

I artikkelen beskrives en studie der målet var å finne frem til reliable og valide instrumenter for måling av skulderfunksjon.

Testing av skulderfunksjon hos pasienter med proksimal humerus fraktur

Reliabilitets- og validitetstesting av fire standardiserte evalueringsverktøy

Sammendrag

Formålet med studien er å finne reliable og valide evalueringsverktøy for vurdering av skulderfunksjon. 24 pasienter med proximal humerusfraktur (33-86 år) deltok i studien. Fire evalueringsverktøy; Constant score, fire standardiserte skulderfunksjonstester og to ADL spørreskjemaer ble vurdert i forhold til intra- og interreliabilitet. Dataene ble analysert ved hjelp av korrelasjon, Altman's agreement index og Kappa. Variasjonen for en enkelt måling for Constant score var stor, 95 prosent konfidensintervall (-10,9 til +10,3 poeng). De fire funksjonstestene: Intrareliabilitet 0,74 til 0,96 (Kappa), interreliabilitet, 0,36 til 0,69 (Kappa). Fire ADL funksjoner: 0,60 til 0,88 (Kappa). Korrelasjon mellom Constant score, to av funksjonstestene og fire ADL funksjonene: 0,60 til 0,88. Undersøkelsen tyder på at det er godt samsvar mellom standardiserte skulderfunksjonstester, ADL spørreskjemaer og Constant score. Det er holdpunkter for at de fire funksjonstestene er hensiktsmessige evalueringsverktøy for å vurdere skulderfunksjon hos den enkelte pasient med proksimal humerus fraktur, forutsatt at vurderingen foretas av samme fysioterapeut. Nøkkelord: Skulder, evalueringsverktøy, reliabilitet og validitet.

Kristin Thuve Dahm, fysioterapeut, Ullevål og Tåsen Fysioterapi og Trening AS

Inger Holm, fysioterapeut, dr philos, professor, fysioterapiavdelingen, Rikshospitalet

Anders Bakke, fysioterapeut, Ullevål og Tåsen Fysioterapi og Trening AS

Jens I. Brox, dr med, ortopedisk avdeling, Rikshospitalet

Det økte kravet til dokumentasjon er en utfordring innenfor helsefagene generelt og fysioterapifaget spesielt. Det stilles først og fremst krav til dokumentasjon av behandlingseffekt. For å kunne dokumentere effekt av fysioterapi er det en forutsetning at det finnes gode måleinstrumenter for registrering av fysisk funksjon.

Gode måleinstrumenter betyr i denne sammenheng valide og reliable instrumenter. Reliabiliteten til en målemetode sier noe om hvor stor feilmargin metoden er beheftet med. Hvis feilmarginene er små sier man at målemetoden er reliabel. Måleinstrumenter som skal benyttes til å evaluere effekt av behandling må kunne fange opp endringer over tid og oppdage interessante kliniske forskjeller når de oppstår (responsivness).

Validiteten defineres blant annet som en målemetodes evne til å gi meningsfull informasjon og hvorvidt målemetoden kan brukes til det formålet den er tiltenkt. En forutset-

ning for at en målemetode er valid, er at den er reliabel. Hvis målemetoden ikke fanger opp forandringene når de skjer, er det en trussel mot validiteten. Validiteten gjenspeiler blant annet i hvor stor grad instrumentet måler det underliggende fenomen (content validitet). Hvis et annet valid instrument finnes som måler det samme, kan det ene måles opp mot det andre (kriterievaliditet). Parametrene som man ønsker å måle, må ha en bakgrunn i den teoretiske forankringen som foreligger (construct) validitet.

Innenfor fysioterapi er det i hovedsak tre ulike former for målemetoder; biofysiologiske målinger, selvrappporterende målinger og målinger basert på observasjon (1). Måleinstrumenter som skal benyttes av fysioterapeuter i klinikken, bør være lette å anvende i praksis.

