disadvantages associated with more conventional uses of experts, such as round-table discussions or other milder forms of confrontation with opposing views. The method employed in the experiment appears to be more conducive to independent thought on the part of the experts and to aid them in the gradual formation of a considered opinion. Direct confrontation, on the other hand, all too often induces the hasty formulation of preconceived notions, an inclination to close one's mind to novel ideas, a tendency to defend a stand once taken or, alternatively and sometimes alternately, a predisposition to be swayed by persuasively stated opinions of others.*[10]

Och i artikelns avslutning formuleras vad som framstår som det grundläggande skälet till att Delfimetoden fyller en funktion. Givet att metodologin vidareutvecklas, skriver Dalkey och Helmer,

> *it may be hoped that a carefully contrived opinion consensus would often turn out to be an acceptable substitute for direct empirical evidence when the latter is unavailable.*[11]

9 Dalkey & Helmer 1963, 458–59.
10 Ibid., 459.
11 Ibid., 467.

Några år efter att metoden prövats för första gången utvecklade Olaf Helmer och filosofen Nicholas Rescher detta resonemang i en gemensam artikel. I denna artikel, som publicerades år 1959, presenteras en välartikulerad uppfattning om de omständigheter under vilka experters bedömningar utgör det bästa underlaget för beslut. Det centrala begreppet i Helmer och Reschers analys är formalisering. Detta begrepp används inte alltid i resonemang om behovet av konsensusmetoder, och hänvisningar till Helmer och Rescher är ovanliga, men deras sätt att förstå expertis och vetenskap utgör fortfarande grunden för tillämpning och utveckling av denna typ av metoder. Redan i denna artikel används för övrigt termerna konsensusprocedur och konsensusmetod.[12]

Konsensus bland utövare av inexakta vetenskaper

Ämnet för artikeln är de "inexakta" vetenskapernas epistemologi.[13] Distinktionen mellan exakta och inexakta vetenskaper syftar på graden av formalisering av de metoder som används för att samla in och analysera data. Även discipliner i vilka formaliserade metoder inte tillämpas alls eller endast fyller en mycket begränsad funktion kan betraktas som vetenskaper, förutsatt att de syftar till att ge förklaringar och förutsägelser och att den kunskap som produceras är intersubjektiv, det vill säga delas av ett kollektiv av kompetenta utövare.[14] Evidens som genereras i exakta vetenskaper utgör ofta en god grund för politiska, affärsmässiga och andra beslut, men det var i de procedurer som kan göra det möjligt att basera beslut på resultat från inexakta discipliner som Helmer och Rescher såg en potential för förbättring. Till stöd för sitt resonemang använde de bland annat exempel från arkitektur och ingenjörsvetenskap, men också från medicin. För att kunna bedöma en patients tillstånd behöver läkare naturligtvis uppgifter om puls, blodtryck och annat som kan mätas, hävdade de, men dessutom måste de kunna bedöma en rad omständigheter kring det enskilda fallet som inte låter sig mätas på något entydigt sätt. Endast den som kan utöva "informal expert judgment", det vill säga besitter det icke-formaliserade omdöme som är kännetecknande för en expert, kan göra adekvat bruk av sådan bakgrundsinformation.[15] Logiken i att utveckla metoder för att etablera konsensus bland utövare av inexakta vetenskaper är att sådan expertis ofta är den bästa kunskap som finns att tillgå.

12 Helmer & Rescher 1959, 46–47.
13 Begreppet epistemologi är nära besläktat med kunskapsteori. Det syftar, enkelt uttryckt, på vetenskapsteoretiska frågor om hur kunskap om världen konstrueras.
14 Helmer & Rescher 1959, 25.
15 Ibid., 40.

Ett instrument för förutsägelser

Delfimetoden är uppkallad efter oraklet i antikens Delfi, som ansågs ha förmågan att förutse framtida skeenden. Betoningen på förutsägelser är tydlig i pionjärstudien om målen för sovjetiska kärnbombsattacker, men även i Helmer och Reschers artikel, i vilken det hävdas att det stöd som experter kan erbjuda beslutsfattare nästan undantagslöst består av förutsägelser av framtida skeenden. Grunden för deras resonemang var följaktligen övertygelsen att beslutsfattares prediktiva instrument kan förbättras avsevärt.[16]

Metoden vann stor spridning under 1960- och 1970-talet, och användes i stor omfattning för detta syfte. "Technological forecasting", förutsägelser om vilka av de nya teknologier som alltid finns på ritbordet som kommer att få genomslag på marknaden, har varit ett vanligt tillämpningsområde. Metoden har använts inom framtidsforskning, i större företag, inom samhällsvetenskap och andra områden.

16 Helmer & Rescher 1959. 41–42.

Nominella grupper

Typer av interaktion

I slutet av 1960-talet utformade André Delbecq och Andrew Van de Ven, båda verksamma vid the Graduate School of Business, University of Wisconsin, en ny metod för beslutsfattande i grupp. I en serie artiklar publicerade under 1970-talets första år presenterade Delbecq, som var professor i management, och Van de Ven, som vid denna tid var doktorand, sin metod. Som beteckning för metodens centrala, mest innovativa komponent valde de termen "nominal group technique". Nominella grupper är liksom Delfimetoden en teknik med vars hjälp interaktion mellan individer struktureras för att de negativa aspekterna av spontana, oreglerade grupprocesser ska kunna undvikas. I en artikel som publicerades år 1971 summerade Delbecq och Van de Ven existerande forskning om beslutsfattande i grupp.

Kommittéer av olika slag fattar vanligtvis beslut under tämligen ostrukturerade former, i vad Delbecq och Van de Ven kallade "interagerande grupper". Kreativiteten i sådana grupper begränsas, liksom kvaliteten i de beslut som fattas, av en rad faktorer. Deltagare med lägre status än övriga medlemmar hämmas ofta i ostrukturerade muntliga överläggningar, dominanta personer kan i otillbörlig grad sätta sin prägel på utbytet av idéer, och diskussionerna kan i långa perioder fastna i ett enda spår. Av dessa och liknande skäl förblir resurser som var och en av gruppens medlemmar har tillgång till outnyttjade. Om deltagarna i stället utför en del av sitt arbete individuellt, och om bestämda procedurer för att samla in och värdera dessa individuella bidrag följs, kan avsevärt bättre resultat uppnås.

Termen nominell grupp syftar på en grupp vars medlemmar var och en, sittande vid samma bord, skriver ned synpunkter och idéer. Socialpsykologiska experiment hade enligt Delbecq och Van de Ven visat att den tysta aktivitet som under kollektivt arbete av detta slag omger varje individ skapar en kreativ spänning som har en gynnsam

inverkan på deltagarnas prestationer. Även den dynamik som uppstår i interagerande grupper fyller dock en funktion, hävdade Delbecq och Van de Ven. Olika typer av mötesteknik, föreslog de, bör därför användas i skilda faser av kommittéers arbete. När problem ska inventeras och möjliga lösningar identifieras är nominella grupper överlägsna interagerande. För att syntetisera och värdera den information som framkommer, och för att gradvis bygga upp konsensus, bör man däremot använda sig av interagerande grupper. När gruppen slutligen ska ta ställning till de förslag som på detta sätt har vuxit fram sker det på nytt lämpligast i nominella grupper.[17]

En modell med fem faser

Innan Delbecq och Van de Ven presenterade sin metod i publikationer hade de provat ut den såväl i privata företag som i politiska organ och i utbildningssektorn. De kunde därför ge detaljerade anvisningar för dess tillämpande.[18] Den nominella gruppens teknik utgjorde en viktig komponent i en modell för utveckling av åtgärdsprogram i organisationer som Delbecq och Van de Ven kallade the Program Planning Model. Modellen var indelad i fem faser. I de två första faserna identifierades problem respektive möjliga åtgärder i nominella grupper. I senare faser prioriterades problem och åtgärder, varpå ett program utformades och utvärderades. Allt detta skedde i interagerande grupper, men på basis av information som genererats i processens två första faser. För att inventera problem i den aktuella organisationens verksamhet förde man i den inledande fasen samman ett representativt urval av konsumenter eller klienter. I en påföljande fas sammankallade man en "knowledge resource panel", en grupp experter som kunde redogöra för de vägar att komma till rätta med existerande problem som stod till buds.[19] I båda fallen samlades information in i nominella grupper med 6–9 medlemmar. Efter att individuellt ha listat problem respektive tänkbara initiativ rapporterade varje gruppmedlem vid sittande bord sina punkter till en moderator som antecknade dem på en tavla eller ett stort block. I en allmän diskussion gavs gruppen så tillfälle att ventilera de punkter som förts fram, att revidera dem och att lägga till nya. Därpå genomfördes ett omröstningsförfarande i vilket varje deltagare enskilt och anonymt skrev ned de punkter på den kollektiva listan som han eller hon ansåg vara viktigast.

17 Van de Ven & Delbecq 1971.
18 I detta sammanhang använde de faktiskt termen "guideline"; se Van de Ven & Delbecq 1971, 210.
19 Delbecq & Van de Ven 1971; för den citerade termen, se s. 478.

Värdet av strukturerad interaktion ansikte mot ansikte

The Program Planning Model kombinerar alltså nominella grupper med interagerande i en noggrant strukturerad process. Några år senare publicerade Delbecq och Van de Ven en studie i vilken deras egen teknik för kollektivt beslutsfattande experimentellt jämförs dels med Delfimetoden, dels med konventionella ostrukturerade gruppdiskussioner. I detta experiment genererades med Delfimetoden och NGT såväl fler som bättre idéer än konventionella grupper. Resultaten gav dessutom vid handen att nominella gruppers beslutsfattande håller högre kvalitet än det Delfiformatet ger upphov till. De negativa aspekter av Delfimetoden som Delbecq och Van de Ven pekade på hängde alla samman med att interaktion ansikte mot ansikte inte ingår i processen. Avsaknaden av fysiska möten innebär att idéer inte kan förklaras eller modifieras i ljuset av invändningar eller frågor. Enkätsvaren gör det möjligt att beräkna majoritetens prioriteringar och bedömningar, men inga konflikter löses, och därmed blir de ståndpunkter som intas mindre väl avvägda än de kunde ha varit.[20]

I sin översikt av forskning om beslutsfattande i grupp hade Van de Ven och Delbecq nämnt att tidigare studier påvisat samma förhållande, nämligen att

> *Final choices made by groups after interaction are better than decisions based simply on the statistical pooling of individual judgments.*[21]

Annorlunda uttryckt ger Delbecq och Van de Vens metod, tack vare att den inbegriper ett element av interaktion, förutsättningar för rationellt beslutsfattande som Delfimetoden saknar.

Den nominella gruppens teknik har i sin ursprungliga form använts i en del studier av hälsorelaterade frågor.[22] Dess betydelse i framväxten av konsensusmetoder inom det biomedicinska området ligger dock främst i att format som introducerats senare delvis bygger på den metod som Delbecq och Van de Ven utvecklade.

20 Van de Ven & Delbecq 1974, 619.
21 Van de Ven & Delbecq 1971, 209.
22 Van De Ven & Delbecq 1972, Horn & Williamson 1977 och Trivedi 1982 är några exempel. Gallagher, Hares, Spencer et al. 1993 är en programmatisk artikel som uppmanar till mer omfattande bruk av NGT inom hälsoområdet.

Konsensuskonferenser

Såväl Delfimetoden som den nominella gruppens teknik har alltså utvecklats inom andra fält, men har länge tillämpats även inom medicin. Konsensuskonferenser, däremot, utformades specifikt för att hantera problem inom hälso- och sjukvården. Formatet utvecklades liksom Delfimetoden och nominella grupper i USA, några år efter att den senare tekniken hade lanserats.

