Evaluering av medisinsk behandling: Hva vet vi, og hvordan bør kunnskapen håndteres?

Jan Magnus Bjordal[1,2], Atle Klovning[3,4] og Lars Slørdal[5,6]

1) Institutt for fysioterapi, Avdeling for helse- og sosialfag, Høgskolen i Bergen
2) Seksjon for fysioterapivitenskap, Institutt for samfunnsmedisinske fag, Universitetet i Bergen
3) Seksjon for allmennmedisin, Institutt for allmenn- og samfunnsmedisin, Universitetet i Oslo
4) Allmennmedisinsk forskningsenhet, Institutt for allmenn- og samfunnsmedisin, Universitetet i Oslo
5) Institutt for laboratoriemedisin, barne- og kvinnesykdommer, NTNU
6) Avdeling for klinisk farmakologi, St. Olavs Hospital

Korrespondanse: Lars Slørdal, Avdeling for klinisk farmakologi, St. Olavs Hospital, 7006 Trondheim
E-post: lars.slordal@ntnu.no Telefon: 73598843

SAMMENDRAG

Kunnskap om effekten av medisinsk behandling fremskaffes i kliniske studier. Godt planlagte og gjennomførte randomiserte, placebokontrollerte studier tillegges størst vekt når data skal vurderes. De datamengdene vi besitter når det gjelder forskjellige behandlinger kan oppsummeres i oversikter eller terapianbefalinger som ofte blir normative for klinisk praksis. Dette stiller store krav til evalueringsprosedyrene. I denne oversiktsartikkelen diskuteres i innledningen noen eksempler på at kunnskapsgrunnlaget er blitt feiltolket. Deretter omtales kvalitetskriterier som legges til grunn i vurderingen av data, og den metodologiske basis for meta-analyser. Usikkerhetsmomenter og potensielle fallgruver ved bruk av meta-analyser omtales. Meta-analysens fortrinn eksemplifiseres ved bruk av data om behandling av korsryggsmerter, hvor det nylig er utarbeidet kliniske retningslinjer som etter vår oppfatning er beheftet med feil eller mangler. Meta-analysen har vesentlige fortrinn fremfor mindre formaliserte metoder for oppsummering av kunnskap om terapi ved sykdom.

Bjordal JM, Klovning A, Slørdal L. **How do we evaluate knowledge about therapeutic efficacy?** *Nor J Epidemiol* 2008; **18** (2): 137-146.

ENGLISH SUMMARY

Data on the effects of medical therapies are available from clinical studies. Well planned and executed randomised, placebo-controlled studies are given the highest level of evidence when study results are reviewed. The available body of data on different therapies lend themselves to synthesis in clinical overviews or therapy recommendations that regularly achieve status as normative for clinical practice. This warrants caution and rigour as far as the underlying procedures are concerned. The current review encompasses a brief discussion of previous failures of pharmacovigilance, including the thalidomide disaster and the COX-2 debacle. The quality criteria for data evaluation and the methodological basis for meta-analyses are presented. Possible sources of error in meta-analyses, such as different reporting of effects, publication bias, heterogeneity, conflicts of interest and problems with evaluation of side effects, are discussed. The advantages of meta-analysis are exemplified with data on therapy of lower back pain, which recently has been subjected to clinical practice guidelines that, in our view, have several shortcomings. Meta-analyses have significant advantages over less formalised methods for the evaluation of the current state of knowledge of therapeutic alternatives.

FRA TRO TIL EVIDENSBASERT TERAPI?

Innføring av et terapeutikum på basis av en systematisk sammenliknende studie skjedde første gang i det 18. århundre. En skipslege i den britiske marinen, James Lind, samlet i 1747 tolv sjømenn med vitaminmangelsykdommen skjørbuk i samme banjer og satte dem på identiske dietter. Sjømennene ble så "randomisert" til å motta én av seks forskjellige behandlinger man trodde hadde effekt ved skjørbuk: sider, svovelsyre, eddik, sjøvann, muskat eller sitrusfrukt (i dette tilfellet både appelsin og sitron). Sitrusfruktsupplementet virket overlegent best (1), og det ble i ettertid standardprosedyre å behandle mannskap i den britiske marinen med sitronsaft på lange sjøreiser – med det resultat at skjørbuk nesten ble borte fra britiske marinefartøyer fra slutten av 1700-tallet. Da Lind gjorde eksperimentet, mente The Royal College of Physicians at svovelsyre var førstevalg ved sykdommen, mens marineadministrasjonen foretrakk eddik; to eksempler på at "autoriteter" – i fravær av harde data – ofte tar feil i slike spørsmål (2).

Et annet eksempel på en systematisk og eksperimentell tilnærming til legemiddeltesting finnes i

William Witherings eksperimenter med revebjelle-ekstrakter (som inneholdt digitalisglykosider) i behandlingen av "dropsy" (en lite presis 1700-tallsdiagnose som i dag best oversettes med "ødem"; de fleste av pasientene hadde etter alt å dømme hjertesvikt) i årene 1775-85 (3).

I forbindelse med framveksten av "moderne" legevitenskap og medisinsk praksis er en lang rekke legemidler og andre behandlingsmodaliteter tatt i bruk. Terapiregimer som er eldre enn 40-50 år har sjelden vært gjenstand for rigorøs testing før lanseringen. Noen ganger er dette en fordel. Penicilliner er for eksempel nyretoksisk hos marsvin, og ville av den grunn trolig aldri ha overlevd prekliniske tester hvis de ble oppdaget i dag (4). Det finnes dessverre mange eksempler på at uhensiktsmessig og/eller farlig farmakoterapi har hatt betydelig og langvarig utbredelse.

Den "nye" æra i farmakoterapien – med rigorøs og sekvensiell testing av legemidlenes yteevne – startet rundt 1960, og faller sammen med tre viktige hendelser; erkjennelsen av placeboeffekten (5), gjenoppdagelsen av den randomiserte kontrollerte utprøvningen som marinelegen James Lind introduserte 200 år tidligere (1,2), og det som i ettertid er blitt kalt thalidomidskandalen (6). Thalidomid var et beroligende og søvnframkallende middel som ble markedsført i mange vestlige land (inkl. Norge, men ikke USA) i årene etter introduksjonen i 1956, og som tidlig på 1960-tallet ble vist å føre til en svært høy frekvens av nerveskader og karakteristiske medfødte misdannelser hos henholdsvis ca. 40.000 og 8-12.000 pasienter (6). Denne tragedien betraktes i ettertid som den hendelsen som framfor noe annet satte fokus på viktigheten av å overvåke legemidlers bivirkninger og sikkerhet.