Hensikten med dette prosjektet er å vurdere måleinstrumenter for registrering av skulderfunksjon i forbindelse med overarmsbrudd (ukomplisert proksimal humerus fraktur). Fysioterapi utgjør en viktig del av behandlingen for disse pasientene. Det er imidlertid uenighet om effekten av slik behandling. Det er anvendt ulike måleinstrumenter for å registrere effekt av fysioterapi-behandling i ulike studier. Det vil derfor være viktig å finne frem til måleinstrumenter som er reproduerbare og samtidig sensitive nok til å registrere en eventuell effekt av behandling.

Det er uenighet om hvorledes vi best måler skulderfunksjon (2). Det er likevel noen parametre som er gjennomgående for de

eksisterende måleinstrumentene; smerte, leddbevegelighet og ADL-funksjon (3). Uenigheten dreier seg først og fremst om det er fornuftig å registrere enkeltparametere, eller om en totalskåre (summasjon av alle parametrene) gir et bedre bilde av skulderfunksjonen (2).

De fleste av måleinstrumentene for registrering av skulderfunksjon er utviklet og brukt av ortopedier og ikke nødvendigvis egnet for å vurdere effekt av fysioterapi. Totalskårer som er mye brukt i tidligere studier, er blant andre Constant score (2) og A standardized method for the assessment of shoulder function, ASES (4). Constant score og ASES skal kunne vurdere skulderfunksjon uavhengig av diagnose (3). Constant score skal kunne reflektere klinisk betydningsfulle forskjeller, brukes til å registrere forandringer etter en intervensjon og er den totalskåren som oftest benyttes i Europa (5).

Det er også utviklet måleinstrumenter spesielt for reumatiske pasienter, herunder «Fire skulderfunksjonstester» (6). Dette måleinstrumentet er også brukt ved vurdering av pasienter med proksimal humerus fraktur og inneklemmingssyndrom.

Det er utviklet flere skjemaer for registrering av skulderfunksjon i forhold til aktiviteter i dagliglivet, ADL. Et av skjemaene er utviklet i forbindelse med en effektstudie på proksimale humerus frakturer (7). Det foreligger ingen reliabilitetstesting av skjemaet. Et annet ADL skjema er utviklet i forbindelse med en studie av inneklemmingssyndrom (8), med en finere gradering av de enkelte funksjonene.

Målingene i denne studien er i hovedsak basert på observasjon og selvrappotering, og vil således kunne la seg påvirke av den som utfører målingene på en annen måte enn ved biofysiologiske målinger. De ulike måleinstrumentene vil derfor heretter bli omtalt som evalueringsverktøy.

I denne artikkelen beskrives en studie der målet var å finne frem til reliable og valide måleinstrumenter som er anvendbare for fysioterapeuter i praksis. Vi har valgt å vurdere de fire evalueringsverktøyene; Constant score, fire skulderfunksjonstester og to ulike ADL skjema, fordi de representerer ulike typer evalueringsverktøy og er mye brukt i praksis.

Materiale og metode

24 pasienter med proksimal humerus fraktur deltok i studien. Pasientene ble rekruttert fra Oslo Legevakt. Pasienter over 18 år som ble konservativt behandlet for proksimal humerus fraktur ble inkludert i studien og ble testet gjennomsnittlig åtte uker (3-21 uker) etter skaden. Gjennomsnittsalderen var

66 år (33-86 år). 15 av pasientene var kvinner og ni var menn. Det var like mange med skade i høyre som i venstre arm.

Fire ulike evalueringsverktøy for registrering av skulderfunksjon ble evaluert: Constant skulderscore (2), fire skulderfunksjonstester (6) og to ulike ADL skjemaer (7,8).

Constant score: Denne er utviklet som en totalskår bestående av et antall enkeltparametre. De ulike parametrene er dels selvrappoterende (smerte og ADL; defineres som delsum 1), dels målt av fysioterapeut (bevegelsesutslag og kraft; defineres som delsum 2). Parametrene blir registrert på ulike nivåer. Skåren er grovt delt inn på følgende måte:

- Smerte 15 poeng
- ADL 20 poeng
- Bevegelsesutslag 40 poeng
- Kraft 25 poeng
- Totalt 100 poeng

De fire skulderfunksjonstestene: Testene er utviklet av en svensk fysioterapeut (6) og består av fire enkle tester for skulderfunksjon:

- a) Løfte hånden mot taket
- b) Hånden mot motsatt skulder bak ryggen.
- c) Hånden i nakken.
- d) Hånden til motsatt skulder foran kroppen.