Problemet med geografisk variation

Under 1970-talet hade avsevärda lokala variationer inom sjukvården dokumenterats, särskilt i USA. Epidemiologen John Wennberg intog en ledande roll bland forskare som visade att de medicinska ingrepp som en och samma diagnos föranledde i olika städer, regioner och delstater kunde variera dramatiskt. Andelen kvinnor som vid 70 års ålder genomgått hysterektomi visade sig till exempel uppgå till 70 procent i ett område i delstaten Maine men bara 20 procent i ett annat, och medan 8 procent av barnen i en del av Vermont hade fått sina tonsiller avlägsnade hade ingreppet i en annan del av samma stat utförts på 70 procent av barnen. Läkarprofessionens status var vid denna tid mycket hög, och medicinska bedömningar antogs allmänt vila stadigt på vetenskaplig grund. Påvisandet av omfattande geografisk variation i medicinsk praktik ledde till att detta antagande ifrågasattes. Den slutsats Wennberg drog var att läkares bedömningar av många tillstånd lika mycket byggde på subjektiva åsikter som på vetenskap. För att begränsa vad han kallade "the practice style factor" krävdes dokumentation av utfallet av olika ingrepp; den typ av forskning Wennberg gick i bräschen för brukar kallas "outcomes research".[23]

23 Wennberg 1984, 1989.

Fenomenet geografisk variation gav inte bara upphov till farhågor beträffande tillförlitligheten i och den vetenskapliga grunden för läkares beslut att ordinera läkemedel och utföra kirurgiska ingrepp. Fenomenets ekonomiska konsekvenser var dessutom besvärande för sjukvårdens huvudmän. Kostnaden per capita för sjukhusvistelser kunde till exempel skilja sig kraftigt åt mellan två städer vid New Englands kust med var sitt framstående akademiskt sjukhus och befolkningar som i de flesta avseenden var mycket lika. Sannolikheten för att en medborgare i Boston skulle läggas in på sjukhus, och därmed kostnaden för stadens sjukvård, var nästan dubbelt så hög som i New Haven.[24] Under samma period som Wennberg och andra forskare i studie efter studie belade en omfattande geografisk variation i medicinsk praktik blev det allt tydligare att kostnaderna för det amerikanska sjukvårdssystemet ökade på ett okontrollerat sätt. USA:s sjukvård hade under hela efterkrigstiden expanderat kraftigt, och bara från mitten av 1960- till mitten av 1980-talet nästan tredubblades den andel av landets BNP som sektorn tog i anspråk.[25] Behovet av politiska initiativ var uppenbart.

Health Technology Assessment

Under 1960-talet förändrades hållningen till vetenskap och teknik i samhället. Användningen av vetenskapsbaserad teknik började under denna period att framstå som problematisk i många sammanhang, och i början av 1970-talet talades det mycket om behovet av att utvärdera ny och existerande teknologi. 1972 inrättades i USA ett särskilt organ, Office of Technology Assessment, vars uppgift var att förse kongressen med information om olika teknologiers säkerhet och effektivitet. Hälso- och sjukvård var ett av de områden som tidigt uppmärksammades, och år 1976 föreslogs från flera håll, bland annat av senator Ted Kennedy, att medicinska innovationer borde bli föremål för samma typ av utvärdering som the Office of Technology Assessment ägnade andra sektorer.[26] Från denna tid har utvärdering av medicinsk teknologi vuxit fram som ett viktigt fält. Termerna "technology assessment of medical technologies" och "medical technology assessment" förekommer, men Health Technology Assessment (HTA) har blivit det etablerade namnet för detta fält. Medicinsk teknologi, eller health technology, definieras i dessa sammanhang mycket brett. Så här lyder en av de explicita definitioner som har givits:

24 Wennberg 1989, 78.
25 Relman 1988, 1221.
26 Perry 1987, 485-86.

> *Health care technologies are drugs, devices, medical and surgical procedures, and the organisational and support systems within which health care is delivered.*[27]

HTA-traditionen är alltjämt stark, trots att EBM i de flesta sammanhang har nått en mycket högre grad av synlighet. De två strömningarna är nära besläktade, vilket kan illustreras av att Statens Beredning för Medicinsk Utvärdering (SBU), som idag allmänt uppfattas som EBM:s högborg i Sverige, grundades som ett svenskt organ för HTA.[28]

National Institutes of Healths roll

Givet att the National Institutes of Health (NIH) har stått som garant för kvaliteten i lejonparten av USA:s offentligt finansierade medicinska forskning var det inte underligt att organisationen förväntades ansvara för den typ av utvärdering som efterlystes. Som nytillträdd chef för NIH i mitten av 1970-talet tog Donald S. Fredrickson initiativet till en ny mekanism för utvärdering av medicinsk teknologi, brett definierad. I september 1977 hölls en första så kallad "consensus development conference", vars ämne var screening av bröstcancer, och året därpå grundades the Office of Medical Applications of Research (OMAR), ett organ med uppgiften att organisera konferenser av samma slag. Konsensuskonferenser, som formatet har kommit att kallas, är en av de viktigaste metoder som har utvecklats inom HTA-traditionen.[29] Konferensernas syfte formulerades ursprungligen i termer av teknik-överföring; när Fredrickson 1976 presenterade formatet för en senats-kommitté betonade han sålunda behovet av att påskynda

> *the transfer of new technology across the 'interface' between biomedical research and the health care community and systems…*[30]

Senare har syftet ofta formulerats utan hänvisning till föreställningen om tekniköverföring, men innebörden tycks i allt väsentligt vara oförändrad. Så här heter det i en god översikt av konferensformatets framväxt:

27 Jörgensen 1995, 27, n. 1.
28 Hult, under utgivning. År 1987, när Beredningen inrättades, hade uttrycket evidensbaserad medicin ännu inte myntats. Myndighetens engelska namn är fortfarande The Swedish Council on Technology Assessment in Health Care.
29 I slutet av 1980-talet beskrevs NIH:s program för konsensuskonferenser till exempel som "the most visible and influential medical technology assessment activity in the United States." Wortman, Vinokur & Sechrest 1988, 495.
30 Mullan & Jacoby 1985, 1068.

> *OMAR's goal was to improve the translation of the results of biomedical research into information that could effectively be employed in the practice of medicine and public health care.*[31]

Formulerat på detta sätt sammanfaller det syfte NIH:s konsensuskonferenser var avsett att tjäna med kärnan i EBM, ett koncept som lanserades 15 år senare.[32]

En form av offentlig hearing

Konsensuskonferenser är en form av offentlig hearing. Formatet har beskrivits som en kombination av tre olika modeller för konfliktlösning. Dels liknar konferenserna rättegångar där expertvittnen framträder inför en jury, dels har de vissa drag gemensamma med vetenskapliga konferenser där forskare presenterar resultat av nya studier och besvarar frågor från kolleger, och dels har de karaktären av stormöten eller bystämmor där varje deltagare har rätt att göra sin röst hörd.[33] OMAR och de enskilda institut vid NIH som delade ansvaret för varje arrangemang följde under de första åren inte något strikt enhetligt format, men år 1982, när ett trettiotal konsensuskonferenser hade genomförts, utarbetades en uppsättning riktlinjer i syfte att formalisera processen. En central roll i det format som så småningom etablerades intas av panelen, som brukar bestå av 10–15 personer representerande medicinsk grundforskning, klinisk erfarenhet, biostatistik eller liknande metodologisk kompetens samt allmänheten.[34] Konferenserna brukar pågå två och en halv dag, ofta inför en publik på hundratals personer. Fram till den andra dagens eftermiddag presenterar talare/experter (speakers) aktuell forskning om de interventioner eller procedurer som konferensen avhandlar, och panel och publik ges tillfälle att ställa frågor och framföra synpunkter. Därefter sammanträder panelen i avskildhet för att formulera ett konsensusdokument innehållande rekommendationer beträffande hur eller i

31 Jörgensen 1995, 17. För en snarlik formulering, se Jacoby 1990, 9.

32 Den översättningsprocess som åsyftas i citatet ovan bör inte förväxlas med den för vilken beteckningen "translational research" på senare år blivit vanlig. De två formerna av översättning är helt olika, men saken kompliceras av att innebörden i begreppet translational research inte är entydig. Som regel avser begreppet vägen från laboratoriebaserad grundforskning till läkemedel och diagnostiska metoder, men det används även för att beteckna vägen från kliniska studier till medicinsk praktik. I denna senare innebörd sammanfaller begreppet sålunda både med målsättningen för NIH:s konsensuskonferenser och med den grundläggande idén i EBM. För en diskussion av de två innebörderna av begreppet translational research, se Woolf 2008.

33 Mullan & Jacoby 1985, 1068.

34 Mullan & Jacoby 1985, 1071.

vilken omfattning tekniken ifråga bör användas. Innan ett preliminärt dokument kan presenteras den tredje dagens morgon krävs som regel nattliga sessioner. En presskonferens brukar följa, och efter viss revidering distribueras och publiceras dokumentet.

Betydelsen av panelens sammansättning

Sannolikheten att konferenser av detta slag kan generera tillförlitliga sammanfattningar av existerande kunskap om olika medicinska innovationer betingas av en rad faktorer. Mycket tyder på att de svårigheter som i första hand måste hanteras gäller panelens sammansättning. Idealet är en panel vars medlemmar är tillräckligt väl insatta i de frågor som behandlas för att kunna ta in och kritiskt bedöma all information som presenteras, men utan att på förhand ha bundit sig för en viss hållning. Att de experter som lägger fram resultat av aktuell forskning tillhör bestämda skolbildningar och företräder partsintressen är naturligt nog; arrangörernas uppgift är att borga för att ett väl avvägt urval av ståndpunkter presenteras. Panelens medlemmar, däremot, bör inte vara alltför intimt knutna till olika intressegrupper. Två principer står här mot varandra. Å ena sidan har NIH eftersträvat ***neutrala*** paneler. Eftersom neutralitet ofta hänger samman med en brist på kunskap har man i andra fall försökt sätta samman paneler som representerat en ***balans*** mellan skilda intressen, helt i enlighet med det sätt på vilket medverkande experter har valts ut. Ju hårdare panelens medlemmar är knutna till partsintressen av olika slag, desto sämre är dock deras förutsättningar att nå konsensus.

Från ett tidigt stadium har NIH:s konsensuskonferenser satts ifråga ur ett flertal aspekter, men särskilt stark tycks den kritik ha varit som gjort gällande att panelernas sammansättning varit skev. I många fall har det inte bara varit så att det urval som gjorts inte framstått som representativt, utan det har dessutom ansetts tjäna dolda politiska syften. Kritiker har alltså uppfattat konferenserna som riggade för att ge ett visst utfall.[35] Enligt en utvärdering av NIH:s program för konsensuskonferenser, som påbörjades några år efter att konsensuskonferenserna hade sjösatts men publicerad först 1988, så reducerade den formalisering av processen som genomfördes några år in på 1980-talet problemets omfattning, men undanröjde det inte. Fortfarande, heter det i denna rapport, fördelas uppdragen i paneler och som experter via "'the good ol' boy' network", det vill säga på basis av

35 Wortman, Vinokur & Sechrest 1988, 490; jfr. Perry 1987, 486–87.

vänskaps- och kollegiala band, och dessutom i oproportionerlig utsträckning utifrån akademisk prestige.[36] Paul Wortman, Amiram Vinokur och Lee Sechrest, rapportens författare, efterlyste därför grundligare, mer formaliserade rutiner för urval av paneler, liksom av de experter som hörs och de frågor som fokuseras i förhandlingarna.[37]

Ett välgjort vetenskapligt underlag är en förutsättning

En lika väsentlig punkt i den kritik som har framförts har gällt konsensusdokumentens vetenskapliga underlag. I den granskning Wortman, Vinokur och Sechrest genomförde framkom att panelerna i de flesta fall inte på förhand hade försetts med material i vilket forskning av relevans för konferensernas frågeställningar sammanfattades. Många panelmedlemmar klagade över att material de på egen hand kunde söka upp var alltför komplicerat för att vara hanterbart, och att de sammanställningar som vanligen delades ut vid konferenserna blev tillgängligt alltför sent för att kunna användas.[38] I en översikt av NIH:s konferensprogram från 1995 bekräftas situationen:

> *In general, there appears to be very little systematic effort to survey the available literature and summarise the state of science on the topic under consideration.*

En naturlig följd av detta sakernas tillstånd är att

> *Very few consensus statements make reference to the scientific literature upon which findings are based.*[39]

NIH:s konsensuskonferenser har uppenbarligen fyllt en viktig funktion, men trovärdigheten i de rekommendationer som har utfärdats har undergrävts av de problem som här har redovisats. Att avsaknaden av systematiskt sammanställt vetenskapligt underlag för rekommendationerna sedan EBM:s genombrott har ställt konsensuskonferenser i en dålig dager behöver knappast påpekas. Behovet av systematiska metoder med vilka vetenskapliga studier som borde beaktas vid konferenserna kunde identifieras och syntetiseras påtalades dock redan på 1980-talet, innan gruppen kring David Sackett hade lanserat

36 Wortman, Vinokur & Sechrest 1988, 478–79, 492.
37 Delfimetoden borde kunna användas för detta syfte, föreslog de. Wortman, Vinokur & Sechrest 1988, 491.
38 Wortman, Vinokur & Sechrest 1988, 480.
39 Jörgensen 1995, 23.

