preklinische en klinische tests zijn uitgevoerd. Van belang is dat zulke tests zijn uitgevoerd voor alle bekende faalscenario's.

Met dank aan prof.dr.T.J.J.H.Slooff, orthopedisch chirurg, voor figuur 1.

abstract

Causes of failure of hip and knee arthroplasties

– Most joint replacement prostheses in the hip or knee survive for longer than 10 years. Failure of a prosthesis may be due to infection or a fracture of bone or prosthesis, but much more often it has no clear cause, after a process of aseptic loosening.

– Mechanisms of aseptic loosening are: bone resorption as a reaction of bone to migrated foreign body particles, material fatigue due to repeated mechanical stress on the prosthesis and its connections with the bone, failed ingrowth due to a poor fit of the (uncemented) prosthesis in the bone bed, resorption of bone mass due to the prosthesis taking over part of the mechanical bone stresses, and wear of the material.

– The risk of early failure depends on the patient's bone quality and life expectation, on the surgical technique, the weight bearing on the prosthesis and the fixation, and the shape and materials of the prostheses. Selection of durable prosthetic types and techniques is being done more and more by means of stepwise introduction, (pre)clinical tests and post-marketing surveillance.

literatuur

1 Malchau H. On the importance of stepwise introduction of new hip implant technology [dissertation]. Göteborg, Zweden, 1995.
2 Huiskes R. Failed innovation in total hip replacement. Diagnosis and proposal for a cure. Acta Orthop Scand 1993;64:699-716.
3 Huiskes R, Verdonschot N. Failure scenario's and the innovation cycle. In: Callaghan JJ, Rosenberg AG, Rubash HE, editors. The adult hip. Philadelphia: Lippincott-Raven, 1997.
4 Verdonschot N, Huiskes R. The effects of cement-stem debonding in THA on the long-term failure probability of cement. J Biomech 1997;30:795-802.
5 Ryd L. Micromotion in knee arthroplasty. A roentgen stereophotogrammetric analysis of tibial component fixation. Acta Orthop Scand 1986;220 Suppl:1-80.
6 Huiskes R, Rietbergen B van. Preclinical testing of total hip stems. The effects of coating placement. Clin Orthop 1995;319:64-76.
7 Faro LM, Huiskes R. Quality assurance of joint replacement. Legal regulation and medical judgement. Acta Orthop Scand 1992;250 Suppl:1-33.
8 Karrholm J. Roentgen stereophotogrammetry. Review of orthopedic applications. Acta Orthop Scand 1989;60:491-503.
9 Ryd L. Roentgen stereophotogrammetric analysis of prosthetic fixation in the hip and knee joint. Clin Orthop 1992;276:56-65.

Aanvaard op 12 september 1997

Voor de praktijk

Dwalingen in de methodologie. VII. Reproduceerbaarheid van metingen

h.c.w.de vet en a.j.h.m.beurskens

Reproduceerbaarheid van metingen is een belangrijk onderwerp in de geneeskunde. Met reproduceerbaarheid wordt bedoeld dat bij herhaalde metingen dezelfde uitkomst wordt gevonden.[1] Als één persoon twee metingen vlak na elkaar uitvoert en als die verschillende uitkomsten opleveren (intrabeoordelaarsvariatie), dan zijn daar diverse verklaringen voor. De discrepantie kan komen, doordat de arts de eerste keer anders keek dan de tweede of doordat hij hetgeen hij zag anders interpreteerde. Als twee verschillende artsen de metingen uitvoeren, is de kans dat de uitkomst verschilt nog groter (interbeoordelaarsvariatie). Het kan ook zijn dat het meetinstrument bij de tweede meting een andere uitslag geeft (meetfouten van het instrument). Tenslotte kunnen de patiëntkenmerken die gemeten worden tussentijds veranderd zijn (biologische variatie).