Hver av de fire testene er delt inn seks kategorier, betegnet 1-6.

Første ADL-skjema: ADL-skjemaet (7) er selvrappoterende og består av ni situasjoner fra dagliglivet: Gre håret, knytte et forkle bak (ta opp en lommebok fra baklommen), ta noe ned fra en hylle (over øyehøyde), henge opp tøy, re en seng, snu en pute om natten, vaske seg under motsatt arm, feste et sikkerhetsbelte og ta på en frakk.

Svarene er delt inn i fem kategorier fra 0-4; 0=ikke relevant, 1=umulig, 2=med stort besvær, 3=med litt besvær, 4=uten besvær.

Andre ADL-skjema: ADL-skjemaet (8) består av seks spørsmål fra dagliglivet: gre håret, ta på skjorte/genser, tørke seg bak, ta på sko/strømper, bære fem kilo varer i bærepose og ta noe ned fra høyt overskap. ADL funksjonene blir vurdert på en skala fra 1-7, hvor 1 er lett og 7 er umulig.

Pasientene ble testet to ganger samme dag av fysioterapeut (A) og (B) (interrater-reliabilitet) og retestet to dager etter av fysioterapeut (A) (intrarater-reliabilitet). Fysioterapeut A og B hadde sammen gjennomgått retningslinjene for de ulike testene og prøvetestet to pasienter med proksimal humerus fraktur. Den selvrappoterende delen av Constant score og ADL skjemaene ble bare testet for intrareliabilitet. For å kontrollere smertenivået i forbindelse med utførel-

sen av testene, ble hvilemerten i skulderen registrert på VAS skala før og etter testene ved de tre testtidspunktene.

De statistiske analysene ble gjennomført ved hjelp av SPSS (Statistical Package for the Social Science, 1997). For de kontinuerlige variablene ble det regnet ut korrelasjon, gjennomsnittlig differanse mellom målingene og standardavviket til differansen mellom målingene (9). Korrelasjonskoeffisienten (r) ble beregnet etter Pearsons standardmetode og måler i hvor stor grad det er lineær sammenheng mellom to målinger.

For å kunne estimere samsvar mellom enkeltmålinger ble først differansen mellom gjennomsnittene beregnet. Denne sier noe om gjennomsnittlige målefeil, bias, av den ene målingen i forhold til den andre. Standardavviket til differansen (S_{diff}) uttrykker samsvaret mellom enkeltmålingene ($2S_{diff}$ =coefficient of repeatability). Verdiene viste seg å være symmetrisk fordelte, og det ble beregnet et 95 prosent konfidensintervall for målingene. Bredden på dette intervallet betegnes som «limit of agreement»=95 prosent grense for enighet.

De ordinale dataene ble vurdert ved hjelp av korrelasjon (Spearman rho) og Kappa. Spearman rho uttrykker det samme som Pearson, men er regnet ut ved hjelp av rangen på observasjonene. Det ble i tillegg regnet enighet i prosent for de ordinale dataene, ved å telle opp antall sammenfallende observasjoner i forhold til totalt antall registreringer: (enighet/enighet + uenighet) x 100.

I Bostrøms artikkel om de fire funksjonstestene (6) omtales også enighet hvor det aksepteres ett poengs uenighet i prosent. For å kunne sammenligne dataene i denne studien med Bostrøms, ble denne prosenten derfor regnet ut når det gjelder interraterenighet for de fire funksjonstestene.

Det ble regnet gjennomsnitt med standardavvik for smerte før og etter testene (VAS skala).

Det ulike måleinstrumentene ble sammenholdt ved korrelasjon (Spearman rho).

Resultater

Vurdering av skulderfunksjonen hos 24 pasienter med proksimal humerus fraktur ved hjelp av Constant score ga en gjennomsnittlig poengsum på cirka 40 poeng. Gjennomsnittet med standardavvik var tilnærmet lik ved de ulike testtidspunktene (tabell 1).