EBM. Metaanalys, som senare har kommit att bli en av de centrala metoderna inom EBM, utvecklades just för att garantera tillförlitligheten i översikter av litteraturen på olika områden. Under 1980-talets andra hälft fick metaanalys sitt genomslag inom medicinsk forskning, och i Wortman, Vinokur och Sechrests rapport rekommenderas de som arrangerar konsensuskonferenser att ta denna teknik i bruk.[40] Samma uppmaning har, med eller utan explicit hänvisning till metaanalys, förts fram i många sammanhang.[41]

Konferensformatets fortsatta användning

Vid mitten av 1990-talet hade NIH genomfört mer än 100 konsensuskonferenser, och verksamheten har fortsatt därefter. Formatet har också spritt sig internationellt. Den första konsensuskonferens som ägde rum utanför USA hölls år 1982 på svensk mark, i regi av Medicinska Forskningsrådet och SPRI. Några år senare hade konsensuskonferenser även arrangerats i Danmark, Storbritannien och Holland.[42] Med undantag för Holland fick konferenserna i dessa länder en bredare inriktning än i USA. I en ovan citerad passage av NIH:s chef Donald Fredrickson formuleras ambitionen att överföra teknik från medicinsk forskning till sjukvården, över den gräns (interface) som åtskiljer dessa områden.[43] NIH:s program för konsensuskonferenser utformades redan från början i enlighet med denna tankefigur, så tillvida att man skilde mellan vad man kallade "teknisk konsensus" och "interface consensus". I den förra processen, som OMAR inrättades för att administrera, stod medicinska innovationers säkerhet och effektivitet i fokus. För den andra processen, som gällde ekonomiska, etiska, juridiska och sociala aspekter av teknikanvändning i vårdapparaten, skulle ett särskilt organ ansvara, som dock lades ned efter några få år.[44] Medan verksamheten vid NIH även fortsättningsvis varit inriktad på teknisk konsensus har konferensformatet på annat håll vidareutvecklats långt bortom HTA-traditionen. Det avgörande steget togs i Danmark under 1980-talets andra hälft, genom att panelerna helt och

40 Wortman, Vinokur & Sechrest 1988, 491.
41 Se till exempel Perry 1987, 487; Jacoby 1988; Breart 1990, 25; Jörgensen 1995, 23; Sauerland & Neugebauer 2000, 910.
42 Vang 1986, Johnsson 1988.
43 Se ovan, s.18.
44 Perry 1987, 486; Jacoby 1990, 9.

hållet sattes samman av lekmän. Denna danska modell av konsensuskonferenser har därefter övertagits av aktörer i ett stort antal länder, som i det modifierade formatet ser ett instrument för demokratisk insyn i och inflytande över politiska frågor med viktiga vetenskapliga och tekniska komponenter.[45]

45 Joss & Durant 1995; Guston 1999; Nielsen, Lassen & Sandøe 2007.

The Rand Appropriateness Method

I mitten av 1980-talet, ett par årtionden efter lanserandet av Delfimetoden, utvecklades en konsensusmetod i ett samarbete mellan forskare vid RAND Corporation och University of California, Los Angeles. Ledaren för samarbetet hette Robert Brook. I ett brett upplagt projekt inventerade han och hans kolleger litteraturen om karotisendarterektomi, coloskopi och fyra andra medicinska och kirurgiska procedurer. Målsättningen var att på basis av de randomiserade kontrollerade kliniska prövningar (RCTs) som publicerats kunna formulera detaljerade kriterier för de kliniska tillstånd vid vilka de olika interventionerna var motiverade eller lämpliga – "appropriate". På ett tidigt stadium i sitt arbete fann gruppen att publicerade RCTs inte gav tillräckligt stöd för lämplighetskriterier av detta slag. Gruppen valde därför ett alternativt tillvägagångssätt för att nå sin målsättning, nämligen att utarbeta en hårt strukturerad metod med vars hjälp experters åsikter om vilka indikationer som motiverar olika ingrepp kan syntetiseras. Den korta benämningen "the appropriateness method" förekommer, men metodens ursprung vid RAND Corporation och UCLA brukar anges. I en del sammanhang används förkortningen RAM, för "the RAND appropriateness method"; för enkelhetens skull kommer denna praxis att följas fortsättningsvis.

Ett grundläggande argument för konsensusmetoder

RAM beskrevs för första gången i en artikel som publicerades 1986, och dess upphovsmän har därefter behandlat den i ett stort antal publikationer. När metoden presenterades var det inte som ett alternativ till Delfitekniken, nominella grupper eller NIH:s konsensuskonferenser, och begreppet konsensusmetod nämns över huvud taget inte i 1986 års artikel.[46] Detta kan förefalla underligt, i synnerhet som samma forskargrupp ett par år tidigare hade publicerat en ofta citerad

46 Brook et al. 1986. I en liknande presentation tre år senare betecknas RAM på ett ställe som en konsensusmetod, dock utan att närmare förklaring ges. Chassin 1989, 25.

översikt av konsensusmetoder.[47] Relationen till existerande konsensusmetoder var dock av sekundär betydelse för gruppen kring Brook. Det primära var att rättfärdiga den nya metoden gentemot RCTs, som vid 1980-talets mitt hade etablerats som den studiedesign som ger mest tillförlitlig kunskap om medicinska interventioners effekter.

RCTs är enligt Brook och hans kolleger inte bara en resurs- och tidskrävande typ av studie, utan ger dessutom sällan den utförliga information om olika patientkategorier som är nödvändig för att detaljerade kriterier för interventioners lämplighet ska kunna formuleras. Väl genomförda kliniska prövningar ger ytterst värdefull kunskap om interventioners effekter, men sådana studier är inte tillräckliga. Problemet gäller inte bara de sex procedurer gruppen granskade, utan är helt allmänt. "We will never", skriver Mark Chassin, en av de seniora forskarna i gruppen, i en programmatisk artikel,

> *have enough rigorous scientific data for common medical and surgical procedures to rely solely on them as a source of information for appropriateness criteria.*[48]

En förutsättning för att kriterier för lämplighet ska kunna utarbetas är följaktligen att relevant kunskap kan utvinnas ur någon annan källa. För Brook och hans kolleger har det varit uppenbart att sådan kunskap existerar. Så här beskriver Chassin denna kunskapskälla, samt den utmaning konsensusmetoder står inför:

> *As clinicians, we extrapolate from existing research data on a daily basis when we treat patients. This is an exercise with which practising clinicians are very familiar. The problem for researchers is how to extract that knowledge in a systematic, reliable and valid manner.*[49]

Två punkter bör omedelbart noteras. För det första sammanfaller huvuddragen i resonemanget ovan på ett slående sätt med de skäl Dalkey, Helmer och Rescher angav för att Delfimetoden fyller en funktion. Som tidigare påpekats utformades denna metod för att göra det möjligt att på ett tillförlitligt sätt extrahera icke-formaliserad kunskap av det slag experter inom många branscher besitter. Ett av de exempel Helmer och Rescher använder i sin artikel om de inexakta

47 Fink et al. 1984.
48 Chassin 1989, 22.
49 Chassin 1989, 23.

vetenskapernas epistemologi gäller, som vi har sett, läkare som inhämtar information om mätbara variabler som blodtryck och puls, men som ständigt måste falla tillbaka på sitt kliniska omdöme. För det andra utgör inte heller för gruppen kring Brook klinisk erfarenhet en kunskapskälla isolerad från vetenskapliga data. Den kunskap som på ett systematiskt sätt behöver samlas in springer, som framgår av citatet från Chassin ovan, ur läkares förmåga att tillämpa vetenskaplig information på individuella patienter. Brook och hans kolleger är mycket tydliga med att de åsikter hos experter som RAM har utvecklats för att samla in inte bygger på erfarenhet *i stället för* publicerade RCTs. Åsikterna bygger på båda typerna av kunskap, flödar ur båda dessa källor.[50]

Tesen att den publicerade vetenskapliga litteraturen i sig inte förmår ge den vägledning som medicinsk praktik av hög kvalitet kräver har blivit ett standardargument för RAM. Metodens ledande förespråkare idag, Paul Shekelle, verksam vid Evidence-Based Practice Center vid RAND Health i Santa Monica, Kalifornien, har till exempel regelbundet använt sig av detta argument.[51] Tesen utgör dessutom ett av de mest grundläggande argumenten för konsensusmetoder helt allmänt. Den mest omfattande översikt av konsensusmetoder som hittills har publicerats inleds till exempel just med detta argument.[52] En del forskare har försökt beräkna den proportion av medicinsk praktik som kan stödjas med rigorösa studier av interventionernas effekter. Evidens av det slag RCTs ger saknas för mellan 50 och 90 procent av alla medicinska procedurer, hävdade en grupp författare häromåret. Endast mellan 15 och 20 procent av de procedurer som tillämpas kan, enligt en annan källa, rättfärdigas med rigorösa vetenskapliga data.[53]

Tillvägagångssätt

De paneler som används i RAM brukar uteslutande bestå av ansedda läkare. Det vanliga är att nio personer ingår i varje panel – enligt Chassin därför att detta visade sig vara det maximala antalet individer som på ett effektivt sätt kan delta i strukturerad interaktion i en ansikte mot ansikte-situation.[54] I de första studierna som genomfördes bad gruppen bakom RAM ledande figurer inom amerikansk medicin att

50 Se Brook et al. 1986, 54.
51 Shekelle et al. 1998, 1888; Shekelle 2004, 228; Shekelle 2009, 517.
52 Murphy et al. 1998, 1; hänvisning ges här till Chassins artikel från 1989.
53 Nicollier-Fahrni et al. 2003, 15; Shekelle et al. 1998, 1888. Flera publikationer anförs till stöd för vardera uppskattningen.
54 Chassin 1989, 24. Numera förekommer det att mellan sex och femton personer ingår i varje panel; Shekelle 2009, 517.

föreslå läkare som skulle kunna medverka, och senare har nomineringar av lämpliga medlemmar inhämtats från respekterade medicinska sällskap inom de områden som undersökts.[55] Man har försökt se till att alla specialiteter som patienter brukar komma i kontakt med i samband med att en given procedur blir aktuell är representerade i panelerna, och en huvudprincip har redan från början varit att olika intressen måste balanseras. Specialister som utför proceduren i fråga får till exempel inte utgöra mer än en minoritet, eftersom det är väl känt att de tenderar att överskatta dess värde. Förutom vad gäller medlemmarnas specialiteter brukar en balans eftersträvas med avseende på geografisk spridning och fördelning mellan akademisk och privat praktik.[56]

Till varje medlem i en panel sänds dels en översikt av litteraturen på området ifråga, dels en lista med kliniska scenarier. Om översikterna sägs inte mycket i publikationer om RAM, utöver att de fokuserar kliniska prövningar inom området ifråga, sammanställs av de forskare som organiserar processen och tjänar som underlag för de listor med scenarier som utarbetas. Dessa listor är avsedda att vara uttömmande, och brukar omfatta från några hundra upp till cirka 2000 kombinationer av förslag till indikationer. I de första studier som gruppen kring Brook genomförde specificerades till exempel 675 förslag till indikationer för karotisendarterektomi och 1086 scenarier för coloskopi. Endast om den population för vilken procedurerna kan bli aktuella indelas i ett så stort antal underkategorier, har metodens upphovsmän och förespråkare resonerat, kan varje kategori patienter bli tillräckligt homogen för att en given procedur ska kunna betraktas som lika lämplig eller olämplig för dem alla. Panelens medlemmar ombeds ge sin bedömning, på en niogradig poängskala, av hur lämplig proceduren ifråga är i varje specificerat kliniskt scenario. Så här lyder den ursprungliga, fullständiga definitionen av nyckelbegreppet "lämplig":

> *'Appropriate' was defined to mean that the expected health benefit (i.e., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity – not necessarily in order of importance) exceeded the expected negative consequences (i.e., mortality, morbidity, anxiety of anticipating the procedure, pain produced by the procedure, time lost from work) by a sufficiently wide margin that the procedure was worth doing.*[57]

55 Brook et al. 1986, 54; Shekelle et al. 1998, 1889.
56 Chassin 1989, 24; Shekelle et al. 1998, 1889; Shekelle 2004, 229.
57 Brook et al. 1986, 55. Definitionen har upprepats många gånger, men ges som regel mindre utförliga formuleringar.