Femti år inne i den "nye" tid skulle man tro at de viktigste spørsmål knyttet til effekt og trygghet av de legemidlene som markedsføres for lengst var avklart, og at skolemedisinens grunnlag for legemiddelbruk er kunnskapsbasert og rasjonell. Slik er det ikke. Markedsføringen av østrogenterapi begynte i 1966, da en gynekolog i New York, Dr. Robert Wilson, publiserte boken "Feminine forever" (7). Boken, som ble en bestselger, lanserte konseptet om menopausen som sykdom ("ovariesvikt"), som det naturligvis fantes et legemiddel for; Ayerst/Wyeths Premarin® (forkortelsen står for "PREgnant MARe urINe" og henspiller på at datidens dominerende østrogenpreparat var utvunnet fra urinen til gravide hester) (8). Fra 1970-årene fikk oppfatningen av østrogenbehandling som en eliksir som holdt eldre kvinner unge, friske og attraktive stor utbredelse, og terapien ble – på basis av mer eller mindre velfunderte tankerekker og data fra kohortstudier som ofte anvendte pseudoendepunkter – særlig tillagt effekt som beskyttelse mot osteoporose og påfølgende beinbrudd og forekomst av kardiovaskulær sykdom. Vi måtte vente til 2002 før en stor randomisert placebokontrollert studie konklusivt viste at østrogenets helsebringende effekter var ren ønsketenkning; behandlingen gav i realiteten en økning i kardiovaskulær sykelighet, og kostnad/nytte-regnskapet for postmenopausal østrogenbruk var totalt sett negativt (9). Salget av østrogenpreparater gikk i ettertid mye tilbake.

I 1999 begynte markedsføringen av en ny klasse smertestillende midler, de såkalte selektive cyklooksygenase (COX)-2-hemmerne, som ble hevdet å være både mer effektive og langt tryggere enn de gamle ikke-steroide antiinflammatoriske legemidlene (NSAIDs) de skulle erstatte. COX-2-hemmerne ble snart markedsledere innen et stort og lukrativt legemiddelsegment. Høsten 2004 ble COX-2-hemmeren rofecoxib (Vioxx®) trukket fra verdensmarkedet av produsenten på grunn av at midlet var assosiert med en økt frekvens av arteriell tromboembolisk sykdom (hjerteinfarkt, hjerneslag og plutselig død). I ettertid er flere andre COX-2-hemmere trukket tilbake, og alle disse midlene har mistet forhåndsgodkjenningen i blåreseptordningen og vært gjenstand for en dramatisk salgssvikt (10). Vi vet ikke sikkert hvor mange pasienter som er blitt skadet eller drept av disse legemidlene, men forsiktige anslag tilsier at COX-2-hemmere er ansvarlig for mer enn ti ganger flere dødsfall og alvorlige senfølger enn det som resulterte fra tildragelsen som nå gjerne omtales som thalidomidskandalen (11).

Det finnes også en rekke andre legemidler hvis virkninger og berettigelse er omdiskutert. Nyere eksempler omfatter blant annet glukosamin ved artrose (12) og kolinesterasehemmere ved Alzheimers sykdom. I 2005 publiserte British Medical Journal en systematisk oversikt over randomiserte kliniske utprøvninger med de nye Alzheimer-midlene donepezil (Aricept®), rivastigmin (Exelon®) og galantamin (Reminyl®) (13). Studien konkluderte med at kunnskapsgrunnlaget var preget av suboptimale studier og at effekten av legemidlene i beste fall var liten. Dette ble imøtegått i bl.a. norske medier, hvor "alle" – som i denne sammenhengen betyr legemiddelmyndighetene, produsentene og det nasjonale geriatriske ekspertmiljøet, som i noen grad har mottatt forskjellige ytelser fra legemiddelprodusentene – tydeligvis var enige om at midlene var effektive (14).

Et annet eksempel gjelder omega-3-fettsyrer (Omacor®, andre). En systematisk oversikt i BMJ, som rapporterte at inntak av omega-3-fettsyrer ikke hadde effekt på total dødelighet eller på forekomst av hjerte-karsykdom eller kreftsykdom (15), ble kritisert – med henvisning til bl.a. inklusjonskriterier og andre, eldre meta-analyser – av en norsk gruppe i en påfølgende kommentar i Tidsskrift for Den norske legeforening (16).

Rasjonell medisinsk terapi forutsetter at vi har kunnskap om hvorvidt behandlingene virker og om hvor sikre de er. Avklaring av slike spørsmål har åpenbare følger for pasientbehandlingen, og har derfor også store helsemessige og økonomiske konsekvenser. Vi skal i det følgende redegjøre for hvordan kunnskap om legemidlers yteevne og sikkerhet innhentes og evalueres.

GRADERING AV DOKUMENTASJONEN OG STYRKEN PÅ ANBEFALINGENE – HVORDAN GJØRES DET?

Vurderinger av nytten av behandlingsformer og produksjon av retningslinjer for behandling av en sykdom kan gjøres på flere måter. Prosessen finner vanligvis sted i et spekter som defineres av to ytterligheter. Man kan

1) basere seg på enighet mellom antatte eksperter, eventuelt også behandlere og pasienter innenfor fagområdet, eller
2) basere seg på det vitenskapelige dokumentasjonsgrunnlaget, som videre kan vektes og graderes etter nivået på den vitenskapelige dokumentasjonen.

Velger man den første varianten, er saken grei; forfattergruppen setter seg sammen og diskuterer til enighet oppnås. Det finnes relativt nye eksempler på bruk av slike konsensusmodeller (17), men de er prinsipielt uvitenskapelig i sin tilnærming og mangler reproduserbarhet, siden resultatet alltid vil avhenge av hvem som inngår i arbeidsgruppen. Alternativ 2 innebærer en mer rigorøs og vitenskapelig tilnærming. Vitenskapeliggjøringen av prosessen med utarbeiding av retningslinjer er imidlertid sjelden uproblematisk.

Den vitenskapelige dokumentasjon graderes ofte med hensyn til kvaliteten og styrken i materialet (Tabell 1). Måten dette gjøres på kan variere. De fleste av graderingssystemene har metodiske svakheter som gjør dem upresise og åpne for skjevheter i vurderingen av datagrunnlaget. Systemene mangler ofte vurdering av klinisk relevans eller avveininger av forholdet mellom nytte og risiko (bivirkninger), samtidig som foretatte vurderinger i varierende grad er reproduserbare. Kvaliteten vurderes vanligvis på grunnlag av studiedesign, studiekvalitet, konsistens, relevans (mht. bl.a. studiepopulasjon og intervensjon), utfall og sammenligninger med andre tilgjengelige studier.