Intraraterenigheten for Constant score, gjennomsnittet av differansen for de to målingene, var 0,3. Det ble regnet ut et 95 prosent konfidensintervall for Constant score fra -10,9 poeng til +10,3 poeng. Det vil si en variasjon på cirka 20 poeng fra en måling til en annen. Gjennomsnittet av differansen for delsum 2 var null. Variasjonen for del-

sum 2 hadde et 95 prosent konfidensintervall fra +5,6 til -5,6 poeng. Det vil si en variasjon på cirka 10 poeng. Interraterenighet, differansen mellom gjennomsnittet av de to målingene, var under 0,5 poeng, 95 prosent konfidensintervall (fra -10,4 til +9,6). Det vil si en variasjon på cirka 20 poeng av maksimalt 65 poeng (tabell 2).

Intraraterenigheten var fra 0,84 til 0,94 for de enkelte bevegelsesutslagene i Constant score. Interraterenigheten var 0,85 for abduksjon og 0,62 for innadrotasjon (tabell 3).

Intraraterenigheten for de fire skulderfunksjonstestene var fra 0,74 og 0,77 for tre av testene og 0,90 for en av testene (hånden mot taket). Interraterenigheten var 0,69 for hånden mot taket og 0,54 for testen hånden mot motsatt skulder (tabell 4).

Pasientens egen vurdering av de ulike ADL funksjonen var fra 0,63-0,67 for fire av ADL funksjonene (tabell 5).

Sammenheng mellom to av skulderfunksjonstestene, fire av ADL testene og Constant score var fra 0,60- 0,88 (tabell 6).

Diskusjon

Differansen mellom gjennomsnittet for de ulike målingene av Constant score var små både når det gjaldt intrarater- og interraterreliabilitet. Det var imidlertid stor variasjon mellom enkeltmålinger. Dette betyr at det var stor usikkerhet knyttet til en måling for den enkelte pasient. Bredden på et 95 prosent konfidensintervall var cirka 20 poeng (av maksimalt 100) når det gjaldt intraraterenighet. For delsum 2 var variasjonen cirka 11 poeng av maksimalt 65 når det gjaldt intraraterenighet og cirka 20 poeng når det gjaldt interraterenighet (tabell 2). Det vil si at Constant score kan variere tilfeldig med cirka 20 poeng fra en måling til en annen, vurdert av samme fysioterapeut. Hvis man med 95 prosent sikkerhet skal kunne si at en forandring i funksjon på individnivå skyldes en reell endring i funksjon og ikke en tilfeldig variasjon i målingene, må forskjellene mellom målingene overskride 20 poeng. Det er ikke definert hva som er en klinisk interessant forskjell i forhold til skåren. En variasjon på 20 poeng av 100 tilsier at Constant score er for upresis til å fange opp klinisk interessante forskjeller på enkelt-pasienter (lav responsivens). Dette betyr i praksis at Constant score kun er egnet til å vurdere effekt av behandling av grupper av pasienter.

Det foreligger få reliabilitets- og validitetsstudier av Constant score. I originalartikkelen ble bare interraterreliabiliteten vurdert. Den gjennomsnittlige «observer error» var tre prosent (0-8 prosent), og interraterreliabiliteten ble vurdert å være

Tabell 1: Constant score (gjennomsnitt ± standardavvik) for de tre test-tidspunktene

	A1	A2	B
Constant score	39,4±19,9	39,7±20,3	
Delsum 1	19,8±8,7	20,0±8,3	
Delsum 2	19,8±12,6	19,8±13,6	20,1±11,7

A1: Tester 1 første gang, A2: Tester 1 andre gang, B: Tester 2

Tabell 2: Intra- og interraterenighet for Constant score

	Intraraterenighet			Interraterenighet		
	Gjsn. Diff.	Sdiff	KI	Gjsn. Diff	Sdiff	KI
Constant score	-0,3	5,3 poeng	-10,9-+10,3			
Delsum 1	-0,2	3,5 poeng	-7,2-+6,9			
Delsum 2	0,0	2,8 poeng	-5,6-+5,6	-0,4	5,0 poeng	-10,4-+9,6

Gj.sn Diff: Gjennomsnitt av differansen mellom de to målingene

Sdiff: Standardavviket til differansen (Altmans)