Den högsta poängsatsen, 9, ges när proceduren anses vara ytterst lämplig, 5 betyder att lika mycket talar för som emot och poängsatsen 1 att proceduren är fullständigt olämplig för scenariet ifråga. Bedömningarna skall vara rent medicinska, det vill säga ekonomiska kostnader skall inte vägas in.

De som medverkar i panelerna brukar hinna klassa mellan 200 och 300 scenarier per timme. Efter att de har skickat in sina bedömningar samlas varje panel till ett möte, som vanligen pågår en eller två dagar. En sammanställning av de poäng som givits för varje scenario delas ut, och diskussioner under ledning av en moderator tar vid, med fokus på de scenarier som bedömts olika. Där oenighet uppstår brukar någon av följande tre situationer råda. För det första kan scenarierna ha formulerats på sätt som är oklara och öppnar för olika tolkningar, för det andra kan nya studier ha kommit till vissa medlemmars kännedom medan andra inte tagit del av dem, och för det tredje kan antingen medlemmarnas erfarenheter eller deras tolkningar av existerande studier divergera. När någon av de två första situationerna är för handen försöker moderator nå enighet i gruppen. I situationer av den tredje typen undviker man att pressa fram konsensus, eftersom även meningsskiljaktigheter ger processens arrangörer viktig information.[58] Benämningen för det konferensformat som NIH lanserade 1977 var som redan nämnts "consensus development conferences". Det format som Brook och hans kolleger utformade fyller även en annan funktion: RAM är en metod som dels mäter, dels utvecklar konsensus.[59] De mätningar som är av vikt görs efter att panelernas diskussioner har avslutats. Medlemmarna ombeds då bedöma scenarierna på nytt, och principer har utformats för tolkning av de poängsatser som summeras. Om samtliga medlemmar har tilldelat ett scenario mellan 7 och 9 poäng anses konsensus råda om att proceduren ifråga är lämplig vid detta scenario, och ligger samtligas poängsatser mellan 1 och 3 är panelen överens om att proceduren är olämplig. Ligger poängen mellan 4 och 6, eller om olika bedömningar gjorts, anses osäkerhet råda om huruvida proceduren är lämplig. Endast för omkring 5 procent av fallen brukar paneler ranka scenarier på ett sådant sätt.[60]

58 Shehelle 2004, 229; Shekelle 2009, 517; Brook et al. 1986, 62.

59 En klar distinktion mellan "consensus development" och "conensus measurement" ges i Jones & Hunter 1995, 376–77.

60 Chassin 1989, 25; Shekelle 2004, 229. Brook och hans kolleger har även laborerat med alternativa sätt att tolka panelernas bedömningar; Brook et al. 1986, 61–62.

Under panelernas sammanträden brukar en hel del uppmärksamhet ägnas specificerandet av kliniska scenarier. Scenarier som inte är entydiga spaltas upp i två eller flera, medan andra scenarier tvärt om förs samman till en enda. I de studier som först utfördes med RAM genomgick de ursprungliga katalogerna med scenarier omfattande förändringar innan mötena avslutades.[61] Den totala bilden av de undersökta procedurernas lämplighet eller värde blev därmed en helt annan. Från ett mycket tidigt stadium i formatets utveckling har det alltså stått klart att panelernas diskussioner fyller en viktig funktion i den konsensusprocess som metoden ger upphov till.

Två tydligt definierade syften

RAM utformades för att tjäna två tydligt definierade syften. De resultat som metoden genererar är för det första avsedda att fungera som beslutsstöd i klinisk praktik. Kriterier för lämplighet av det slag som RAM ger skiljer sig avsevärt från de riktlinjer för klinisk praktik som myndigheter idag utfärdar i stor skala, men lika uppenbart är att vägledning för kliniska beslut ges i båda fallen. Släktskapet framgår till exempel av att resultaten av en konsensusprocess strukturerad enligt RAM, som vanligtvis kallas "appropriateness ratings" eller "appropriateness criteria", även benämns "appropriateness guidelines".[62]

För det andra kan kriterier för olika procedurers lämplighet användas som instrument för utvärdering av klinisk praktik. Varje sammanfattning av kunskapsläget på ett givet område som kan tjäna som beslutsstöd, det vill säga vägleda beslut i realtid, kan även användas retrospektivt, för att granska hur välgrundade de beslut som fattats har varit. Att så är fallet noteras till exempel i den publikation ur vilken en auktoritativ auktion återgavs i inledningen av denna rapport. Omedelbart på denna definition följer kommentaren att riktlinjer mycket väl kan användas för kvalitetssäkring, likaväl som kriterier som utformats för utvärderingsändamål kan användas som riktlinjer för klinisk praktik, och att dokument av dessa två slag i vissa sammanhang är praktiskt taget identiska.[63] Vad RAM beträffar är det inte i efterhand, utan upphovsmännens kännedom eller avsikter, som de kriterier som genererats har använts för att utvärdera medicinsk praktik.

61 Brook et al. 1986, 58.
62 Se Brook 1994, 219, och Hicks 1994, 733.
63 Field & Lohr 1992, 27.

I publikationer av gruppen kring Brook har syftet att utvärdera kvaliteten i den sjukvård som ges tvärt om ofta varit primärt.[64] Metoden har presenterats som ett nytt redskap med vars hjälp den geografiska variationens problem skulle kunna lösas. För att metoden ska kunna tillämpas på detta sätt krävs ett ytterligare steg efter att kriterier för en given procedurs lämplighet har definierats. I metodens andra steg klassificeras ett stort antal patientjournaler med avseende på de indikationer på vilka beslut har fattats om att antingen utföra eller avstå från att utföra olika ingrepp. Genom att ställa dessa indikationer mot dem som definierats på basis av panelers rankingar kan andelen överanvändning respektive underanvändning och lämplig användning av proceduren ifråga bestämmas.

Det är dessutom värt att nämna att Brook och hans kolleger i den artikel i vilken metoden först presenterades även nämner ett tredje syfte. När konsensus inte kan nås i panelerna om huruvida en procedur är lämplig eller olämplig vid vissa indikationer visar detta att kunskapen om procedurens effekter är bristfällig. RAM kan alltså även användas för att peka ut ämnen som bör bli föremål för kliniska prövningar.[65]

Relationen till metodens föregångare

När studier utförda med RAM publiceras heter det ofta att en modifierad Delfimetod har tillämpats. Det format som gruppen kring Brook utvecklade på 1980-talet kombinerar dock komponenter från mer än en av de konsensusmetoder som existerade vid denna tid, och att den främst skulle bygga på Delfitekniken är alls inte uppenbart. Metoden har sålunda beskrivits som en modifierad form av den nominella gruppens teknik,[66] som en hybrid mellan Delfimetoden och NGT,[67] och till och med som en vidareutvecklad form av NIH:s konsensuskonferenser.[68] Den sistnämnda karakteristiken är knappast berättigad, men gott fog finns för de två förra.

64 Se till exempel Kahn et al. 1988; Winslow et al. 1988; och Brook 1994.
65 Brook et al. 1986, 63.
66 Murphy et al. 1998, 5.
67 Raine, Sanderson & Black 2005, 631.
68 Wortman, Smyth, Langenbrunner & Yeaton 1998, 110.

Såväl RAM som Delfimetoden föreskriver:

– att ett skriftligt material sänds ut till dem som valts ut för och accepterat att medverka i processen,
– att deltagarna ombeds avge och returnera kvantitativa omdömen,
– att de i ett senare skede får tillgång till sammanställningar av gruppens samlade bedömningar i vilka deras egna framgår,
– att de därefter ges tillfälle att revidera sina bedömningar,
– och att de som organiserar processen summerar deltagarnas bedömningar enligt bestämda kvantitativa rutiner.

Också i den typ av konsensusprocess som Delbecq och Van de Ven utvecklade bedömer deltagarna en situation enskilt vid två tillfällen, men den inledande omgången struktureras inte av detaljerade formulär, och processens resultat brukar inte kvantifieras. Den avgörande skillnaden mellan Delfimetoden och nominella grupper är dock en annan, nämligen att NGT är en mötesbaserad metod. Dess upphovsmän ansåg, som tidigare nämnts, att den möjlighet formatet ger deltagarna att förklara egna och ta del av andras ståndpunkter leder till bättre resultat än dem som uppnås med Delfitekniken. På denna centrala punkt har gruppen bakom RAM anslutit sig till Delbecq och Van de Vens hållning.

Relationen mellan RAM, NGT och Delfimetoden kan summeras i tabellform. I tabell 1 nedan ingår även NIH:s konsensuskonferenser samt ad hoc-paneler, det vill säga kommittéer som beslutsfattare inrättar för att inhämta experters åsikter utan att formella konsensusmetoder tillämpas.

	Ad hoc-paneler	Delfi	Nominella grupper	Konsensus-konferenser	RAM
Frågeformulär distribueras	Nej	Ja	Nej	Nej	Ja
Bedömningar görs enskilt	Nej	Ja	Ja	Nej	Ja
Deltagarna får ta del av sammanställningar av gruppens bedömningar	Nej	Ja	Ja	Nej	Ja
Interaktion ansikte mot ansikte	Ja	Nej	Ja	Ja	Ja
Formaliserad aggregering av resultat	Nej	Ja	Ja*	Nej	Ja

Tabell 1. Förenklad summering av metodernas centrala egenskaper.[69]

** Gemensamt för Delfimetoden, RAM och NGT, så som Delbecq och Van de Ven utformade den, är att processens slutresultat utgörs av sammanställningar av deltagarnas enskilda prioriteringar. Graden av formalisering varierar dock, i och med att nominella grupper inte brukar tillämpa kvantitativa procedurer för sammanställning av resultaten.*

Kritik

Att panelernas sammansättning är av avgörande betydelse för de kriterier för interventioners lämplighet som RAM ger har länge varit uppenbart, och som vi har sett formulerade metodens arkitekter på ett tidigt stadium normer för hur professionella intressen bör balanseras i grupperna. Redan i de ursprungliga studierna visade det sig till exempel att samtliga kirurgiska procedurer som ingick rankades avsevärt högre av de kirurger som medverkade än av panelernas övriga medlemmar.[70] Att specialister som själva utför ett ingrepp anser det vara mer användbart och värdefullt än andra läkare tycks idag vara allmänt accepterat. Tidigt intresserade sig Brook och hans kolleger dessutom för nationella variationer. I en studie som publicerades 1988 lät de en brittisk och en amerikansk panel bedöma indikationer för två kardiovaskulära procedurer, öppen kranskärlsoperation (CABG) och koronarangiografi. Bedömningarna gick vitt isär. Den brittiska panelens omdömen var genomgående mycket hårdare, och för koronarangiografi