In dit artikel bespreken wij enkele veelgebruikte en geschikte maten om de reproduceerbaarheid te kwantificeren.[1] Wij kijken naar metingen waarbij patiënten worden ingedeeld in categorieën, bijvoorbeeld op basis van de aan- of afwezigheid van kraakbeenerosies van de handgewrichten bij reumapatiënten of op basis van de ernst van de erosies. Daarna beschrijven wij hoe de reproduceerbaarheid van een continue variabele, bijvoorbeeld de grijpkracht, bij deze patiënten bepaald kan worden.

samenvatting

– Reproduceerbaarheidsmetingen zijn belangrijk om medische gegevens te interpreteren.

– Om de reproduceerbaarheid van categoriale variabelen te bepalen is kappa de geschiktste maat. Kappa meet de overeenkomst, gecorrigeerd voor de toevalsovereenkomst. De interpretatie is niet simpel. De kappawaarde wordt namelijk beïnvloed door het aantal categorieën waarin ingedeeld moet worden en de prevalentie van de scores van de beoordelaars.

– Voor continue variabelen is de correlatiecoëfficiënt van Pearson goed bruikbaar, mits men zich realiseert dat deze voorbijgaat aan systematische fouten en sterk afhankelijk is van de heterogeniteit van de gegevens.

– Voor continue variabelen kunnen ook de grenzen van overeenkomst vastgesteld worden. Deze methode is geschikt om systematische verschillen en toevalsfouten te onderscheiden en om de grootte van de verschillen te kwantificeren.

– In het algemeen geldt dat voor een goede interpretatie van de diverse reproduceerbaarheidsmaten een visuele presentatie van de gegevens in de vorm van een tabel of een figuur een duidelijke meerwaarde heeft.

Universiteit Maastricht, vakgroep Epidemiologie, Postbus 616, 6200 MD Maastricht.
Mw.dr.ir.H.C.W.de Vet en mw.dr.A.J.H.M.Beurskens, epidemiologen.
Correspondentieadres: mw.dr.ir.H.C.W.de Vet.

reproduceerbaarheid van categoriale gegevens

Kappa is een maat voor de reproduceerbaarheid van categoriale gegevens (gegevens die in categorieën kunnen worden ingedeeld (bijvoorbeeld ziek en niet-ziek) in tegenstelling tot continue gegevens, zoals lichaamsgewicht). Als twee reumatologen 100 röntgenfoto's van de hand beoordelen op de aan- of afwezigheid van kraakbeenerosies, kunnen de resultaten weergegeven worden in een 2×2-tabel (tabel).[2] Eenzelfde tabel kan totstandkomen als één reumatoloog de foto's 2 keer beoordeelt. In de tabel is te zien dat er in 75 van de 100 gevallen overeenkomst is: voor 50% van de gevallen oordelen beiden positief, voor 25% beiden negatief. Deze percentageovereenkomst houdt geen rekening met de toevalsovereenkomst. Zelfs als de tweede reumatoloog met zijn ogen dicht had gescoord, zou hij in een aantal gevallen hetzelfde antwoord hebben gegeven als de eerste. De coëfficiënt kappa (κ) houdt hier wel rekening mee.[3] Kappa geeft de mate van extra overeenkomst boven de toevalsovereenkomst, als fractie van wat maximaal aan overeenkomst tussen de reumatologen haalbaar zou zijn. In de tabel is het percentage geobserveerde overeenkomst (p_o) 75. Het percentage toevalsovereenkomst (p_e) wordt berekend op basis van de randtotalen van beide reumatologen: $65/100 \times 60 = 39\%$ van de röntgenfoto's zou door beiden alleen al op grond van het toeval positief gescoord worden, en $35/100 \times 40 = 14\%$ negatief. Het totale percentage toevalsovereenkomst is dus 53%. Dat betekent ook dat er maar een ruimte van 47% (100% – 53%) beschikbaar is voor de maximale overeenkomst die bereikt kan worden door goed te kijken. De waarde van kappa wordt in dit geval:

$$\kappa = (p_o - p_e)/(1 - p_e) = (75\% - 53\%)/(100\% - 53\%) = 0{,}47.$$

De waarden van κ liggen normaliter tussen 0 en +1. Bij een perfecte overeenkomst is κ gelijk aan 1, de waarde 0 betekent dat er niet meer overeenkomst is dan op grond van toeval te verwachten is.