KI: Konfidensintervall

Tabell 3: Intrarater- og interraterenighet for enkeltparametere av Constant score

Bevegelsesutslag	Intraraterenighet		Interraterenighet	
	Kappa	Prosent enighet	Kappa	Prosent enighet
Fleksjon	0,84	86	0,48	63
Abduksjon	0,94	90	0,85	88
Utdrotasjon	0,87	92	0,4	50
Innadrotasjon	0,89	92	0,62	71

Tabell 4: Intrarater- og interraterenighet for funksjonstestene

Funksjonstester	Intraraterenighet		Interraterenighet	
	Kappa	Prosent enighet	Kappa	Prosent enighet
Hånden mot taket	0,90	92	0,69	75
Hånden bak på ryggen	0,77	83	0,36	50
Hånden i nakken	*	83	*	54
Hånden mot motsatt skulder	0,74	79	0,54	63

* Kappa lar seg ikke regne ut, da en av kategoriene manglet ved det ene test-tidspunktet

Tabell 5: Pasientens egen vurdering ved testdag 1 og 2. Bære 5 kilo i en bærepose inneholder syv kategorier. De andre funksjonene inneholder fire kategorier

ADL funksjoner	Kappa	Prosent enighet
Vaske seg under motsatt arm	0,66	79
Ta noe ned fra en hylle	0,67	97
Gre hår	*	91
Bære fem kilo i en bærepose	0,63	69

* Kappa lar seg ikke regne ut, da en av kategoriene mangler ved det ene tidspunktet.

Tabell 6: Sammenhengen mellom funksjonstestene/ASL testene og Constant score. Spearman rho

Funksjons-/ADL testene	Constant score
Hånden mot taket	0,88
Hånden mot motsatt skulder	0,79
Gre håret	0,78
Vaske seg under motsatt arm	0,60
Ta noe ned fra en hylle	0,88
Bære fem kilo i en bærepose	0,70

svært god. Det er vanskelig å forstå hvorledes tallene er kommet frem. Artikkelen gir gjennomsnittlige forskjeller og sier ikke noe om usikkerheten rundt en enkelt måling. Studien ble gjennomført på hundre pasienter som ble vurdert i grupper på 10 og 10. Pasientene ble vurdert av tre fysioterapeuter (2). Variasjonen når det gjelder en enkelt måling, var stor. En annen undersøkelse (3) underbygger at skåren er for upresis til klinisk å følge opp den enkelte pasienten. Det var stor variasjon både når det gjaldt intrarater- og interraterreliabilitet; 16-20 poeng for en enkelt måling. Undersøkelsen ble utført på pasienter med artritt, innklemmingssyndrom og instabilitet. Intrarater- og interraterreliabiliteten angitt i poeng hadde en variasjon på henholdsvis 16 og 18 poeng for en enkelt måling. Variasjonen for en enkelt måling var altså litt mindre i denne studien enn i vår studie. Dette gjaldt når alle pasientene ble vurdert under ett. Variansen i de ulike gruppene var forskjellig, slik at de ulike diagnosegruppene burde vært vurdert hver for seg. Gruppene ble da svært små og resultatene ble usikre. Variasjonen for en enkelt måling var størst for pasienter med instabilitet.

Resultatene av vår studie viste at intraraterreliabiliteten for de fire funksjonstestene var god for tre av testene (Kappa 0,74-0,77) og svært god for den ene av testene (Kappa 0,90) (hånden mot taket). Interraterreliabiliteten derimot var god for kun en av funksjonstestene (Kappa 0,69) (hånden mot taket) og moderat (Kappa 0,54) for testen hånden mot motsatt skulder (tabell 4).

De fire funksjonstestene kan brukes til å vurdere effekt av behandling for den enkelte pasient. De fire testene inneholder til sammen informasjon om skulderens aktive bevegelsesutslag i alle plan. Det er imidlertid en forutsetning at det er samme fysioterapeut som vurderer skulderfunksjonen før og etter behandlingen. To av testene, hånden mot taket og hånden mot motsatt skulder, vurderes også relativt likt av to forskjellige fysioterapeuter. Disse to testene gir begrenset informasjon om skulderens aktive bevegelsesutslag (kun fleksjon og fleksjon med adduksjon) og vil derfor alene sannsynligvis ikke være egnet til å vurdere effekt av behandling. Det ser ut til at funksjonstesten hånden bak på ryggen bør erstattes av delparameteren innadrotasjon av Constant score. Interraterreliabiliteten for denne parameteren var god (Kappa 0,62) og høyere enn for funksjonstesten hånden bak på ryggen (Kappa 0,36) (tabell 6).