69 En liknande tabell ges i Murphy et al. 1998, 3.
70 Park et al 1986, 770.

låg britternas medianvärde hela 2,35 poäng under amerikanernas. Tillämpade på en population av drygt 1600 patienter från 65 år och uppåt som genomgått proceduren i USA indikerade panelernas divergerande bedömningar att 74 respektive 39 procent av ingreppen varit lämpliga medan 17 respektive 42 procent varit olämpliga.[71]

Mot bakgrund av sådana siffror drogs, i en ledarartikel i JAMA samma år, slutsatsen att man helt enkelt måste räkna med att expertpaneler som sätts samman på olika vis kommer att ge varierande bedömningar av procedurers lämplighet. De resultat som Brook och hans grupp lagt fram angående andelen lämpliga respektive olämpliga ingrepp som görs inom sjukvården är viktiga, heter det, men oberoende test av de kriterier gruppen tillämpar efterlyses.[72] Sex år senare publicerades en betydligt mer utförlig diskussion av omständigheter som kan undergräva tillförlitligheten i de resultat RAM genererar. Nicholas Hicks, som författaren heter, hävdade att alltför många av de antaganden metoden bygger på är implicita. Han betonade bland annat att läkare som i identiska kliniska situationer sätter in identiska åtgärder kan fatta sina beslut av olika skäl, men att RAM inte gör avsedda utfall explicita. Vilka risker och fördelar med en given intervention som panelens medlemmar beaktat vid sina bedömningar framgår inte heller. Kriteriernas innebörd blir därmed oklar, varför de bör användas med försiktighet. Deras legitimitet undergrävs dessutom av att panelerna uteslutande består av medicinska experter; Hicks föreslog att även patienter och andra relevanta lekmannagrupper borde vara representerade. Vidare pekade han på att den roll moderatorer intar vid mötena kan vara av stor betydelse för förhandlingarnas utfall. Genomgående betonas i denna artikel alltså den osäkerhet som kriterier för lämplighet som utarbetats med hjälp av RAM är förknippade med. Hur systematisk metoden än är, heter det, är den i hög grad subjektiv.[73]

Givet dessa osäkerhetsfaktorer infinner sig frågan om huruvida en vetenskaplig grund existerar för de kriterier för lämplighet som Brook och hans kolleger arbetar fram. I vilken mån bygger de åsikter som deras metod syntetiserar på vetenskapliga studier av hög kvalitet? frågar Hicks. "The most sincerely held opinion of the most eminent physician", fortsätter han, "can still be wrong if it is not supported

71 Brook et al. 1988, 751–53.
72 Mulley & Eagle 1988.
73 Hicks 1994. David Naylor framförde några år senare en liknande serie reservationer beträffande tillförlitligheten i RAM; se Naylor 1998.

by science."[74] I en kritisk granskning ett år tidigare av metodologiska antaganden som RAM kan förmodas bygga på hade en annan forskare, Charles Phelps, intagit samma hållning. Metoder av denna typ, ansåg Phelps,

> *merely provide a refined way of recording conventional wisdom about the efficacy of medical therapies, wisdom that often stands without strong scientific support.*

Endast väl genomförda vetenskapliga studier, är budskapet, kan blotta och kullkasta etablerade villfarelser och ersätta dem med säker kunskap om medicinska interventioners effekter. Denna målsättning kan varken uppnås med RAM eller med andra konsensusmetoder:

> *Methods based on reaching a consensus among experts do not create new scientific data, they only codify old beliefs.*[75]

Nuvarande status

Inget tyder på att de forskare som utvecklat RAM har låtit sig nedslås av det faktum att metodologiska problem föreligger. År 1994, när flera viktiga invändningar mot metoden hade framförts, var Robert Brook tillräckligt övertygad om dess värde för att i en ledarartikel i British Medical Journal kraftfullt plädera för den.[76] I artikeln återger han, från en del av de studier hans grupp hade publicerat, resultat som tyder på att andelen ingrepp som utförs på olämpliga eller osäkra grunder på vissa håll har legat kring 50–60 procent. Metoder med vars hjälp lämplighet kan bestämmas existerar, framhåller Brook, och avslutar med en uppmaning till läkare att göra bruk av dessa metoder, så att problemet med över- och underanvändning av kliniska interventioner kan elimineras.

Ingen tvekan råder om att RAM är en avancerad metod av stort potentiellt värde, men lika tydligt är att den i flera avseenden är problematisk. En ledande roll bland dem som på senare år har bringat klarhet kring metodens egenskaper intas av Paul Shekelle, som 1998

74 Hicks 1994, 733.
75 Phelps 1993, 1244.
76 Brook 1994. Hicks kritik förelåg inte i tryck vid denna tidpunkt; den publicerades som en respons på Brooks ledare.

publicerade en viktig artikel med en grupp kolleger. Utgångspunkten för det resonemang som förs i artikeln är att debatten kring metodens styrka och begränsningar har varit teoretisk och byggt på åsikter. För att säkra slutsatser ska kunna dras om hur den fungerar krävs i stället empiriska studier.[77] Den grundläggande invändning som framförts av Hicks, Phelps och andra är att man med andra deltagare i panelerna skulle ha nått andra kriterier för lämplighet än dem som presenterats, och att metodens resultat därför inte är att lita på. Som en respons på denna kritik utförde Shekelle och hans kolleger därför ett experiment avsett att testa metodens reproducerbarhet.

Tre parallella paneler sattes samman för att utarbeta lämplighetskriterier för koronar revaskularisering och lika många för att bedöma lämpligheten av hysterektomi vid olika kliniska scenarier. Samma relativt detaljerade rutiner tillämpades vid urvalet av samtliga paneler. I vardera panelen för koronar revaskularisering ingick tre kardiologer som utför ballongvidgning av kranskärlen (PTCA), en kardiolog som inte utför detta ingrepp, tre kardiovaskulära kirurger samt två primärvårdsläkare. I vardera panelen för hysterektomi ingick fyra gynekologer som utför ingreppet, två gynekologer som inte utför det, och tre primärvårdsläkare. Samtliga deltagare hade nominerats av relevanta, respekterade specialistsällskap. De klassificerades med avseende på specialitet, geografisk region, kön samt huruvida deras praktik var privat eller akademisk, och ett stratifierat randomiserat urval genomfördes. Kriterierna för panelernas sammansättning var alltså i hög grad formaliserade. I övrigt följdes i huvudsak det för RAM gängse formatet: översikter av litteraturen på respektive område sändes ut till panelernas medlemmar, tillsammans med en katalog av kliniska scenarier (948 för koronar revaskularisering och 1718 för hysterktomi); efter att bedömningarna returnerats och sammanställts samlades varje panel för ett tvådagars möte; och när diskussionerna avslutats rankades scenarierna på nytt. Den enda avvikelsen bestod i att panelerna inte tilläts modifiera scenarierna, eftersom detta skulle ha begränsat möjligheterna att jämföra deras bedömningar. Avslutningsvis tillämpades de kriterier för lämplighet som processen lett fram till på data om några tusen patienter som bedömts för hjärtbesvär respektive genomgått hysterektomi.

Panelernas bedömningar sammanföll till 94–96 procent med avseende på överanvändning av koronar revaskularisering, till 92–93 procent med avseende på underanvändning av samma ingrepp, och till mellan

77 Shekelle et al. 1998, 1894.

70 och 88 procent med avseende på överanvändning av hysterektomi. Uttryckt som trevägs kappa låg överensstämmelsen på 0,52 respektive 0,83 och 0,51. Om än metoden långtifrån är perfekt, kommenterar Shekelle och hans kolleger, visar resultaten att dess reproducerbarhet ligger på samma nivå som den som gäller för ett antal accepterade diagnostiska tester.[78] I redogörelsen för experimentet pekas även några begränsningar i studien ut. Till dessa hör att de procedurer processens arrangörer följt för att syntetisera litteraturen på ett givet fält och för att utforma kataloger av indikationer, inte har undersökts. Värt att notera i sammanhanget är att författarna här använder begreppet "systematiska översikter".[79] Detta är en förpliktande beteckning, givet den status den har uppnått inom EBM, främst till följd av verksamheten vid the Cochrane Collaboration. De rutiner enligt vilka översikterna har genomförts beskrivs inte närmare här än i tidigare publikationer om RAM, undantaget upplysningen att de granskats av externa experter, men denna gång påtalar författarna alltså själva att detta är en brist.[80]

Metoders reproducerbarhet diskuteras ofta i termer av reliabilitet. Vad frågan i fallet RAM gäller är alltså huruvida konsensusprocesser organiserade enligt detta format är möjliga att formalisera i så hög grad att de ger samma resultat oavsett vilka personer som medverkar. Kan formaliseringen av metodologiska procedurer drivas så långt att samma resultat regelbundet reproduceras, inom det rimligas gränser, bör RAM accepteras som ett tillförlitligt instrument för denna typ av mätningar. Lika väsentlig är emellertid frågan om validitet, det vill säga huruvida metoden mäter det den är avsedd att mäta. Från ett tidigt stadium har randomiserade experiment föreslagits som det idealiska sättet att pröva metodens prediktiva validitet.[81] Ur en population av patienter för vilka interventioner som med hjälp av RAM definierats som lämpliga för deras kliniska tillstånd skulle en grupp slumpvis väljas ut för att genomgå interventionen medan en annan grupp inte skulle behandlas. Genom att följa upp experimentets utfall skulle man kunna bestämma lämplighetskriteriernas värde som beslutsstöd. Enighet har rått om att tillvägagångssättet vore problematiskt ur etisk synvinkel, och några studier av detta har inte identifierats.

78 Shekelle et al. 1998, 1893. Senare har Shekelle hävdat att ett kappa-värde över 0,8, en nivå som alltså överskreds med avseende på underanvändning av coronary revascularization i denna studie, brukar betecknas som en nära nog perfekt överensstämmelse; Shekelle 2009, 518.

79 Shekelle et al. 1998, 1888 och 1894.

80 Shekelle et al. 1998; upplysningen ifråga ges på s. 1889.

81 Mulley & Eagle 1988, 541; Chassin 1989, 27; Hicks 1994, 733.

Ett alternativt tillvägagångssätt är att forskare som inte är involverade i de kliniska besluten klassar en interventions lämplighet för en grupp patienter antingen dessa genomgår ingreppet eller inte, och därefter följer upp utfallet. En serie observationella studier av detta slag har genomförts. Bland annat publicerades förra året en studie ledd av Harry Hemingway i vilken drygt 9300 patienter ingick.[82] Två paneler utarbetade i denna studie var sin uppsättning kriterier för lämpligheten av koronarangiografi. Överensstämmelsen mellan kriterierna var bara medelhög (kappa 0,58), men detta har inte varit av större betydelse för studiens resultat. Diskrepansen var stor mellan behandlande läkares beslut och bedömningar, på basis av panelernas kriterier, av vilka patienter som var lämpliga kandidater för ingreppet. Underanvändningen av koronarangiografi fastställdes enligt panelernas kriterier till 54 procent respektive 71 procent. Oavsett vilken panels kriterier som tillämpades var dödligheten i den underbehandlade patientgruppen tre år senare avsevärt högre än i den som hade genomgått interventionen.