De κ-coëfficiënt kan ook toegepast worden voor observaties waarbij meer dan twee uitkomsten mogelijk zijn.[3] In plaats van de aan- of afwezigheid van erosies kan men ook de ernst van de erosies beoordelen. Naarmate het aantal categorieën van ernst dat men wil onderscheiden toeneemt, wordt het moeilijker om twee beoordelingen in precies dezelfde categorie te krijgen. De κ-waarde zal dan in het algemeen lager zijn.

In genoemd voorbeeld waarbij de meetschaal ordinaal is (meer klassen met een logische volgorde), kan een gewogen kappa (κ_w) berekend worden.[4] Daarbij worden misclassificaties tussen aan elkaar grenzende categorieën minder zwaar meegeteld dan fouten tussen categorieën die verder van elkaar liggen.

Interpretatieproblemen met κ. Kappa is een algemeen geaccepteerde maat voor de reproduceerbaarheid van categoriale variabelen. In de literatuur wordt aangegeven bij welke waarden van κ een overeenkomst slecht, matig, redelijk of goed genoemd mag worden.[5] Toch zitten er aan de interpretatie van κ-waarden veel haken en ogen.

Allereerst zagen wij dat κ lager wordt naarmate er meer categorieën gebruikt worden. Daarnaast is κ afhankelijk van de prevalentie van het verschijnsel dat gemeten wordt. Hoge prevalenties zorgen voor een hoge toevalsovereenkomst, waardoor weinig ruimte voor extra overeenkomst (te behalen door de deskundige beoordeling) overblijft. Stel dat in de tabel beide reumatologen 80% van de röntgenfoto's positief hadden gescoord (in plaats van 65 en 60%), dan zou het percentage verwachte overeenkomst op grond van het toeval (p_e) 64% + 4% = 68% zijn geweest. Uitgaande van dezelfde waargenomen overeenkomst (p_o) van 75% wordt de κ-waarde dan 0,22. De κ-waarde alleen zegt dus niet zoveel. Presentatie van de hele tabel geeft meer inzicht in de reproduceerbaarheid. Daaruit wordt duidelijk om hoeveel categorieën het gaat, wat de prevalenties van de scores zijn en dus hoe hoog de toevalsovereenkomst is en, in geval van meer categorieën, welke categorieën het moeilijkst onderscheiden worden.

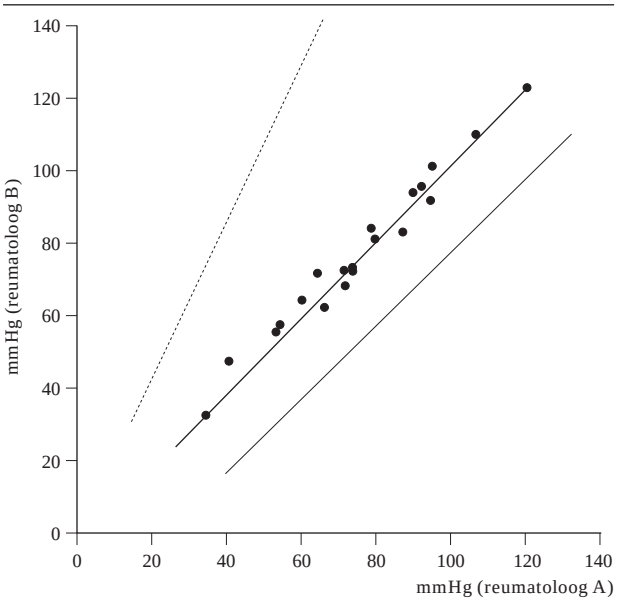
Overeenkomst tussen twee reumatologen die röntgenfoto's van de handen van 100 patiënten beoordeelden op de aan- of afwezigheid van kraakbeenerosies[2]

	reumatoloog B: erosies		
	aanwezig	*afwezig*	*totaal*
reumatoloog A: erosies			
aanwezig	50	15	65
afwezig	10	25	35
totaal	60	40	100

