Intertester reliabilitet av fysiske tester
-testet på arbeidstakere med muskel- og skjelettplager

Lise Krohn-Hansen

Masteroppgave i helsefag - Fysioterapivitenskap

Faggruppe for fysioterapivitenskap
Institutt for global helse og samfunnsmedisin
Universitetet i Bergen

Vårsemester, 2015

Forord

For et par år siden hadde jeg aldri trodd at jeg skulle sitte i Rio de Janeiro og skrive forord til masteroppgaven min. En drøm har gått i oppfyllelse. To spennende år på masterstudiet kombinert med å bo i det store utland med er snart over. Det har vært to krevende år; ny kultur, nytt språk, nytt nettverk og ikke minst et masterstudiet som skulle gjennomføres, men for en berikelse. Jeg føler meg ekstremt privilegert. Det å ha mulighet for å ta høyere utdannelse er ingen selvfølge. Selv om min jobberfaring har vært innenfor nevrologisk fysioterapi og rehabilitering er jeg veldig takknemlig for at det åpnet seg opp en mulighet for å knytte masteroppgaven min opp mot FAktA-prosjektet ved Institutt for global helse og samfunnsmedisin, UiB. Tross mange hundre mils avstand har prosjektet praktisk latt seg gjennomføre med god veiledning på epost og skype, samt flere turer hjem i forbindelse med samlinger på masterstudiet.

En stor takk til mine veiledere Tove Ask og Tove Dragesund som har gitt uvurderlig veiledning og tilbakemeldinger i prosessen. Ikke minst en stor takk til jobben dere har gjort som testere i prosjektet sammen med Alice Kvåle.

Tusen takk til mine ledere Else Sterndorff og Ingrid Thorsen som har gjort det mulig å få utdanningspermisjon fra fysioterapiavdelingen på Haukeland sykehus. En takk til Fond til Etter- og Videreutdanning av Fysioterapeuter for økonomisk støtte.

Takk til studievenner og kollegaer som har delt av sine erfaringer og vært en stor inspirasjonskilde til å gjennomføre dette prosjektet.

Til slutt vil jeg takke familie og venner. Dere betyr alt for meg. En spesiell takk til tvillingsøsteren min Nina som har ”backet” opp de dagene det har gått litt trått. Ikke minst vil jeg rette en stor takk til min kjære Eirik som først og fremst er en fantastisk ektemann i tillegg til at han har vært en super mentor og inspirator underveis.

Rio de Janeiro, mai 2015
Lise Krohn-Hansen

Innholdsfortegnelse

Vedlegg

Figurer

Tabeller

Illustrasjoner

Sammendrag

Bakgrunn: I prosjektet Funksjon, Aktivitet og Arbeid (FAktA) gikk arbeidstakere med muskel- og skjelettplager som var sykemeldt eller stod i fare for å bli sykemeldt gjennom en funksjonsundersøkelse. **Hensikt:** Hensikten med denne studien var å undersøke intertester reliabilitet mellom tre testere for seks fysiske tester som inngår i funksjonsundersøkelsen i FAktA-prosjektet. **Materiale og metode:** Et tverrsnittdesign ble brukt. De seks fysiske testene var; seks deltester fra Global Kroppsundersøkelse (GBE-Fleksibilitet), Back Performance Scale (BPS), høy løftetest, Biering-Sørensen test, dynamisk sit-up test og 18 definerte punkt etter kriterier fra American College of Rheumatology (ACR-18). Førti-åtte deltagere ble testet ved to anledninger. Testerne var tre fysioterapeuter med lang erfaring med bruk av testene. Intraclass correlation coeffisient ($ICC_{2.1}$) med 95% konfidens intervall, Within-subject standard deviation (S_w), Smallest detectable change (SDC) og Bland-Altman plot ble kalkulert. **Resultater:** Testene viste høy til svært høy intertester reliabilitet og $ICC_{2,1}$ varierte mellom 0.80 to 0.94. Absolutt reliabilitet uttrykt i målefeil ble rapportert, og viste moderat grad av målefeil for fire av testene og stor grad av målefeil for to av testene. **Konklusjon:** Basert på resultatene i denne studien er intertester reliabilitet mellom testere i FAktA-prosjektet god for alle de 6 fysiske testene når de er testet som et testbatteri, men det er stor grad av målefeil for dynamisk sit-up test og ACR-18.