Kvaliteten av dokumentasjonen kan graderes ned hvis det foreligger publikasjonsbias, som for eksempel arter seg ved at studier som er positive til en intervensjon publiseres og legges til grunn, mens data som viser ingen eller til og med negativ effekt av intervensjonen ikke gjenfinnes i faglitteraturen, eller hvis de tilgrunnliggende studiene gir et upresist estimat av effekten og/eller at det viser seg at de data man har å støtte seg til er for begrensede eller sparsomme. Tilstedeværelse av forvekslingsfaktorer (”confounding factors”) vil også bidra til å trekke evidensstyrken ned. Motsvarende kan kvaliteten oppgraderes hvis det foreligger en sterk sammenheng mellom intervensjon og utfall, og hvis de tilgrunnliggende data viser dose/respons-sammenhenger (jo høyere dose, jo mer uttalt respons).

Kliniske retningslinjer skal oppsummere kunnskap om epidemiologi, etiologi, diagnostikk, forebygging, behandling og oppfølging av den eller de tilstander det fokuseres på. Avveininger mellom positive effekter som fremkommer i randomiserte kontrollerte undersøkelser og negative effekter som bivirkninger (som ofte innrapporteres i form av kasuistikker) må tydeliggjøres. I tillegg må man ta hensyn til andre typer nytte- og kostnadsvurderinger, ofte i form av helseøkonomiske avveiinger, og integrere individuelle og kollektive verdivalg som grunnlag for konkrete anbefalinger.

Tabell 1. Én av mange måter å gradere beste evidens på, etter anbefalingsstyrke (A–D) og evidensnivå (1–5). Tabellen baserer seg på webstedet til Centre for Evidence-Based Medicine, Oxford University (1) og er publisert i en norsk fagbok (2). Andre veiledere bruker *, **, og ***, eller A, B og C.

A	1a	Systematisk oversikt over randomiserte, kontrollerte studier
	1b	Minst én randomisert kontrollert studie med smalt 95% konfidensintervall
	1c	«Alle eller ingen»-effekt
B	2a	Systematisk litteraturoversikt: kohortstudier
	2b	Minst én kohortstudie eller 1 lavkvalitets randomisert kontrollert studie
	2c	«Outcomes»-forskning eller helsetjenesteforskning
	3a	Systematisk litteraturoversikt over kasus-kontroll-studier
	3b	Minst én kasus-kontroll-studie
C	4	Kasuistikkserier, lavkvalitets kohort / kasus-kontroll-studier
D	5	Ekspertvurdering uten kritisk vurdering, eller basert på ekstrapolert basalforskning

1. Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, et al. Levels of evidence and grades of recommendations. Oxford, 2001. Available from: www.cebm.net/levels_of_evidence.asp#levels.
2. Klovning A. Fagutvikling. I: Misvær N, Oftedal G, red. Håndbok for helsestasjoner 0-5 år, 2. utg. Oslo: Kommuneforlaget, 2006: 39-46.

META-ANALYSEN: HVA ER DET, OG HVORDAN UTFØRES DEN?

En meta-analyse er en statistisk metode som sammenholder resultater fra flere studier for å beregne den totale gjennomsnittseffekten av en behandlingsform. Meta-analyser kan også brukes for å beregne risiko (insidens) for at en sykdom eller at en bivirkning skal oppstå.

Utfall av en intervensjon kan registreres som kontinuerlige (for eksempel målt på en skala fra 1 til 10), kategoriske (for eksempel "mye verre-verre-uendret-bedre-mye bedre") eller dikotomiserte (for eksempel "mer enn 50% smerte-mindre enn 50% smerte","frisk-syk" eller "levende-død") data. I studier av effekten av smertestillende behandlingsregimer vil som regel smerte målt på en visuell analog skala (VAS) på 100 mm være det primære effektmålet. Den statistiske metoden som man bruker til å beregne effekten kalles for "weighted mean difference" (WMD) mellom aktiv behandling og placebo målt i mm på en 100 mm VAS, og innebærer en vekting basert på effekten dividert med verdien av variansen. Dette betyr at jo mindre variansen er, jo sterkere vektes den aktuelle studien i analysen. De fleste statistikkprogrammer for meta-analyser angir den prosentvise vektingen av enkeltstudiene. Hvis vi har kontinuerlige utfallsmål på ulike skalaer, er det vanlig å beregne det vi kaller "standardised mean difference" (SMD), som er en arbitrær enhet og i realiteten et forholdstall mellom effekten og variansen (jo mindre varians, jo høyere SMD). Det er generelt enklere å relatere WMD til klinisk relevante effekter på en 100 mm VAS skala. Et eksempel: For uspesifikke ryggsmerter er gjennomsnittlig ("baseline") smerteintensitet på ca. 60 mm i det som foreligger av studier. Med et slikt utgangspunkt kan effektmål operasjonaliseres slik at en "liten/klinisk ubetydelig" klinisk effekt tilsvarer 9 mm (6-11,9) mm) bedring i forhold til placebo, en "moderat" effekt tilsvarer 15 mm (12-17,9 mm) mens en "meget god" effekt tilsvarer mer enn 18 mm. For relativ risiko (RR) beskrives gjerne en "liten" effekt 1,2–1,5, en "moderat" effekt 1,6–1,9 og en "meget god" effekt 2 og mer.

Når det gjelder dikotomiserte data, brukes vanligvis relativ risiko (RR) fordi dette angir det relative styrkeforholdet mellom effektene av placebo og intervensjonen. I noen sammenhenger angir man for eksempel at RR-verdier over 2 med smale konfidensintervaller anses som statistisk meget sterke, og at disse vil trekke opp evidensgraden. Man kan også subgruppere materialet etter dose i meta-analyser for å undersøke eventuelle dose-respons sammenhenger.

USIKKERHETSMOMENTER I META-ANALYSER

Effektmål

Relativt ofte oppgir teksten i forskningsrapportene ikke den egentlige effektstørrelsen, men p-verdier og statistisk signifikans. Dette er spesielt vanlig i industrifinansierte studier (se nedenfor). P-verdier sier imidlertid intet om effektstørrelse, og slik resultatrapportering vanskeliggjør meta-analyse, som vanligvis krever gjennomsnittsverdier for endring ("mean change") og standarddeviasjonen (SD) for effektberegning.