Enigheten av skulderfunksjonstestene ble vurdert ved hjelp av Kappa. Ingen av de tidligere reliabilitetsstudiene som omhandler de

fire funksjonstestene har brukt Kappa. Kappa tar hensyn til hvor stor enigheten er og tar også hensyn til at enighet kan forekomme ved tilfeldighet. Et problem ved bruk av Kappa er at den er avhengig av prevalensen av de ulike kategoriene, slik at ulike forventede verdier gir ulike verdier for Kappa. Pasientene i denne studien har relativt dårlig funksjon og det er ikke forventet lik fordeling i de ulike kategoriene. Dette kan ha vært med på å redusere verdien for Kappa. Et annet problem er at Kappa avhenger av antall kategorier. De fire funksjonstestene inneholder mange kategorier, totalt seks. Enighet kan bli bedre med færre kategorier. Antall kategorier kan således også ha vært med på å redusere verdien for Kappa. Disse forhold må man ta i betraktning når man analyserer verdiene av Kappa. Vanligvis vil en koeffisient større enn 0,61 være en tilfredstillende verdi for Kappa.

Intraraterenigheten i prosent i denne studien var betydelig høyere enn i Bostrøms undersøkelse (6), henholdsvis 79-92 prosent mot 53-75 prosent. Dette kan skyldes at Bostrøms studie ble gjennomført på reumatiske pasienter. Det synes å være stor variasjon i bevegelsesutslag fra dag til dag hos reumatiske pasienter, selv om de ikke er i en akutt fase av sykdommen.

Bostrøm (6) vurderte enighet for de fire funksjonene i prosent og godtok ett poengs uenighet. Interraterreliabiliteten var da fra 80-95 prosent i Bostrøms studie. Interraterreliabiliteten i denne studien var fra 92-100 prosent med ett poengs uenighet, altså noe høyere enn i Bostrøms studie. Interreliabiliteten var relativt lik i de to studiene, mens intrareliabiliteten i prosent i denne studien var betydelig høyere enn i Bostrøms undersøkelse, henholdsvis 79-92 prosent mot 53-75 prosent.

To av de fire funksjonstestene (hånden i nakken og hånden bak på ryggen) ble reliabilitetstestet i en studie hvor funksjon ble vurdert i forhold til smerte (10). Fem pasienter med proksimal humerus fraktur ble vurdert av to fysioterapeuter ved fire anledninger. Dataene var således ikke uavhengige, men ble analysert som uavhengige og vurdert i forhold til interraterenighet (prosent enighet). Enigheten var god for hånden bak på ryggen (85 prosent) og dårlig for hånden i nakken (40 prosent). I vår studie var interraterreliabiliteten dårlig for begge disse testene (50 prosent og 54 prosent). En forklaring på de ulike resultatene i de to studier

ene kan være at dataene i Westerbergs studie ikke var uavhengige. Det er dessuten vanskelig å trekke konklusjoner fra målinger fra så få som fem pasienter.

Det er ikke definert hva som er en klinisk interessant forskjell i forhold til de fire funksjonstestene. Hvis testene skal kunne brukes til å vurdere effekt av behandling kan man ikke akseptere en uenighet på ett poeng slik som i Bostrøms undersøkelse. Det vil si at vurderingen av en pasient kan variere tre kategorier uten at dette skyldes annet enn tilfeldig variasjon i målingene. Hvis man med rimelig grad av sikkerhet skal kunne si at en forandring i funksjon for den enkelte pasient skyldes en reell endring i funksjon og ikke tilfeldig variasjon i målingene, må forskjellen mellom målingene være mer enn tre kategorier. Det er derfor nødvendig at man stiller strenge krav til enighet, slik at man med rimelig grad av sikkerhet kan si at en forandring på en kategori skyldes en reell forandring i funksjon og ikke en tilfeldig variasjon i målingene.