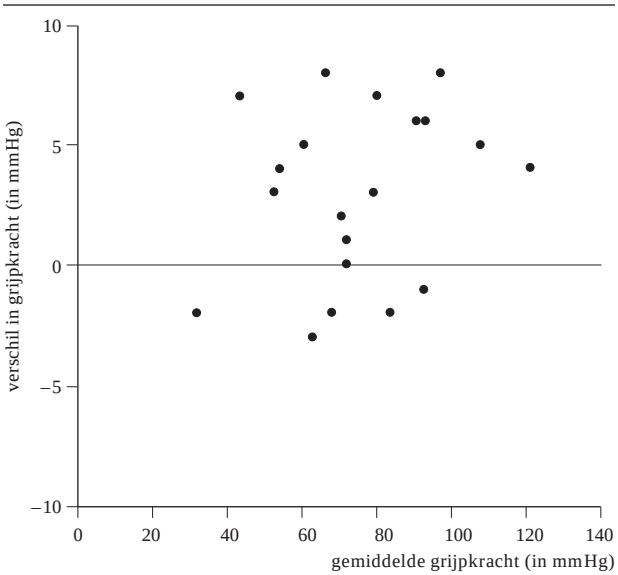
bepaling van de reproduceerbaarheid van continue gegevens

Correlatiecoëfficiënt van Pearson. Om de reproduceerbaarheid van gegevens vast te stellen die gemeten zijn op een continue schaal, wordt vaak gebruikgemaakt van de correlatiecoëfficiënt van Pearson (r).[1] Deze kan waarden tussen 0 en 1 aannemen. Het voordeel van r is dat hij makkelijk uit te rekenen is, maar hij heeft twee belangrijke nadelen. Ten eerste houdt hij geen rekening met systematische afwijkingen tussen twee beoordelingen. In figuur 1 zijn enkele voorbeelden gegeven van grijpkrachtmetingen. In alle gevallen is r nagenoeg gelijk aan 1: de punten liggen vlak bij een rechte lijn. Dit geldt natuurlijk voor de middelste lijn waar beoordelaar A en B ongeveer dezelfde waarde meten, maar dit gaat ook op als beoordelaar B steeds 2 keer zo hoge waarden meet (gestippelde lijn) als beoordelaar A, en eveneens als beoordelaar B systematisch steeds 10 eenheden minder meet dan beoordelaar A. De correlatiecoëfficiënt geeft dus alleen aan in hoeverre er een lineaire relatie bestaat.

Het tweede nadeel van r is dat hij sterk afhankelijk is van de spreiding van de uitkomsten, met andere woorden van de heterogeniteit van de populatie. Uitschieters hebben veel invloed op de waarde van r. Daarom is het zinvol om de gegevens waarop r berekend is in een figuur te zien. Men moet er altijd op bedacht zijn dat de r die bereikt werd in een bepaalde populatie niet automatisch geldt voor een andere populatie.



figuur 1. Grijpkrachtmetingen bij 20 patiënten uitgevoerd door 2 onafhankelijke reumatologen (A en B);[2] de grijpkracht wordt aangegeven in mmHg. De waarnemingspunten liggen bijna op een rechte lijn, de correlatiecoëfficiënt is nagenoeg gelijk aan 1. Dat is eveneens het geval als beoordelaar B steeds 2 keer zo hoge waarden meet als beoordelaar A (gestippelde lijn) en eveneens als beoordelaar B systematisch steeds 10 eenheden minder meet dan beoordelaar A (naar rechts verschoven doorgetrokken lijn).



figuur 2. Verschil in grijpkracht bij 20 patiënten gemeten door 2 onafhankelijke reumatologen (reumatoloog A en B; het verschil is B – A), uitgezet tegen het gemiddelde van de 2 metingen per patiënt ((A + B)/2).[2]

Intraklassecorrelatiecoëfficiënt. Een andere maat die soms gebruikt wordt in plaats van de r van Pearson is de intraklassecorrelatiecoëfficiënt (ICC).[6] Deze ondervangt het eerstgenoemde nadeel van r. De ICC bereikt namelijk alleen zijn maximale waarde (dat is 1) als de metingen van beoordelaar A en B precies overeenkomen (de middelste lijn in figuur 1). Van het tweede nadeel, de afhankelijkheid van de spreiding van de uitkomsten, heeft de ICC net zoveel last als r. Het grote nadeel van de ICC is dat de berekening erg ingewikkeld is.