I en nyligen utgiven ledare i en kardiologisk tidskrift sammanfattar Shekelle resultaten av empiriska studier som genomförts för att pröva reliabiliteten och validiteten i RAM, däribland hans eget experiment från slutet av 90-talet och flera studier med resultat liknande dem som Hemingway publicerade förra året.[83] Shekelle reserverar sig mot observationella studiers begränsningar, och betonar att diagnostiska test alltid måste användas med försiktighet och urskiljning i klinisk praktik. Trots dessa förbehåll är hans slutsats att

> *the appropriateness criteria category for a particular patient probably represents the 'default' treatment option, and… should the cardiologist decide on a different course it is incumbent on the doctor to be explicit about why this choice is made – and to document it.*[84]

En liknande hållning intas i en ledare i Annals of Internal Medicine i vilken Hemingways ovan nämnda studie kommenteras. Författaren förklarar, efter att ha framfört flera reservationer, att lämplighetskriterier "have great promise, especially if integrated into practice

82 Hemingway et al. 2008. Paul Shekelle ingår i författarkollektivet bakom studien.

83 Shekelle 2009. Ytterligare studier av metodens reliabilitet och validitet summeras i Shekelle 2004, 229–30.

84 Shekelle 2009, 519–20. Elva år tidigare hade Shekelle med en grupp kolleger dragit försiktigare slutsatser; se Shekelle et al. 1998, 1893–94.

through the electronic medical record."[85] Lämplighetskriterier utarbetade med hjälp av RAM har under de senaste 20 åren använts i stor omfattning, både som beslutsstöd i klinisk praktik och som ett redskap för utvärdering, och en hel del forskning ägnas möjligheten att förbättra metoden. Även dess kritiker framhåller att metoden är avancerad och väl underbyggd. Nicholas Hicks kallar den till exempel "one of the leading tools for measuring appropriateness of care. " David Naylor beskriver RAM som "arguably the most respected approach to defining appropriate care," och dessutom som ovanlig i det att den utgör en "meticulously tested and systematic [method] for leavening limited evidence with expert opinion and inference."[86]

I Sverige har RAM tillämpats i åtminstone två studier. Ann Bengtsson och några kolleger vid Sahlgrenska i Göteborg använde sig i början av 1990-talet av en förenklad version av metoden för att undersöka lämpligheten av koronarangiografi och revaskularisering i en svensk population.[87] I en större studie, utförd av en grupp forskare bland vilka Bengt Brorsson, Lars Werkö och Robert Brook ingick, tillämpades några år senare lämplighetskriterier på svenska patienter som genomgått koronarangiografi.[88] I båda fallen har metoden alltså använts för att utvärdera vårdinsatser. Däremot har RAM, såvitt jag vet, ännu inte använts för att utarbeta beslutsstöd i Sverige.

85 Faxon 2008, 277.
86 Hicks 1994, 730; Naylor 1998, 1918, 1920.
87 Bengtsson et al. 1994.
88 Bernstein et al. 1999.

Konsensusmetoders funktioner

Sedan formella konsensusmetoder började användas inom hälsosektorn har metaanalys etablerats som en metod för syntes av vetenskapliga studier inom många fält. Tillförlitligheten av metaanalys har satts ifråga, särskilt när stora RCTs har gett avvikande resultat.[89] Teknikens förtjänster har dock vägt betydligt tyngre, och metaanalys är idag, liksom det generaliserade formatet systematiska översikter, en helt central metod inom EBM. En av effekterna av den stigande status metaanalys och systematiska översikter kommit att åtnjuta är att reservationerna gentemot konsensusmetoder har tilltagit. Användningen av konsensusmetoder ökar, men deras funktioner har förändrats.

En kort kommentar om relationen mellan metaanalys och konsensusmetoder är på sin plats här. Metaanalys har faktiskt beskrivits som en konsensusmetod, men denna karakteristik är missvisande.[90] Sant är att metaanalys liksom konsensusmetoder syntetiserar kunskap. Sant är också att en enkel typ av konsensusteknik som bland annat inom samhällsvetenskap är ganska vanlig även tillämpas inom metaanalys. Tekniken ifråga bygger på att minst två observatörer oberoende av varandra klassificerar eller extraherar data ur något material, och därefter jämkar samman sina bedömningar på de punkter där de går isär.[91] Skillnaden mellan metaanalys å ena sidan och Delfimetoden, nominella grupper och RAM å den andra är trots detta så stor att det inte finns något skäl att betrakta den förra tekniken som en konsensusmetod.

NIH:s konsensuskonferenser utformades för att summera tillgänglig kunskap om olika medicinska interventioner, och ett centralt kriterium

89 Se till exempel Higgins & Spiegelhalter 2002.

90 Wortman 2004, 2612.

91 Ett exempel ur litteraturen om konsensusmetoder ges i Nicollier-Fahrni et al. 2003, 16: “In case of disagreement between the two reviewers, a consensus decision was achieved after a re-reading of the study and a discussion.” För ett exempel inom metaanalys, se Leichsenring & Rabung 2008, 1553: “Disagreements were resolved by consensus.”

för urval av ämnen var att en god vetenskaplig bas existerade.[92] Ett utbrett antagande idag är att konsensusmetoder tvärtom kan fylla en funktion i lägen där god evidens *saknas*. Jag har tidigare kallat detta ett standardargument för RAM, och det är samtidigt ett grundläggande argument för konsensusmetoder över huvud taget. Argumentet ges ibland en form vars innebörd är att en avsevärd andel av klinisk praktik aldrig kommer att kunna ges ett robust vetenskapligt stöd. Det är dock betydligt vanligare i en annan form. Så här lyder den alternativa versionen i en relativt försiktig formulering:

> *Consensus methods may be more appropriate during the early or 'emerging' phase of the intervention when there are few high-quality research studies and a viable [research synthesis] cannot be conducted. As the intervention develops and sufficient research studies accumulate, research synthesis would seem more appropriate.*[93]

I denna version av argumentet antas alltså den brist på kliniska prövningar som föranleder användningen av andra metoder än metaanalys och systematiska översikter vara ett övergående problem. Den betydelse konsensusmetoder tillskrivs betingas givetvis av om det tillstånd som gör dem nödvändiga förmodas vara kortvarigt eller permanent. Gemensamt för argumentets båda versioner är dock att konsensusmetoder antas fylla en sekundär funktion i relation till evidensbaserade metoder: det är inom områden där den vetenskapliga basen är svag som metoder krävs för att man att på ett tillförlitligt sätt ska kunna syntetisera klinisk erfarenhet.

Som påpekades i rapportens inledning hänger genombrottet för evidensbaserade metoder samman med en djup skepsis mot erfarenhet och åsikter. Denna skepsis kommer till uttryck i en passage av Hicks som återgivits i ett tidigare avsnitt: "The most sincerely held opinion of the most eminent physician can still be wrong if it is not supported by science."[94] Att förmodat auktoritativa översikter av kunskapsläget på olika områden i flera fall inte beaktat nyare forskning påvisades år 1992, i en ofta citerad artikel som starkt bidragit till att skapa respekt för metaanalys som en metod för syntes av evidens.[95]

92 Mullan & Jacoby 1985, 1070; Jacoby 1988; Wortman et al 1988, 476.
93 Wortman, Smyth, Langenbrunner & Yeaton 1998, 120. "Research synthesis" är en samlande beteckning för metaanalys och systematiska översikter. En lika försiktig formulering ges i Naylor 1998, 1920.
94 Hicks 1994, 733.
95 Antman et al. 1992.

Den misstro mot experters åsikter som redan från början utgjort en central komponent i EBM kommer mycket klart till uttryck i Edward Huths formulering, återgiven i inledningen till denna rapport, om en lång marsch från erfarenhet och expertis till evidens.

En spänning existerar alltså inom det evidensbaserade konceptet. Å ena sidan görs stora ansträngningar att ersätta åsikter och erfarenhet med robust vetenskaplig kunskap. För dem som ger sådana ansträngningar sitt stöd framstår konsensusmetoder rimligen som kvarlevor från en historisk fas vi snarast bör lägga bakom oss. Å andra sidan anses klinisk erfarenhet på många håll utgöra en oundgänglig källa till kunskap. I den mån detta senare antagande accepteras kan ambitionen att på ett tillförlitligt sätt syntetisera kunskapen ifråga knappast ifrågasättas. De strävanden att formalisera konsensusmetoder som pågår är nära besläktade med de insatser som gjorts för att göra metaanalys, systematiska översikter och kliniska prövningar mer rigorösa. Den avgörande frågan gäller huruvida klinisk erfarenhet är en nödvändig kunskapskälla eller ej. Annorlunda uttryckt gäller den huruvida expertpanelers bedömningar, där vetenskapliga studier av hög kvalitet inte föreligger i tillräcklig omfattning, bör accepteras som den näst bästa formen av extern evidens.[96] Skilda hållningar i denna fråga ger upphov till en spänning inom den evidensbaserade rörelsen som, för det mer renläriga lägret, kan te sig som en klyfta mellan ett vetenskapligt och ett snarast förvetenskapligt ideal.

Rekommendationer beträffande RAM

Som tidigare framgått kritiserades NIH:s konsensuskonferenser redan från början för att lämna alltför stort utrymme för subjektiva och godtyckliga prioriteringar. Detta ledde, som vi har sett, bland annat till krav på rutiner avsedda att garantera att tillförlitliga sammanställningar av resultaten av relevant forskning gjordes tillgängliga för panelerna. Kravet har kommit att bli karakteristiskt för den evidensbaserade rörelse som vid denna tid började ta form, och har ställts på många områden. Som nämnts har till exempel kritiker av RAM på samma sätt ifrågasatt den vetenskapliga basen för de kriterier för kliniska interventioners lämplighet som utarbetas med denna metod.

Ståndpunkten att riktlinjer för klinisk praktik utfärdade av specialistsällskap och nationella organ måste bygga på evidens har likaså

96 Se Nicollier-Fahrni et al. 2003, 20.

framförts i många sammanhang.[97] När det framkommer att detta krav rutinmässigt nonchaleras betraktas det som mycket allvarligt. I en kritisk granskning från slutet av 1990-talet av 279 publicerade riktlinjer, i vilken ett flertal brister påvisades, framhölls sålunda i första hand att ett vetenskapligt underlag för rekommendationerna sällan hade utarbetats på ett systematiskt sätt. Granskningen visade bland annat att de metoder som använts för att identifiera evidens specificerades i mindre än 17 procent, och att formella metoder för syntes av data endast tillämpats i 7,5 procent, av detta urval av riktlinjer.[98] Nyligen gav en granskning av riktlinjer utfärdade av två amerikanska specialistsällskap för kardiologi ett liknande resultat. Vad som i detta fall ansågs oroväckande var att cirka hälften av rekommendationerna i gällande riktlinjer från sällskapen ifråga tillhör den lägsta av tre nivåer av evidens, det vill säga uteslutande bygger på experters åsikter, fallstudier eller vedertagna normer i vården. Ändå, kommenterar författarna, är fältet kardiologi väl försett med vetenskapliga studier. Om andelen evidensbaserade riktlinjer är så begränsad inom denna specialitet, hur illa är det ställt där tillgången till kliniska prövningar är långt mindre?[99]

Kriterier för lämplighet utarbetade med hjälp av RAM har blivit föremål för liknande studier. Anne Nicollier-Fahrni och några av hennes kolleger vid universitetet i Lausanne jämförde för några år sedan indikationer för coloskopi som utarbetats av en expertpanel med klinisk forskning på området. Studien är intressant i flera avseenden, och jag har redan hänvisat till den mer än en gång. Bland annat betonar författarna att den vetenskapliga litteraturen på många områden ger otillräckligt stöd för de kliniska beslut som måste fattas. De bedömer till exempel att endast cirka 5 procent av de möjliga indikationerna för coloskopi har behandlats i publicerade RCTs. Det var lämplighetskriterier för denna bråkdel av indikationer som det schweiziska forskarlaget ställde mot publicerade studier. Panelernas bedömningar för mellan 20 och 30 procent av dessa indikationer, fann de, övrensstämde inte fullständigt med existerande klinisk forskning. Författarna konstaterar att resultatet inte är optimalt, pekar på möjliga orsaker, och föreslår åtgärder för att förbättra situationen.[100]

97 Se till exempel Grimshaw & Russell 1993; Sauerland & Neugebauer 2000; Burgers & van Everdingen 2004.
98 Shaneyfelt, Mayo-Smith & Rothwangl 1999, 1902, 1904.
99 Tricoci, Allen, Kramer et al. 2009, 835. De två sällskapen är American College of Cardiology och American Heart Association. Artikeln har redan givit upphov till debatt; se Norris et al. 2009.
100 Nicollier-Fahrni et al. 2003, 20–21.