Pasientens egen vurdering av de ulike ADL-funksjonene fra første til andre testdag var god for fire av funksjonene; vaske seg under motsatt arm, ta noe ned fra en hylle, gre håret og bære fem kilo i en bærepose. Det så ikke ut til å være noen systematisk variasjon i vurderingene fra den ene gangen til den andre. De ADL-funksjonene som har best reproducerbarhet ser ut til å være de funksjonene som stiller størst krav til funksjon. Dag til dag variasjonen på spørreskjemaet var lavere enn for funksjonstestene gjennomført av to ulike fysioterapeuter. Skjemaene er enklere å bruke enn funksjonstestene og er dessuten observerfri. Bære fem kilo i en bærepose er gradert på en skala fra 1-7 og kan således være egnet til å vurdere effekt av behandling, mens de tre andre funksjonene bare inneholder fire kategorier og derfor muligens er for «grove» til å kunne vurdere effekt av behandling.

Det finnes ingen «gullstandard» for registrering av skulderfunksjon. De tre evalueringsverktøyene ble derfor vurdert mot hverandre. Constant score ble behandlet som kontinuerlige data. De fire funksjonstestene og ADL-testene inneholder ordinale data. Constant score ble sammenholdt med de kategoriske testene ved korrelasjon (Spearman rho). Det var god korrelasjon mellom Constant score og to av funksjonstestene (hånden mot taket og hånden mot motsatt skulder) målt på «impairment» nivå og fire ADL-funksjoner (ta noe ned fra en hylle, vaske seg under motsatt arm, gre håret og bære fem kilo i en bærepose) målt på «activity» nivå. Dette er med på styrke validiteten av de tre evalueringsverktøyene. To andre studier viser at Constant score er

sammenfallende med pasientens vurdering (11,12). Den ene studien ble gjennomført på pasienter med forskjøvede humerus frakturer (12).

Utvalget i studien er lite (n=24) og gruppen er heterogen både med hensyn til alder og funksjonsnivå. Dette resulterte i relativt få data på de ulike funksjonsnivåene. I tillegg er det mulige feilkilder ved testprosedyrene og unøyaktighet i anvendelsen av en del av Constant score. Dette tilsier en forsiktig tolkning av resultatene.

Konklusjon og eventuell nytte for klinikken og den enkelte kliniker

Intraraterreliabiliteten var god og interraterreliabiliteten var relativt god for Constant score, differansen mellom gjennomsnittet for de ulike målingene var små både når det gjaldt intrarater- og interraterreliabilitet. Det var imidlertid stor usikkerhet knyttet til en måling for den enkelte pasient. Bredden på et 95 prosent konfidensintervall var 20 poeng, når det gjaldt intraraterreliabilitet. Det vil si at hos pasienter med proksimal humerus fraktur må en forandring i funksjon overstige 20 poeng for at man med 95 prosent sikkerhet skal kunne si at forandringen ikke skyldes en tilfeldig variasjon i målingene. Dette betyr i praksis at Constant score er best egnet til å vurdere effekt av behandling på grupper av pasienter.

Det ble også funnet at intraraterreliabiliteten for de fire funksjonstestene var god for tre av testene og svært god for en av testene (hånden mot taket). Interraterreliabiliteten var derimot god for kun en av funksjonstestene (hånden mot taket) og moderat for hånden mot motsatt skulder. Delparameteren innadrotasjon av Constant score kan med fordel erstatte funksjonstesten hånden bak på ryggen. Dette betyr i praksis at tre av de fire funksjonstestene og delparameteren innadrotasjon av Constant er egnet til å vurdere effekt av behandling for den enkelte pasient i alle fall når vurderingene foretas av samme fysioterapeut ved de ulike testtidspunktene.