Imidlertid avhenger p-verdien både av gjennomsnittlig endring, variansen og utvalgsstørrelsen (n). Med statistiske metoder kan vi dermed kalkulere SD fra andre variansmål for endring hvis vi kjenner utvalgsstørrelsen, gjennomsnittlige før/etter data for gruppene og p-verdier. Slik transformering av data er arbeidskrevende og gjøres sjelden i praksis, men vi har selv benyttet denne løsningen i flere av våre publiserte meta-analyser (18,19). Ved manglende variansdata, kan man foreta en rimelighetsvurdering av antatt SD, og flere metoder for dette har vært brukt i meta-analyser (20). En annen kilde til frustrasjon når det gjelder effektmål, er at forfattere av forskningsrapporter ofte roter med variansmålene; en gjenganger er at standard error of the mean (SEM) feilaktig oppgis som standarddeviasjon (SD), noe som fører til at studien får en feilaktig høy SMD. Kontrollregning med de oppgitte data for å finne WMD eller prosentvis forskjell mellom aktiv behandling og placebo vil som regel avsløre slike variansfeil.

Noen ganger omdefineres utfallene i en studie underveis. Hvis studien ikke oppnådde de resultatene man forventet, kan det for eksempel være fristende å tildekke dette ved å lansere nye måter å telle resultater på etter at undersøkelsen er gjennomført. Et særdeles godt eksempel på slike uetterrettelige "post hoc"-analyser finnes i CLASS; Pfizers studie av COX-2-hemmeren celecoxib fra 2000 (21), hvor sentrale funn ble drastisk fortegnet gjennom blant annet å redusere observasjonstiden (og dermed frekvensen av alvorlige bivirkninger, noe som favoriserte celecoxib) i ettertid (22). Det tjener forskningsmiljøet til lite ære at CLASS ikke er blitt gjenstand for tilbaketrekking.

Publikasjonsbias

Publikasjonsrutiner kan i seg selv forårsake skjevheter i datagrunnlaget. I en studie av den videre skjebnen til 42 industrifinansierte, placebokontrollerte undersøkelser av selektive serotonerge reopptakshemmere (SSRI) påviste man at studier som viste signifikante, positive resultater av SSRI-medikasjonen nesten uten unntak endte opp som selvstendige artikler i faglitteraturen, mens studier som ikke viste signifikante resultater i favør av intervensjonen forble upublisert, eller endte opp i rapporter hvor de ble omtalt sammen med andre undersøkelser (23). Motsvarende ble positive studier ofte omtalt i flere publikasjoner; i det svenske materialet fant man at 21 undersøkelser bidro med data til to publikasjoner eller flere, mens tre enkeltstudier hver bidro til fem publikasjoner (23). Både utelatelse og dobbeltpublisering av data ("Tordenskjolds soldater") bidrar til å svekke grunnlaget for påfølgende vurderinger.

En annen vanlig årsak til publikasjonsbias er at man kun inkluderer engelskspråklige studier i datagrunnlaget for meta-analyser.

Heterogenitet

Hvis en effekt ikke er konsistent i hele studiepopulasjonen (ved at den f.eks. kun er til stede hos ett av kjønnene eller hos en bestemt aldersgruppe), eller i alle de forskjellige populasjonene som inngår i en metaanalyse, er data heterogene. Heterogenitet er altså et mål på hvor ensartede data er. Ulike tester brukes for å angi heterogenitet, og graden av heterogenitet tallfestes ofte som Q-verdier (Z-test) eller som en prosentandel (I^2 test). Hvis man finner statistisk heterogenitet, anbefales det at man bruker en såkalt "random effects model", mens en ved fravær av statistisk heterogenitet anbefaler bruk av en såkalt "fixed effects model". Imidlertid er avviket i resultat mellom de to metodene marginalt og i størrelsesorden 1-2 mm på 100 mm VAS.

Bruken av heterogenitetstester er blitt kritisert fordi slike mål kan være et resultat av en sirkelargumentasjon. Årsaker til heterogeniteten kan vanligvis finnes enten i studiedesign, egenskaper ved selve intervensjonen eller pasientutvelgelsen.

Et eksempel som illustrerer at heterogenitetsanalyser kan gi interessant informasjon, finnes i glukosaminforskningen: Data om effekt av glukosamin ved artrose er hovedsakelig framkommet i industrifinansierte studier, mens uavhengige forskere ikke har funnet holdepunkter for smertelindrende eller sykdomsmodifiserende effekter. I en systematisk analyse av randomiserte, dobbeltblindete, placebokontrollerte studier av glukosamin påviste Vlad og medarbeidere (24) at de uavhengige studiene var ensartede i sine resultater, mens industristudiene – altså de som fant effekt av glukosamin – var heterogene, noe som i denne sammenhengen ble oppfattet som uttrykk for at det forelå skjevheter i måten forsøkene var gjennomført på, som favoriserte intervensjonen – altså at det heftet "noe" ved disse resultatene.

Industrirelasjoner og andre interessekonflikter

Innenfor feltet muskel/skjelettsmerter er 4 av 5 legemiddelstudier industrifinansierte, mens bare 1 av 5 studier med ikke-patenterte fysiske virkemidler har en slik finansieringskilde. Det finnes noen typiske trekk for industrifinansiert forskning som skiller den fra studier som er finansiert av uavhengige kilder.

Kriterier for pasientutvelgelse og måletidspunkt ser ut til å avhenge av finansieringskilden. I tidligere meta-analyser (18,19) har vi vist at den pasientutvelgelsesprosedyren som for eksempel brukes i smertestudier med NSAIDs favoriserer intervensjonen. En vanlig metode er kun å inkludere etablerte NSAID-brukere, som per definisjon er "respondere" som alle kan forventes å reagere favorabelt på eksposisjon for et nytt NSAID-preparat. Vi beregnet at en slik seleksjonsprosedyre gav en urettmessig forbedring av effektstørrelsen på 38%. I et uselektert spansk pasientmateriale måtte 1/3 av pasientene avbryte NSAID-behandling på grunn av bivirkninger eller manglende effekt (25). Samme tilnærming er for eksempel også brukt i industrifinansierte opioidstudier, hvor man først gir opioider i et par uker og så ekskluderer alle pasienter som ikke tolererer midlet før selve studien starter (26).

En annen prosedyre i samme gate er å kreve at faste NSAID-brukere skal ha en smerteforverring over et visst minimum på VAS når de en kort periode (vanligvis inntil en uke) stopper med å ta NSAID i den såkalte "wash out"-perioden før en klinisk studie starter. Dette sikrer at studiepopulasjonen utelukkende befolkes med NSAID-"respondere" (27-30).