Nicollier-Fahrni och hennes medförfattare ger värdefulla synpunkter i sin artikel, men den fråga de fokuserar är i viss mån fel ställd. Eftersom litteraturen om coloskopi är begränsad har författarna endast funnit 16 publicerade studier, varav 8 RCTs, att ställa panelens bedömningar mot. Inte heller för de 5 procent av indikationerna för coloskopi som här diskuteras existerar alltså klinisk forskning av nämnvärd omfattning. Om de har rätt som hävdar att experters åsikter bör ges hög prioritet där den publicerade litteraturen ger knapphändig information – och Nicollier-Fahrni och hennes kolleger argumenterar själva för den ståndpunkten – är en diskrepans av den storleksordning de har påvisat mellan en panels bedömningar och ett fåtal kliniska prövningar knappast anmärkningsvärd. I den mån konsensusmetoder – vare sig de kallas RAM, modifierad Delfiteknik, hybrider av Delfimetoden och NGT, eller något annat – används för att syntetisera åsikter som delvis bygger på klinisk erfarenhet, på områden där evidensen är begränsad, kan resultatens tillförlitlighet inte rimligen utvärderas med utgångspunkt i deras överensstämmelse med denna bristfälliga evidens. Att granska konsensusmetoder med en sådan måttstock vore rimligt om deras funktion var den som NIH:s konsensuskonferenser utformades för att fylla, men som metaanalys och systematiska översikter senare har övertagit – nämligen att syntetisera vetenskaplig kunskap. I den mån konsensusmetoder fyller den funktion Nicollier-Fahrni och hennes grupp, Paul Shekelle och andra hävdar, är det rimligare att i utvärderingar fokusera deras reliabilitet och prediktiva validitet.

Detta betyder inte att frågan om vilket vetenskapligt underlag expertpanelers bedömningar bygger på är oväsentlig. Tvärtom är det anmärkningsvärt att RAM:s arkitekter har tagit så lätt på frågan om hur översikter av relevant forskning har utarbetats att de inte brytt sig om att beskriva vilka procedurer de har tillämpat. År 1987, när RAM just hade lanserats och Brooks grupp var i färd med att avsluta ett stort antal studier, publicerade Cynthia Mulrow en svidande uppgörelse med den rådande normen för översiktsartiklar om medicinsk litteratur.[101] Brook och hans grupp uppvisar en häpnadsväckande nonchalans när de, åratal efter det genomslag Mulrows artikel fick, fortsätter att sammanställa litteratur som vore detta en oproblematisk uppgift. Tar man ambitionen att formalisera konsensusmetoder allvarligt måste den givetvis även omfatta översikter av tillgänglig forskning. I en artikel från 1990, i vilken RAM jämförs med två modeller för att utarbeta riktlinjer för klinisk praktik, formuleras vad som borde vara

101 Mulrow 1987.

en självklarhet. Författarna argumenterar för formalisering av sådana metoder, och infogar därefter följande kommentar:

> *For example, a literature review conducted according to explicit, well-crafted criteria is preferred to one conducted without such criteria.*[102]

Ingen tvekan råder om att RAM, bland de metoder som identifierats i arbetet med denna rapport, är den som bäst motsvarar den funktion som uppdraget har gällt. Lika tydligt är att två av de metodologiska problem som kriterier för lämplighet är förknippade med måste ges prioritet när metoden tillämpas. Problemen är desamma som främst har uppmärksammats i debatten om konsensuskonferenser, nämligen panelernas sammansättning och de procedurer som följs när relevant vetenskaplig litteratur identifieras och sammanställs. Hur det förra problemet har hanterats för RAM:s vidkommande har diskuterats i ett tidigare avsnitt. Vad gäller det senare problemet delar jag till fullo, trots mina reservationer mot en viss aspekt av det resonemang Nicollier-Fahrni och hennes medförfattare för, deras principiella ståndpunkt angående litteraturöversikternas betydelse för de resultat som kan uppnås med RAM:

> *Further improvements in the process of developing appropriateness criteria… should better combine published evidence and complementary judgement from multidisciplinary expert panels. In particuar, existing published evidence should be more systematically presented and integrated into the process…*[103]

Övriga funktioner

Föreliggande framställning har tlll stora delar kretsat kring behovet av att på systematiska sätt samla in klinisk erfarenhet på områden som bristfälligt täcks av publicerade studier. Avslutningsvis vill jag kortfattat peka på att konsensusmetoder även fyller andra funktioner. Dels blir den bild av metodernas användning som rapporten ger därmed mer fullständig, trots att utrymme för att närmare beskriva dessa ytterligare funktioner inte gives. Dels är en del av de förslag som framförts om hur metoder för dessa syften kan vidareutvecklas av

102 Audet, Greenfield & Field 1990, 711.
103 Nicollier-Fahrni et al. 2003, 21.

relevans även för det användningsområde som har fokuserats i rapporten.

Till att börja med är det värt att nämna att konsensuskonferenser fortfarande används, trots att man för deras ursprungliga funktion nu litar till andra metoder. Behovet av offentliga hearings som ger representanter för allmänheten möjlighet att fråga ut experter och framföra egna synpunkter på teknologi som de berörs av har inte minskat med tiden, och som nämndes i ett tidigare avsnitt har det konferensformat som NIH introducerade 1977 modifierats för att bättre motsvara detta behov. Dessutom samlas experter inom olika medicinska specialiteter regelbundet för att i konsensuskonferenser etablera eller revidera standardiserade kriterier för registrering av data om insatser för återupplivning vid hjärtstillestånd, för bedömning av psykiatriska tillstånd, etc.[104]

Viktigare för rapportens syften är att formella konsensusmetoder även används på områden där tillgången till vetenskapliga studier av hög kvalitet är god. Att på basis av systematiska översikter av RCTs formulera riktlinjer för klinisk praktik är inte oproblematiskt, och många är övertygade om att steget kräver avvägningar som nödvändigtvis bygger på omdömen. Rosalind Raine vid London School of Hygiene and Tropical Medicine, som med en grupp kolleger specialiserat sig på de metoder som används för detta ändamål, hävdar att

> *Most professional societies and national agencies in North America, Australia and Europe recognise that guidelines cannot be based on research evidence alone.*[105]

Behovet av metoder som kan strukturera de processer som organiseras för utarbetande av riktlinjer är därför stort, och detta är idag det primära användningsområdet för formella konsensusmetoder inom medicin. Metoderna fyller alltså inte en, utan två centrala funktioner. De respektive funktionerna följer av att konsensusförfaranden förefaller krävas i situationer som på sätt och vis är varandras motsatser: dels för att erfarenheter ska kunna syntetiseras där tillgången till vetenskapliga studier är knapp, dels för att kliniska riktlinjer ska kunna härledas ur en omfattande evidensbas. I realiteten tycks gränsen mellan de två situationerna ofta vara diffus, eftersom även en omfattande evidensbas alltid innehåller luckor. Däremot är skillnaden klar mellan detaljerade

104 Timmermans & Berg 2003, 1–3.
105 Raine, Sanderson & Black 2005, 631.

indikationer för interventioners lämplighet, av det slag RAM utformats för att ge, och riktlinjer för klinisk praktik, som brukar vara långt mer grovmaskiga.

Enligt Raines forskarlag har professionella organisationer i åtminstone åtta länder använt en modifierad form av den nominella gruppens teknik för att utarbeta riktlinjer. Av författarnas beskrivning framgår att metoden är snarlik RAM, men i en del av sina publikationer nämner de inte Brook och hans grupp, utan hänvisar endast till en av Delbecq och Van de Vens artiklar från tidigt 70-tal.[106] I en annan artikel gör de dock gällande att tre metoder brukar användas: Delfitekniken, NGT och en hybrid av dessa båda. Hybriden är RAM, framgår det snart, och relevanta referenser ges.[107]

Vad metoderna än kallas sammanfaller de i de flesta avseenden med dem som har diskuterats i denna rapport. Litteraturen om metoder för utarbetande av kliniska riktlinjer överlappar med den om hur kriterier för interventioners lämplighet kan bestämmas. Diskussionen om metodologiska problem i det förra sammanhanget är därför av uppenbar relevans för de frågor denna rapport har fokuserat, men här begränsas diskussionen till att nämna ett konkret förslag till förbättring av de rutiner som tillämpas. Raine med kolleger föreslog nyligen att valda delar av riktlinjer som utformats i en expertpanel av det slag vi nu är bekanta med kan sändas ut för kommentarer till en större grupp. Panelen skulle därefter mötas på nytt för att ta ställning till de synpunkter som framkommit, och eventuellt revidera sina rekommendationer. Tillvägagångssättet skulle kombinera den nominella gruppens styrka med Delfimetodens: skälen till att bedömningar går isär kan förklaras och redas ut i en grupp liten nog att samlas kring ett bord, samtidigt som en bredare grupps medverkan borgar för att resultaten är mer representativa, och därmed även stärker deras legitimitet bland potentiella mottagare.[108] Delfitekniken och NGT har länge tillämpats för att mäta och etablera konsensus inom medicin, men trots många års diskussioner är det fortfarande inte uppenbart hur komponenter av de två metoderna bäst kan kombineras.

106 Raine et al. 2004, 429; Hutchings et al 2006, 218.
107 Raine, Sanderson & Black 2005, 631.
108 Raine, Sanderson & Black 2005, 632; Hutchings et al 2006, 223.

Referenser

Antman, Elliott M., Joseph Lau, Bruce Kupelnick et al. 1992. A Comparison of Results of Meta-Analyses of Randomized Controlled Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction. *Journal of the American Medical Association* 268, 240–48.

Audet, Anne-Marie, Sheldon Greenfield & Marilyn Field 1990. Medical Practice Guidelines: Current Activities and Future Directions. *Annals of Internal Medicine 113*, 709–14.

Bengtson, Ann, Johan Herlitz, Thomas Karlsson et al. 1994. The Appropriateness of Performing Coronary Angiography and Coronary Artery Revascularization in a Swedish Population. *Journal of the American Medical Association* 271, 1260–65.

Bernstein, S.J., B. Brorsson, T. Åberg et al. 1999. Appropriateness of Referral of Coronary Angiography Patients in Sweden. *Heart* 81, 470–77.

Black, Nick 2006. Consensus Development Methods. In C. Pope & N. Mays (eds.), *Qualitative Research in Health Care*. London: BMJ Books, 3rd edn, 132–41.

Breart, Gérard 1990. Documentation and Use of Evidence in the Consensus Conference Process. In C. Goodman & S.R. Baratz (eds.), *Improving Consensus Development for Health Technology Assessment:* An International Perspective. Washington, D.C.: National Academy Press, 23–31.

Brook, Robert H. 1994. Appropriateness: The Next Frontier. *British Medical Journal* 308, 218–19.

Brook, Robert H., Mark R. Chassin, Arlene Fink et al. 1986. A Method for the Detailed Assessment of the Appropriateness of Medical Technologies. *International Journal of Technology Assessment in Health Care* 2, 53–63.

Brook, Robert H., Jacqueline B. Kosecoff, R.E. Park et al. 1988. Diagnosis and Treatment of Coronary Disease: Comparison of Doctors' Attitudes in the USA and the UK. *Lancet* 1, 750–53.

Burgers, Jako S. & Jannes J.E. van Everdingen 2004. Beyond the Evidence in Clinical Guidelines. Lancet 364, 392–93.

Chassin, Mark 1989. How Do We Decide Whether an Investigation or Procedure is Appropriate? In A. Hopkins (ed.), *Appropriate Investigation and Treatment in Clinical Practice*. London: Royal College of Physicians of London, 21–29.

Dalkey, Norman & Olaf Helmer 1963. An Experimental Application of the Delphi Method to the Use of Experts. *Management Science* 9, 458–67.

Delbecq, André L. & Andrew H. Van de Ven 1971. A Group Process Model for Problem Identification and Program Planning. *Journal of Applied Behavioral Science* 7, 466–92.

Faxon, David P. 2008. Assessing Appropriateness of Coronary Angiography: Another Step in Improving Quality. *Annals of Internal Medicine* 149, 276–78.