Omdat het bij reproduceerbaarheid overwegend om toevalsfouten gaat en meestal niet om systematische fouten, heeft de r van Pearson vaak de voorkeur boven de ICC. De r is dus een goede maat om de reproduceerbaarheid te kwantificeren in het geval dat er geen systematische verschillen verwacht worden.

Grenzen van overeenkomst. Als de reproduceerbaarheid niet optimaal is, is de volgende vraag of de gevonden verschillen acceptabel zijn in de geneeskundige praktijk of het geneeskundig onderzoek. In de grootte van de verschillen geeft de r van Pearson slecht inzicht. Bland en Altman hebben een methode ontwikkeld die dat inzicht wel levert.[7] Daartoe worden de verschillen (d) tussen de waarnemingen van twee beoordelaars per patiënt uitgezet tegen de gemiddelde waarneming bij die patiënt. Dit resulteert in figuur 2. Ongeveer 95% van de verschillen zal tussen d – 2SD en d + 2SD liggen (SD = standaarddeviatie). Deze grenzen worden grenzen van overeenkomst ('limits of agreement') genoemd. Bij deze methode worden de verschillen in overeenkomst uitgedrukt in dezelfde dimensie als de metingen. Dat maakt ze beter interpreteerbaar. In figuur 2 kan men zich afvragen of bijvoorbeeld 3,6 mmHg bij dergelijke metingen een acceptabele afwijking is. Welke verschillen nog acceptabel zijn, hangt af van de klinische toepassing. Daar is geen statistische toets voor, dat is een klinische beoordeling gebaseerd op medisch gezond verstand.

klinische relevantie

Reproduceerbaarheidsmetingen zijn belangrijk in de geneeskunde. Inzicht in de reproduceerbaarheid is essentieel voor de interpretatie van klinische gegevens, zowel in het wetenschappelijk onderzoek als in de medische praktijk. Voor dat doel is de berekening van een enkele coëfficiënt meestal onvoldoende en verdient een visuele presentatie, in tabellen of figuren, de voorkeur. Strategieën om de reproduceerbaarheid te verhogen kunnen worden gezocht in de standaardisatie van metingen en in consensusbijeenkomsten met beoordelaars. Als de reproduceerbaarheid daardoor niet toeneemt, kan herhaling van metingen uitkomst bieden, hetzij door dezelfde beoordelaar, hetzij door verschillende beoordelaars. Verbetering van de reproduceerbaarheid van klinische metingen kan een belangrijke bijdrage leveren aan de kwaliteit van de geneeskunde.

abstract

Roaming through methodology. VII. Reproducibility of measurements

– Reproducibility measurements are important for a proper interpretation of medical data.

– Kappa is the most adequate measure for categorical variables. Kappa adjusts the observed agreement for chance agreement. The interpretation of kappa is rather difficult. The

kappa value is influenced by the number of categories used for classification and the prevalence of scores of the observers.
– For continuous variables the Pearson correlation coefficient can be used, keeping in mind its ignoring systematic errors and its dependence on the heterogeneity of the data.
– Another method to assess reproducibility for continuous variables is the method of limits of agreement. This method distinguishes systematic and random errors, and quantifies the differences in the dimension of the measurements.
– In general, the interpretation of the various measures of agreement is helped by a visual presentation of the data in a table or figure.

literatuur
[1] Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Oxford: Oxford University Press, 1995.
[2] Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304:1491-4.
[3] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
[4] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-20.
[5] Fleiss JL. Statistical methods for rates and proportions. 2nd ed. Wiley series in probability and mathematical statistics. New York: Wiley, 1981.
[6] Shrout PE, Fleiss JL. Intra class correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
[7] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307-10.