En tredje metode kan anvendes med legemidler som gir hyppige bivirkninger, og går ut på å bruke en såkalt "last-value-carried-forward"-prosedyre for å håndtere frafall i observasjonsperioden. Mange av pasientene i legemiddelstudier kan for eksempel oppleve en viss effekt initialt, men kommer så til et punkt hvor de må kutte legemiddelbruken fordi ulempene med bivirkninger blir dominerende. Det positive bidraget fra initialfasen vil i så fall telle i den statistiske analysen, på tross av at behandlingseffekten egentlig burde være lik null ved behandlingsslutt siden behandlingen måtte avbrytes.

Et fjerde grep er å begrense observasjonene til den perioden hvor legemiddelet er mest effektivt. Et eksempel illustreres av bruken av glukokortikosteroidinjeksjoner mot tennisalbue, som de første 6 uker gir signifikant smertelindring. Det var først etter publiseringen av to uavhengig finansierte studier med lengre observasjonstid at man ble klar over at pasienter som fikk kortisonsteroidinjeksjoner ble signifikant verre enn kontrollgruppen etter 6 og 12 måneder (31,32). Generelt er det en påfallende mangel på langtidsstudier av effekter av forskjellige terapeutiske intervensjoner ved kroniske tilstander.

Interessekonflikter i form av forskjellige samarbeidsrelasjoner med produsenter av behandlingene som evalueres er i seg selv assosiert med økt frekvens av positiv rapportering av den testede intervensjonen i studier (33,34). Dette kan bety at vi selv ved fravær av identifiserbare tvilsomme metoder eller grep bør være skeptiske til studier hvor opphavsmennene har hatt interesser ut over kunnskapsgenerasjon. I den sammenheng er ikke bare økonomiske interessekonflikter relevante; personlige, politiske og profesjonsmessige interessekonflikter kan også introdusere skjevheter. Eksempler på slike interessekonflikter inkluderer konkurransesituasjoner mellom forskningsmiljøer og situasjoner hvor en profesjon slår ring rundt behandlingsformer som de selv har eksklusive rettigheter i forhold til (f.eks. leger og legemiddelforskrivning). Vi ser heldigvis at bevisstheten rundt disse spørsmålene er økende, og eksklusjon av aktører med interessekonflikter er en indikator på kvalitet og seriøsitet i arbeid med kunnskapsoversikter og terapianbefalinger.

Bivirkninger

De fleste randomiserte, kontrollerte studier er styrkeberegnet til å vise positive effekter på utfall som ofte er relativt enkle å tallfeste, som for eksempel millimeter smertereduksjon på VAS. Studiene er sjelden formatert med tanke på bivirkninger, som ofte rapporteres som mer eller mindre tilfeldige tilleggsfunn, eller i form av kasuistikker. Effektmålene vil da ofte bli tilkjent en høy evidensstyrke, mens bivirkningsdata i form av kasuistikkserier vektes langt mindre (Tabell 1). Dessuten vanskeliggjør denne måten å samle informasjon på generering av kunnskap om intervensjonens reelle nytteverdi. Det finnes ingen "fast grense" for hva som tolereres av bivirkninger fra for eksempel et legemiddel; toleransegrensen er snarere en funksjon av midlenes yteevne og anvendelsesområde. Hvis et tenkt medikament kurerer en betydelig andel pasienter med en ellers letal kreftform, vil man måtte godta at midlet har omfattende og i noen tilfeller til og med dødelige bivirkninger – og dette avspeiles da også i måten cytostatikaterapi ved kreft drives på. Hvis midlet derimot har beskjedne og utelukkende symptomatiske effekter, slik tilfellet er med NSAIDs, vil de fleste ikke tolerere at midlet kan gi opphav til alvorlige bivirkninger, selv om disse opptrer sjelden. Bivirkningsdata som tillater en kostnad/nytte-vurdering av legemidlene ut fra slike kriterier foreligger sjelden før et legemiddel har vært på markedet i lengre tid.

Bivirkninger opptrer hyppigst hos pasienter som tidligere har erfart bivirkninger, hos barn og eldre (> 60 år), hos de som allerede er syke og hos kvinner. Dette gjenspeiles ikke i de populasjonene som brukes i kliniske forsøk, som typisk består av friske menn i alderen 20-60 år. Når studiepopulasjonen er en helt annen enn den som senere vil gjenfinnes som mottaker av behandlingen, og når studien i tillegg ikke er av et omfang som gir statistisk styrke på bivirkningssiden, burde det være innlysende at bruk av p-verdier i sikkerhetsvurderinger blir misvisende. Mange har ikke tatt dette inn over seg.

EKSEMPEL: META-ANALYSER AV NOEN BEHANDLINGSFORMER MOT USPESIFIKKE KORSRYGGSMERTER

For å belyse nytteverdien av meta-analyser og hvordan de utføres, tar vi utgangspunkt i retningslinjer for behandling av uspesifikke ryggsmerter. Både de amerikanske (35) og de tverrfaglige norske retningslinjene for behandling av uspesifikke korsryggsmerter (NOR) fra oktober 2007 (36) kan tjene som eksempler.

Innenfor ryggforskningen er vi privilegerte i den forstand at det finnes en lang rekke randomiserte kontrollerte studier av de relevante intervensjonene, slik at vurderingen av vitenskapelig dokumentasjon kan baseres på randomiserte kontrollerte studier eller systematiske oversikter. På denne bakgrunn skulle også forholdene ligge vel til rette for å utføre en systematisk oversikt med meta-analyser av de randomiserte kontrollerte studiene med korsryggsmerter. Imidlertid er denne vurderingsmetoden for vitenskapelig dokumentasjon valgt bort både i USA og i NOR. Begge har valgt en vurderingsform etter "overview"-modellen med kvalitativ analyse som åpner for subjektive og ikke-reproduserbare konklusjoner basert på konsensus. Etter vår mening burde begge heller tatt utgangspunkt i de meta-analysene det var mulig å gjøre.

I denne sammenhengen skal vi eksemplifisere dette ved å se på retningslinjenes vurderinger av NSAIDs ved akutte ryggsmerter og laserterapi ved langvarige og uspesifiserte ryggsmerter. Vi skal granske det som finnes av randomiserte, placebokontrollerte studier for disse to intervensjonene, meta-analysere data og sammenlikne det vi finner med retningslinjenes konklusjoner. Robustheten i meta-analysene vil også bli vurdert i lys av studier publisert etter retningslinjenes sluttdato for litteratursøk.