Denne studien gir holdepunkter for at tre av funksjonstestene og innadrotasjon av Constant score er hensiktsmessig evalueringsverktøy for å vurdere skulderfunksjon hos pasienter med proksimal humerus fraktur. Testene synes å kunne registrere en eventuell effekt av behandling for den enkelte pasient, forutsatt at målingene foretas av samme fysioterapeut. Testene kan være et hjelpemiddel for å kartlegge skulderfunksjon hos disse pasientene og være et arbeidsverktøy i forhold til progresjon og effekt av behandling. De er enkle å gjennomføre og krever ingen hjelpemidler.

Abduksjon målt med goniometer har både god intrarater- og interraterreliabilitet og bør derfor brukes i tillegg. Dag til dag variasjonen på de fire ADL-spørsmålene; gre hår, ta noe ned fra en hylle, vaske seg under motsatt arm og bære fem kilo i en bærepose, er lavere enn for de fire funksjonstestene gjennomført av to ulike fysioterapeuter og bør derfor også brukes.

Litteratur

1. Domholdt E: Physical Therapy Research and Instruments, Principles and Application. Philadelphia, W.B.Saunders Company, 1993.
2. Constant CR, Murley AGH: A Clinical Method of Functional Assessment of the Shoulder. *Clinical Orthopaedics* 1985, 160-4.
3. Conboy VB, Morris RW, Kiss J, Carr AJ: An evaluation of the Constant-Murley shoulder assessment. *The Journal of Bone and Joint Surgery* 1996, 78-B, 229-32.
4. Richards RR, Biglani LU, Friedman R, Gartsman G, Gristina A, Iannotti J, Sidles J, Zuckerman J: A standardized method for the assessment of shoulder function. *J.Shoulder Elbow Surgery* 1994, 3, 347-52.
5. Romeo AA, Bach Jr BR, O'Halloran KL: Scoring Systems for Shoulder Conditions. *The*

American Journal of Sports Medicine 1996, 24, 472-6.

6. Bostrøm C, Harms-Ringdahl K, Nordemar R: Clinical Reliability of Shoulder Function Assessment in Patients With Rheumatoid Arthritis. *Scand.J.Rheumatol* 1991, 20, 36-48.

7. Solem-Bertoft E, Lundh I, Ringqvist I: Physiotherapy after Fracture of the Proximal End of the Humerus. *Scand.J.Rehab.Med.* 1984, 16, 11-16.

8. Brox JI: Rotator tendinosis of the shoulder. Universitetet i Oslo 1997.

9. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* 1986, 8, 307-10.

10. Westerberg C-E, Solem-Bertoft E, Lundh I: The reliability of three active motor tests used in painful shoulder disorders. Presentation of a method of general applicability for the analysis of reliability in the presence of pain. *Scand.J.Rehab.Med.* 1996, 28, 63-70.

11. Iannotti JP, Bernot MP, Kuhlman JR, Kelley MJ, Williams GR: Postoperative assessment of shoulder function: A prospective study of full-thickness rotator cuff tears. *J.Shoulder Elbow Surgery* 1996, 5, 449-57.

12. Zyto K, Kronberg M, Brostrøm L-Å: Shoulder function after displaced fractures of the proximal humerus. *Journal of Shoulder and Elbow Surgery* 1995, 4, 331-36.

Abstract

The objective of the study was to find a reliable and valid assessment tool for evaluating shoulder function. 24 patients with proximal fractures of the humerus (33-86yrs) took part in the study. Four assessment tools were used; Constant score, four standardised shoulder function tests and two ADL questionnaires were evaluated in relation to intra- and inter-reliability. The data was analysed by correlation, Altman's agreement and kappa. There were large variations in single measurements for the Constant score, 95 percent confidence interval (-10,9 - +10,3 points). The four function tests: intra-reliability 0,74-0,96 (Kappa), inter-reliability 0,36-0,69 (Kappa). Four ADL functions: 0,60-0,88 (Kappa). The correlation between Constant score, two function tests and four ADL functions: 0,60-0,88. The investigation suggests that there is good correlation between standardised shoulder function tests, ADL questionnaires and Constant score. There is reason to believe that the four function tests are useful tools in assessing shoulder function in the individual patient with proximal fracture of the humerus, providing the evaluation is undertaken by the same physiotherapist.

Key words: Shoulder, assessment tool, reliability and validity.

Du finner flere fagartikler på
www.fysioterapeuten.no