Aanvaard op 23 januari 1998

Oorspronkelijke stukken

Het aantal carotisoperaties in het Medisch Spectrum Twente vóór en na de publicatie van belangrijk internationaal onderzoek naar het nut van A.-carotisdesobstructie

l.m.dijkema, r.h.geelkerken, p.j.a.m.brouwers, p.de smit en r.j.van det

samenvatting
Doel. Het effect nagaan van goed wetenschappelijk onderzoek op de behandeling van de extracraniële stenose van de A. carotis interna.
Opzet. Retrospectief en vergelijkend.
Plaats. Medisch Spectrum Twente, Enschede.
Methode. De relevante gegevens van 2 jaar carotischirurgie vóór (1989-1990; periode I) en na de publicatie van twee gerandomiseerde multicentrische onderzoeken (1994-1995; periode II) werden met elkaar vergeleken.
Resultaten. Het aantal operatief behandelde patiënten en het aantal carotisdesobstructies was in periode II met respectievelijk 339 en 319% toegenomen. In periode I had 25% van de patiënten een asymptomatische ipsilaterale stenose van de A. carotis interna; in periode II was dit gedaald tot 11%. In periode I had 65% van de patiënten een stenose groter dan 70% van de vatdiameter; in periode II bedroeg dit 85%. De gecombineerde sterfte en blijvende invaliderende morbiditeit na 30 dagen was in periode I 6% en in periode II 3%.
Conclusie. Na de publicatie van twee hooggekwalificeerde onderzoeken in 1991 nam het aantal carotisoperaties met meer dan 300% toe. De indicaties voor de operatieve behandeling van de stenose waren in periode II eerder strikter dan ruimer geworden. De toename van het aantal carotisdesobstructies kan verklaard worden doordat huisarts en neuroloog anders naar de vaatchirurg zijn gaan verwijzen. Deze verandering van het verwijspatroon kan het gevolg zijn van het toepassen van 'evidence-based' geneeskunde.

Wetenschappelijk onderbouwde geneeskunde ('evidence-based medicine') wordt gepropageerd als instrument waarmee beslissingen in de gezondheidszorg rationeel worden ondersteund.[1] In een interview hield ook de minister van Volksgezondheid, Welzijn en Sport een warm pleidooi voor evidence-based geneeskunde. Dit zou moeten leiden tot een verhoging van de kwaliteit van het medisch handelen en tot efficiëntere en goedkopere geneeskunde. Zonder richtlijnen zou de geneeskunde kunnen ontsporen door overschatting van de eigen vermogens. Als voorbeeld werd de carotischirurgie genoemd.[2]

Evidence-based geneeskunde betekent het werken volgens richtlijnen die zijn gebaseerd op goed wetenschappelijk onderzoek.[3] In de jaren tachtig stond het nut van een operatieve interventie bij een symptomatische stenose van de A. carotis interna ter discussie. Het ontbrak aan gedegen onderzoek naar het effect van een carotisdesobstructie.[4] Begin jaren negentig zijn de resultaten bekend geworden van 2 grote onderzoeken, de 'North American symptomatic carotid endarterectomy trial' (NASCET) en de 'European carotid surgery trial' (ECST),[5 6] die de conservatieve en de operatieve behandeling met elkaar vergeleken bij patiënten met een symptomatische stenose van het extracraniële deel van

Medisch Spectrum Twente, Postbus 50.000, 7500 KA Enschede.
Afd. Vaatchirurgie: mw.L.M.Dijkema, co-assistent (thans: assistent-geneeskundige, Ziekenhuis De Weezenlanden, afd. Thoraxanesthesiologie en Intensive Care, Zwolle); dr.R.H.Geelkerken, dr.P.de Smit en R.J.van Det, vaatchirurgen.
Afd. Neurologie: dr.P.J.A.M.Brouwers, neuroloog.
Correspondentieadres: dr.R.H.Geelkerken.