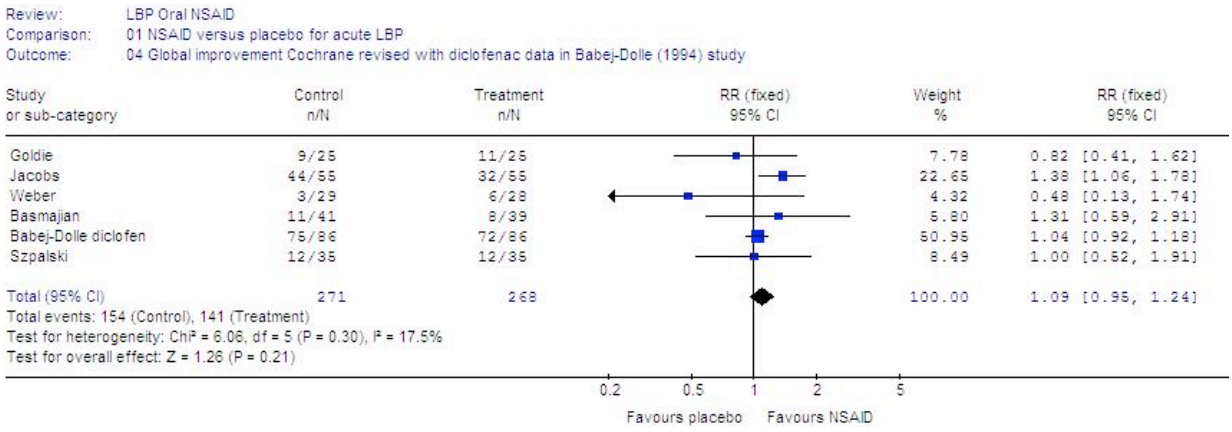
NSAIDs

De amerikanske og norske retningslinjene konkluderer begge med at tre systematiske oversikter viser moderat dokumentasjon for at NSAIDs reduserer akutte uspesifikke ryggsmerter i lett grad. Dette resulterte i følgende anbefalinger: "Det anbefales (……) å starte med paracetamol eventuelt med NSAIDs hvis paracetamol allerede er prøvd" (36) og "For most patients, first-line medication options are acetaminophen or nonsteroidal anti-inflammatory drugs" (35).

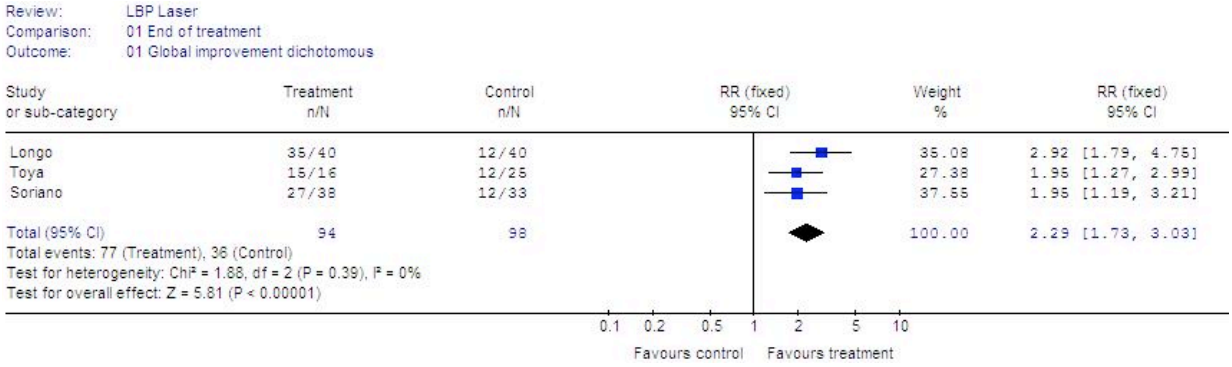
De tre systematiske oversiktene det vises til, er en serie utgaver av en Cochrane-oversikt som er trukket tilbake av opphavsmennene (37), og to påfølgende og svært like oversikter som i stor grad er basert på denne Cochrane-oversikten. Cochrane-oversikten inneholdt meta-analyser.

For kontinuerlige data fra tre studier som brukte ulike smerteskalaer, fant man ingen signifikante forskjeller mellom NSAIDs og placebo. For relativ risiko for forbedring viste meta-analyse av seks studier en liten effekt på 1,24 (95% konfidensintervall 1,10 til 1,41), og kun to av de seks tilgrunnliggende studiene viste signifikant smertelindring. Data fra meta-analysene viser altså at den smertelindrende effekten av NSAIDs er lik placebo eller i beste fall liten.

Når resultatene av enkeltstudier spriker, bør man lete etter årsaker til denne variasjonen: Den studien som er vektet tyngst (38) i meta-analysen er blitt gjenstand for en inklusjonsfeil, idet man har brukt resultatene fra en gruppe som ble behandlet med et obsolet smertemiddel som heter dipyron i stedet for en gruppe behandlet med NSAID-midlet diklofenak, som for øvrig ikke var signifikant forskjellig fra placebo. Når vi korrigerer for denne feilen og rekalkulerer resultatene, blir heller ikke utfallsmålet i meta-analysen for relativ risiko signifikant forskjellig fra placebo (Figur 1). Meta-analysen viser effekten av NSAID ved akutte



Figur 1. Gjennomsnittlig behandlingseffekt med 95% konfidensintervall av NSAID ved akutte uspesifikke korsryggsmerter.



Figur 2. Gjennomsnittlig behandlingseffekt ved behandlingsslutt målt som angivelse av global forbedring med 95% konfidensintervall for laserbehandling ved langvarige uspesifikke korsryggsmerter.

ryggsmerter, målt som relativ risiko for forbedring. Siden det samlete resultatet (svart rombe nederst i figuren) krysser midtlinjen, er det ingen signifikant effekt av NSAID (relativ risiko for forbedring: 1,09 [95% KI 0,96–1,23], p=0,19).

Resultatene fra meta-analysene er såpass konsistente i negativ retning at det etter vår mening ville være lite sannsynlig at fremtidig forskning skulle vise noe annet. Dette er nå også bekreftet av en nylig publisert høykvalitets-studie i Lancet, som heller ikke fant noen signifikant effekt av NSAID ved akutte uspesifikke ryggsmerter (39).