Field, Marilyn J. & Kathleen N. Lohr (eds.) 1992. Guidelines for Clinical Practice: From Development to Use. Washington, D.C.: National Academy Press.

Fink, Arlene, Jacqueline Kosecoff, Mark Chassin & Robert H. Brook 1984. Consensus Methods: Characteristics and Guidelines for Use. *American Journal of Public Health* 74, 979–83.

Gallagher, Morris, Tim Hares, John Spencer et al. 1993. The Nominal Group Technique: A Research Tool for General Practice? *Family Practice* 10, 76–81.

Grimshaw, Jeremy & Ian Russell 1993. Achieving Health Gain Through Clinical Guidelines. I: Developing Scientifically Valid Guidelines. *Quality in Health Care* 2, 243–48.

Guston, David 1999. Evaluating the First U.S. Consensus Conference: The Impact of the Citizens' Panel on Telecommunications and the Future of Democracy. *Science, Tecnology and Human Values* 24, 451–82.

Helmer, Olaf & Nicholas Rescher 1959. On the Epistemology of the Inexact Sciences. *Management Science* 6, 25–52.

Hemingway, Harry, Ruoling Chen, Cornelia Junghans et al. 2008. Appropriateness Criteria for Coronary Angiography in Angina: Reliability and Validity. *Annals of Internal Medicine* 149, 221–31.

Hicks, Nicholas R. 1994. Some Observations on Attempts to Measure Appropriateness of Care. *British Medical Journal* 309, 730–33.

Higgins, Julian P.T, & David J. Spiegelhalter 2002. Being Sceptical About Meta-Analyses: A Bayesian Perspective on Magnesium Trials in Myocardial Infarction. *International Journal of Epidemiology* 31, 96–104.

Horn, Susan Dadakis & John W. Williamson 1977. Statistical Methods for Reliability and Validity Testing: An Application to Nominal Group Judgments in Health Care. *Medical Care* 15, 922–28.

Hult, Erika, under utgivning. Mellan HTA och EBM: SBU som vetenskapens väktare i svensk hälso- och sjukvård.

Hutchings, Andrew, Rosalind Raine, Colin Sanderson & Nick Black 2006. A Comparison of Formal Consensus Methods Used for Developing Clinical Guidelines. Journal of Health Services Research & Policy 11, 218–24.

Huth, Edward J. 2008. The Move Toward Setting Standards for the Content of Medical Review Articles. James Lind Library (**www.jameslindlibrary.org**).

Jacoby, Itzhak 1988. Evidence and Consensus. *Journal of the American Medical Association* 259, 3039.

Jacoby, Itzhak 1990. Sponsorship and Role of Consensus Development Programs within National Health Care Systems. In C. Goodman & S.R. Baratz (eds.), *Improving Consensus Development for Health Technology Assessment: An International Perspective*. Washington, D.C.: National Academy Press, 7–17.

Johnsson, Monica 1988. Evaluation of the Consensus Development Program in Sweden: Its Impact on Physicians. *International Journal of Technology Assessment in Health Care* 4, 89–94.

Jones, Jeremy & Duncan Hunter 1995 Consensus Methods for Medical and Health Services Research. *British Medical Journal* 311, 376–80.

Joss, Simon & John Durant 1995. Introduction. In S. Joss & J. Durant (eds.), *Public Participation in Science: The Role of Consensus Conferences in Europe*. London: Science Museum, 9–13.

Jörgensen, Torben 1995. Consensus Conferences in the Health Care Sector. In S. Joss & J. Durant (eds.), *Public Participation in Science: The Role of Consensus Conferences in Europe*. London: Science Museum, 17–29.

Kahn, Katherine L., Jacqueline B. Kosecoff, Mark R. Chassin et al. 1988. Measuring the Clinical Appropriateness of the Use of A procedure: Can We Do It? *Medical Care* 26, 415–22.

Lambert, Helen 2006. Accounting for EBM: Notions of Evidence in Medicine. Social Science & Medicine 62, 2633–45.

Landeta, Jon 2006. Current Validity of the Delphi Method in Social Sciences. *Technological Forecasting & Social Change* 73, 467–82.

Leichsenring, Falk & Sven Rabung 2008. Effectiveness of Long-term Psychodynamic Psychotherapy: A Meta-analysis. *Journal of the American Medical Association* 300, 1551–65.

Linstone, Harold A. 1978. The Delphi Technique. In J. Fowles (ed.), *Handbook of Futures Research*. London: Greenwood Press, 273–300.

Mulley, Albert G. & Kim A. Eagle 1988. What is Appropriate Care? *Journal of the American Medical Association* 260, 540–41.

Mullan, Fitzhugh & Itzhak Jacoby 1985. The Town Meeting for Technology: The Maturation of Consensus Conferences. *Journal of the American Medical Association* 254, 1068–72.

Mulrow, Cynthia D. 1987. The Medical Review Article: State of the Science. *Annals of Internal Medicine* 106, 485–88.

Murphy, M.K., N.A. Black, D.L. Lamping et al. 1998. Consensus Development Methods, and Their Use in Clinical Guideline Development. *Health Technoly Assessment* 2 (3), 1–88.

Naylor, C. David 1998. What is Appropriate Care? *New England Journal of Medicine* 338, 1918–20.

Nicollier-Fahrni, Anne, John-Paul Vader, Florian Froehlich et al. 2003. Development of Appropriateness Criteria for Colonoscopy: Comparison Between a Standardized Expert Panel and an Evidence-Based Medicine Approach. *International Journal for Quality in Health Care* 15, 15–22.

Nielsen, Annika Porsborg, Jesper Lassen & Peter Sandøe 2007. Democracy at Its Best? The Consensus Conference in a Cross-National Perspective. *Journal of Agricultural and Environmental Ethics* 20:13–35.

Norris, Susan L. et al. 2009. Clinical Practice Guidelines and Scientific Evidence. *Journal of the American Medical Association* 302, 142–47.

Park, Rolla Edward, Arlene Fink, Robert H. Brook et al. 1986. Physician Ratings of Appropriate Indications for Six Medical and Surgical Procedures. *American Journal of Public Health* 76, 766–72.

Perry, Seymour 1987. The NIH Consensus Development Program. A Decade Later. *New England Journal of Medicine* 317, 485–88.

Phelps, Charles E. 1993. The Methodologic Foundations of Studies of the Appropriateness of Medical Care. *New England Journal of Medicine* 329, 1241–45.

Raine, Rosalind, Colin Sanderson, Andrew Hutchings et al. 2004. An Experimental Study of Determinants of Group Judgments in Clinical Guideline Development. *Lancet* 364, 429–37.

Raine, Rosalind, Colin Sanderson & Nick Black 2005. Developing Clinical Guidelines: A Challenge to Current Methods. *British Medical Journal* 331, 631–33.

Relman, Arnold S. 1988. Assessment and Accountability: The Third Revolution in Medical Care. *New England Journal of Medicine* 319, 1220–22.

Rinchuse, Donald J., Daniel J. Rinchuse, Sanjivan Kandasamy & Marc B. Ackerman 2008. Deconstructing Evidence in Orthodontics: Making Sense of Systematic Reviews, Randomized Clinical Trials, and Meta-Analyses. *World Journal of Orthodontics* 9, 167–76.

Sauerland, S. & E. Neugebauer 2000. Consensus Conferences Must Include a Systematic Search and Categorization of the Evidence. *Surgical Endoscopy* 14, 908–10.

Shaneyfelt, Terrence M., Michael F. Mayo-Smith & Johann Rothwangl 1999. Are Guidelines Following Guidelines?: The Methodological Quality of Clinical Practice Guidelines in the Peer-Reviewed Medical Literature. *Journal of the American Medical Association* 281, 1900–05.

Shekelle, Paul 2004. The Appropriateness Method. *Medical Decision Making* 24, 228–31.

Shekelle, Paul G. 2009. Appropriateness Criteria: A Useful Tool for the Cardiologist. *Heart* 95, 517–20.

Shekelle, Paul G., James P. Kahan, Steven J. Bernstein et al. The Reproducibility of a Method to Identify the Overuse and Underuse of Medical Procedures. *New England Journal of Medicine* 338, 1998, 1888–95.

Timmermans, Stefan & Marc Berg 2003. The Gold Standard: *The Challenge of Evidence-Based Medicine and Standardization in Health Care*. Philadelphia, Pa.: Temple University Press.

Tricoci, Pierluigi, Joseph M. Allen, Judith M. Kramer et al. 2009. Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines. *Journal of the American Medical Association* 301, 831–41.

Trivedi, Vandan M. 1982. Measurement of Task Delegations Among Nurses by Nominal Group Process Analysis. *Medical Care* 20, 154–64.

Van de Ven, Andrew & Andre L. Delbecq 1971. Nominal Versus Interacting Group Processes for Committee Decision-Making Effectiveness. *The Academy of Management Journal* 14, 203–12.

Van De Ven, Andrew H. & André L. Delbecq 1972. The Nominal Group as a Research Instrument for Exploratory Health Studies. A*merican Journal of Public Health* 62, 337–42.

Van De Ven, Andrew H. & André L. Delbecq 1974. The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes. *The Academy of Management Journal* 17, 605–21.

Vang, Johannes 1986. The Consensus Development Conference and the European Experience. *International Journal of Technology Assessment in Health Care* 2, 65–76.

Wennberg, John E. 1984. Dealing with Medical Practice Variations: A Proposal for Action. *Health Affairs* 3, No 2, 6–32.

Wennberg, John E. 1989. The Agenda for Outcomes Research. In A. Hopkins (ed.), *Appropriate Investigation and Treatment in Clinical Practice*. London: Royal College of Physicians of London, 77–90.

Winslow, Constance Monroe, Jacqueline B. Kosecoff, Mark Chassin et al. 1988. The Appropriateness of Performing Coronary Artery Bypass Surgery. *Journal of the American Medical Association* 260, 505–09.

Woolf, Steven H. 2008. The Meaning of Translational Research and Why It Matters. *Journal of the American Medical Association* 299, 211–13.

Wortman, Paul M. 2004. Consensus Panels: Methodology. In N.J. Smelser & P.B. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Elsevier, 2609–13.

Wortman, Paul M., Joshua M. Smyth, John C. Langenbrunner & William H. Yeaton 1998. Consensus Among Experts and Research Synthesis: *A* Comparison of Methods. *International Journal of Technology Assessment in Health Care* 14, 109–22.

Wortman, Paul M., Amiram Vinokur & Lee Sechrest 1988. Do Consensus Conferences Work? A Process Evaluation of the NIH Consensus Development Program. *Journal of Health Politics, Policy and Law* 13, 469–98.

Konsensusmetoder inom hälso- och sjukvård

En kunskapsöversikt

Allt arbete i hälso- och sjukvården ska vara grundat i vetenskap och beprövad erfarenhet. Detta gäller för varje beslut om åtgärd, exempelvis att utfärda recept, besluta om röntgenundersökning eller operation. För många åtgärder är det vetenskapliga kunskapsunderlaget bristfälligt och beprövad erfarenhet blir då den kunskapskälla man har att förlita sig på. Att på ett systematiskt sätt fånga och mäta beprövad erfarenhet är inte enkelt. Denna kunskapsöversikt, som framtagits inom ramen för utvecklingsprojektet Nationella medicinska indikationer, beskriver och analyserar olika metoder som använts för att fånga och mäta erfarenhetsbaserad kunskap.

Författare till kunskapsöversikten är docent Ingemar Bohlin vid Avdelningen för teknik- och vetenskapsstudier, Sociologiska institutionen, Göteborgs universitet.

TRYCKSAKER FRÅN SVERIGES KOMMUNER OCH LANDSTING
BESTÄLLS PÅ WWW.SKL.SE ELLER PÅ
TFN 020-31 32 30, FAX 020-31 32 40.

ISBN 978-91-7164-486-2

Sveriges
Kommuner
och Landsting

118 82 Stockholm, *Besök* Hornsgatan 20
Tfn 08-452 70 00, *Fax* 08-452 70 50
info@skl.se, www.skl.se