Laserbehandling av kroniske uspesifiserte ryggsmerter

Om laserbehandling skriver NOR at "de europeiske retningslinjene konkluderer med at det både ved akutte og langvarige korsryggsmerter foreligger få studier på disse modalitetene og at de gjennomgående er av lav metodisk kvalitet" (36), og de amerikanske retningslinjene er omtrent likelydende (35). Under "anbefalinger" står det at "foreliggende dokumentasjon ikke gir grunnlag for anbefaling av laserbehandling" mens amerikanerne skriver at "there is insufficient evidence to recommend low-level laser therapy". Det er ellers verdt å merke seg at ingen av laserstudiene var industrisponsede. Ingen av retningslinjene legger meta-analyser til grunn for de manglende anbefalingene, men de amerikanske retningslinjene hevdet at man var "unable to estimate net benefit" av denne intervensjonen (35). I motsetning til ordlyden i konklusjonen, skriver de ellers at 3 av 4 positive studier holdt høy kvalitet, mens kvaliteten også var høy for den ene studien som viste et negativt resultat av laserterapi. Det begge retningslinjene unnlater å nevne, er at stråledosen i den ene negative laserstudien var 1/10 eller mindre enn i de øvrige studiene. For NORs del er det uklart hvor mange studier som inngår i den skjønnsmessige konsensusvurderingen av dokumentasjonen.

Vi har brukt dette siste eksempelet for å vise hvordan de skjønnsmessige konklusjonene og anbefalingene står i sterk kontrast til det foreliggende tallmaterialet. Påstanden i de amerikanske retningslinjene om at det ikke var mulig å estimere effekt av laserbehandling er ikke korrekt. Meta-analyse viser en meget god effekt av laserbehandling med relativ risiko for forbedring lik 2,3 [95% KI: 1,7–3,0, p=0,00001] (Figur 2). Denne meta-analysen viser en meget god samleeffekt av laserbehandling ved langvarig uspesifikke rygg-

merter (svart rombe nederst er plassert godt til høyre for midtlinjen).

Med forbehold om at materialet er begrenset (n=192), foreligger det altså data for positiv effekt av laserbehandling ved denne type smerter. Dette støttes – på samme måte som fravær av positive effekter av NSAIDs – av nytilkommet forskning, som viser signifikant positive og vedvarende effekter av laserbehandling på smerte og fysisk funksjon (40). Dette samsvarer også med eksperimentelle studier, som har vist at laserbehandling har en antiinflammatorisk virkningsmekanisme i laboratorieforsøk (41,42).

OPPSUMMERING

Vitenskapelig dokumentasjon for terapeutiske intervensjoner kan evalueres på en systematisk måte. I det ovenstående har vi vist hvordan bruk av meta-analyser kan ha en praktisk nytte ved å øke presisjonen for gradering av vitenskapelig dokumentasjon i utarbeidelsen av retningslinjer.

Eksemplene fra ryggretningslinjene viser at konsensusprosesser med utspring i grupper av eksperter kan gi resultater som avviker betydelig fra meta-analyser. Resultatene i meta-analyser fremkommer gjennom fastlagte prosedyrer og statistiske analyser som i liten grad har ytre påvirkning, og prosessen er transparent med potensielle svakheter som vesentlig er knyttet til eventuell inklusjon eller eksklusjon av studier.

Det er tankevekkende at retningslinjene for behandling av ryggsmerter når det gjelder disse og andre behandlingsalternativer går på tvers av kvantitative data fra meta-analyser. Fenomenet er imidlertid utbredt; de tidligere nevnte eksemplene på vurderingen av Alzheimer-midler (13,14) og omega-3-fettsyrer (15, 16) føyer seg begge inn i rekken av eksempler på at meta-analyser og ekspertråd kan gi ulike konklusjoner. Oversiktsmodellen gir med sine omfattende litteratursøk og en tilsynelatende rigorøs metodologi et inntrykk av vitenskapelighet, men er i realiteten en upresis og lite transparent metode. Meta-analyser representerer etter vårt skjønn en sikrere og mer reproduserbar måte å evaluere vitenskapelig dokumentasjon, effektstørrelse og klinisk relevans på enn oversiktsmodeller. Sistnevnte framstår ofte som ikke-transparente når det gjelder utvelgelse av deltakere og oppviser ofte svakheter både i grunnlaget for anbefalingene og gjennom nærvær av potensielle interessekonflikter.

REFERANSER

1. Lind J. A treatise of the scurvy in three parts. Containing an inquiry into the nature, causes and cure of that disease, together with a critical and chronological view of what has been published on the subject. Edinburgh: Printed by Sands, Murray and Cochran for A. Kincaid and A. Donaldson, 1753.
2. Evans I, Thornton H, Chalmers I. Testing treatments – better research for better healthcare. The British Library, 2006.
3. Withering W. An account of the foxglove and some of its medical uses: With practical remarks on dropsy and other diseases. London: Robinson, 1785.
4. Li JJ. Laughing gas, Viagra and Lipitor: the human stories behind the drugs we use. New York: Oxford University Press, 2006.
5. Bostrøm H, Dahlgren H, red. Placebo. Stockholm, I samarbete med Statens beredning för utvärdering av medicinsk metodik (SBU), 2000.
6. Stephens T, Brynner R. Dark Remedy. The impact of thalidomide and its revival as a vital medicine. Cambridge, MA: Perseus Publishing, 2001.
7. Wilson RA. Feminine forever. New York: Evans, 1966.
8. Avorn J. Powerful medicines. The benefits, risks, and costs of prescription drugs. New York: Random House Publishers, 2004.
9. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA* 2002; **288** (3): 321-33.
10. Skurtveit S, Furu K, Slordal L. Use of selective COX-2-inhibitors in Norway before and after the withdrawal of rofecoxib – Norwegian prescription database. *Pharmacoepidemiol Drug Saf* 2006; **15**: S316-7.
11. Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* 2005; **365** (9458): 475-81.
12. Roland PD, Bjordal JM, Klovning A, Slordal L. Glukosamin – den store sukkerpillebløffen. *Tidsskr Nor Lægeforen* 2007; **127** (16): 2121-2.
13. Kaduszkiewicz H, Zimmermann T, Beck-Bornholdt HP, van den Bussche H. Cholinesterase inhibitors for patients with Alzheimer's disease: systematic review of randomised clinical trials. *BMJ* 2005; **331** (7512): 321-7.
14. Henriksen K. «Alle» er enige om effekten. *Dagens Medisin* 2006 (19). Available from: http://www.dagensmedisin.no/nyheter/2006/11/10/n-alle--er-enige-om-effekten/.

15. Hooper L, Thompson RL, Harrison RA, Summerbell CD, Ness AR, Moore HJ, et al. Risks and benefits of omega 3 fats for mortality, cardiovascular disease, and cancer: systematic review. *BMJ* 2006; **332** (7544): 752-60.
16. Landmark K, Aursnes I, Reikvam A, Alm CS. Omega-3-fettsyrer er fortsatt gunstig ved hjertesykdom. *Tidsskr Nor Lægeforen* 2007; **127** (2): 202-3.
17. Reikvam A, Hexeberg S, Kvien TK, Slordal L, Aabakken L, Engebretsen L, et al. Klinisk bruk av COX-hemmere – en konsensus. *Tidsskr Nor Lægeforen* 2006; **126** (5): 591-5.
18. Bjordal JM, Ljunggren AE, Klovning A, Slordal L. Non-steroidal anti-inflammatory drugs, including cyclo-oxygenase-2 inhibitors, in osteoarthritic knee pain: meta-analysis of randomised placebo controlled trials. *BMJ* 2004; **329** 1317-22.
19. Bjordal JM, Klovning A, Ljunggren AE, Slordal L. Short-term efficacy of pharmacotherapeutic interventions in osteoarthritic knee pain: A meta-analysis of randomised placebo-controlled trials. *Eur J Pain* 2007; **11** (2): 125-38.
20. Wiebe N, Vandermeer B, Platt RW, Klassen TP, Moher D, Barrowman NJ. A systematic review identifies a lack of standardization in methods for handling missing variance data. *J Clin Epidemiol* 2006; **59** (4): 342-53.
21. Silverstein FE, Faich G, Goldstein JL, Simon LS, Pincus T, Whelton A, et al. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: A randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. *JAMA* 2000; **284** (10): 1247-55.
22. Juni P, Rutjes AW, Dieppe PA. Are selective COX 2 inhibitors superior to traditional non steroidal anti-inflammatory drugs? *BMJ* 2002; **324** (7349): 1287-8.
23. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *BMJ* 2003; **326** (7400): 1171-3.
24. Vlad SC, LaValley MP, McAlindon TE, Felson DT. Glucosamine for pain in osteoarthritis: why do trial results differ? *Arthritis Rheum* 2007; **56** (7): 2267-77.
25. Arboleya LR, de la Figuera E, Soledad Garcia M, Aragon B. Management pattern for patients with osteoarthritis treated with traditional non-steroidal anti-inflammatory drugs in Spain prior to introduction of Coxibs. *Curr Med Res Opin* 2003; **19** (4): 278-87.
26. Schnitzer TJ, Gray WL, Paster RZ, Kamin M. Efficacy of tramadol in treatment of chronic low back pain. *J Rheumatol* 2000; **27** (3): 772-8.
27. Birbara CA, Puopolo AD, Munoz DR, Sheldon EA, Mangione A, Bohidar NR, et al. Treatment of chronic low back pain with etoricoxib, a new cyclo-oxygenase-2 selective inhibitor: improvement in pain and disability – a randomized, placebo-controlled, 3-month trial. *J Pain* 2003; **4** (6): 307-15.
28. Coats TL, Borenstein DG, Nangia NK, Brown MT. Effects of valdecoxib in the treatment of chronic low back pain: results of a randomized, placebo-controlled trial. *Clin Ther* 2004; **26** (8): 1249-60.
29. Katz N, Ju WD, Krupa DA, Sperling RS, Bozalis Rodgers D, Gertz BJ, et al. Efficacy and safety of rofecoxib in patients with chronic low back pain: results from two 4-week, randomized, placebo-controlled, parallel-group, double-blind trials. *Spine* 2003; **28** (9): 851-8; discussion 9.
30. Pallay RM, Seger W, Adler JL, Ettlinger RE, Quaidoo EA, Lipetz R, et al. Etoricoxib reduced pain and disability and improved quality of life in patients with chronic low back pain: a 3 month, randomized, controlled trial. *Scand J Rheumatol* 2004; **33** (4): 257-66.
31. Smidt N, Lewis M, DA VDW, Hay EM, Bouter LM, Croft P. Lateral epicondylitis in general practice: course and prognostic indicators of outcome. *J Rheumatol* 2006; **33** (10): 2053-9.
32. Smidt N, Lewis M, Hay EM, Van der Windt DA, Bouter LM, Croft P. A comparison of two primary care trials on tennis elbow: issues of external validity. *Ann Rheum Dis* 2005; **64** (10): 1406-9.
33. Kjaergard LL, Als-Nielsen B. Association between competing interests and authors' conclusions: epidemiological study of randomised clinical trials published in the BMJ. *BMJ* 2002; **325** (7358): 249.
34. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003; **326** (7400): 1167-70.
35. Chou R, Qaseem A, Snow V, Casey D, Cross JT, Jr., Shekelle P, et al. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. *Ann Intern Med* 2007; **147** (7): 478-91.
36. Lærum E, Brox JI, Storheim K, Espeland A, Haldorsen E, Munch-Ellingsen J, et al. Nasjonale kliniske retningslinjer. Korsryggsmerter – med og uten nerverotaffeksjon. Oslo: FORMI, Formidlingsenheten for muskel- og skjelettlidelser/Sosial- og helsedirektoratet, 2007.
37. van Tulder MW, Scholten RJ, Koes BW, Deyo RA. WITHDRAWN: Non-steroidal anti-inflammatory drugs for low-back pain. *Cochrane Database Syst Rev* 2006; (2): CD000396.

38. Babej-Dolle R, Freytag S, Eckmeyer J, Zerle G, Schinzel S, Schmeider G, et al. Parenteral dipyrone versus diclofenac and placebo in patients with acute lumbago or sciatic pain: randomized observer-blind multicenter study. *Int J Clin Pharmacol Ther* 1994; **32** (4): 204-9.
39. Hancock MJ, Maher CG, Latimer J, McLachlan AJ, Cooper CW, Day RO, et al. Assessment of diclofenac or spinal manipulative therapy, or both, in addition to recommended first-line treatment for acute low back pain: a randomised controlled trial. *Lancet* 2007; **370** (9599): 1638-43.
40. Djavid GE, Mehrdad R, Ghasemi M, Hasan-Zadeh H, Sotoodeh-Manesh A, Pouryaghoub G. In chronic low back pain, low level laser therapy combined with exercise is more beneficial than exercise alone in the long term: a randomised trial. *Aust J Physiother* 2007; **53** (3): 155-60.
41. Albertini R, Aimbire FS, Correa FI, Ribeiro W, Cogo JC, Antunes E, et al. Effects of different protocol doses of low power gallium-aluminum-arsenate (Ga-Al-As) laser radiation (650 nm) on carrageenan induced rat paw ooedema. *J Photochem Photobiol B* 2004; **74** (2-3): 101-7.
42. Gur A, Karakoc M, Cevik R, Nas K, Sarac AJ. Efficacy of low power laser therapy and exercise on pain and functions in chronic low back pain. *Lasers Surg Med* 2003; **32** (3): 233-8.