Nøkkelord: intertester reliabilitet, muskel- og skjelettplager, fysiske tester, måleinstrument

Abstract

Background: In the project named Function, Activity and Work (FAktA) workers with musculoskeletal pain, who either were on sick leave or at risk of being sick-listed, were examined by a functional evaluation. **Objective:** The purpose of this study was to examine the intertester reliability between three testers with regards to six physical tests included in the functional evaluation of the FAktA-project. **Method:** A cross-sectional design was used. The six physical tests studied are; six subtests from Global Body Examination (GBE-Flexibility), Back performance scale (BPS), high lifttest, Biering-Sørensen test, dynamic sit-up test and 18 tender points examined by criterias' from the American College of Rheumatology (ACR-18). Forty-eight subjects were tested on two occasions. The testers were three therapists who had long experience with using the physical tests. Intraclass correlation coefficients ($ICC_{2.1}$) with 95% confidence interval, Within-subject standard deviation ($S_{w),}$ Smallest detectable change (SDC) and Bland-Altman plots were calculated. **Results:** The intertester reliability was found to be high to very high for all the tests and $ICC_{2,1}$ varied between 0.80 to 0.94. The absolute reliability expressed as measurement errors have a moderate degree of measurement error for four of the tests and high degree of measurement error for two tests. **Conclusion:** This study indicates that the intertester reliability between three testers are good for all the six physical testes when they are tested as a test battery, but dynamic sit-up test and ACR-18 showed a high level of measurement error.

Keywords: Intertester reliability, musculoskeletal disorders, physical tests, measurement instrument

Bruk av referanser og forkortelser

I referanselisten er APA 6th brukt. Utrykk som er forkortet i teksten, blir skrevet fullt ut første gang, deretter blir forkortelsen skrevet i parentes. Oppgaven er paginert fortløpende fra og med introduksjon. Vedlegg er nummerert, og satt inn bakerst i oppgaven.

Oversikt over forkortelser som er brukt i teksten:

ACR	American College of Rheumatology
BPS	Back Performance Scale
GBE	Global Kroppsundersøkelse
GFM	Global Fysioterapi Metode
CI	Konfidens intervall
ICC	Intraclass correlation coefficients
ICF	Internasjonal klassifikasjon av funksjon, funksjonshemming og helse
MCD	Minimal detectable difference
NPRS	Numeric Pain Rating Scale
SDC	Smallest detectable change
SDD	Smallest detectable difference
SEM	Standard error of measurement
S_w	Within-subject standard deviation
WHO	Verdens helseorganisasjon

1 INTRODUKSJON

Bakgrunn

Muskel- og skjelettplager er den hyppigste medisinske årsak til sykefravær og uføreytelser i Norge. I fjerde kvartal av 2014 var muskel– og skjelettplager den medisinske årsaken i omlag en tredjedel av 33,1% av legemeldt sykefravær (NAV, 2014a). Muskel- og skjelettplager fører ofte til redusert funksjon (Ihlebaek, Brage, Natvig, & Bruusgaard, 2010). For å fange opp ressurser og begrensninger knyttet til plagene vil fysiske tester med gode måleegenskaper, være av betydning for å gi råd og tilrettelagt behandling (de Vet, Terwee, Mokkink, & Knol, 2011).

Ved Universitetet i Bergen, Forskningsgruppe i fysioterapi, pågår forskningsprosjektet "Muskel-skjelettplager – Funksjon, aktivitet og arbeid (FAktA)". Hoved hensikten med prosjektet er å undersøke om tidlig intervensjon på arbeidsplassen og i primærhelsetjenesten, kan bidra til redusert utvikling av langvarige muskel- og skjelettplager og eventuelt redusere sykefravær. Som en del av FAktA-prosjektet inviteres arbeidstakere med muskel- og skjelettplager til å gjennomføre en funksjonsundersøkelse hos fysioterapeut ved Institutt for global helse og samfunnsmedisin, UiB. De fysiske testene som inngår i funksjonsundersøkelsen er valgt for å fange opp ulike aspekter ved funksjon hos arbeidstakerne. Det er tidligere utført reliabilitetsstudier av de fleste fysiske testene, men det er ikke utført reliabilitetstester på de seks fysiske testene utført samlet som et testbatteri. Hensikten med denne studien er å undersøke intertester reliabilitet mellom tre testere på de seks fysiske testene som inngår i funksjonsundersøkelsen i FAktA-prosjektet.

Oppbygning og avgrensning av oppgaven

I første del av oppgaven presenteres teori om muskel- og skjelettplager. I denne studien avgrenses muskel og skjelettplager til å gjelde rygg-, nakke- og skulderplager, samt utbredte smerter. Videre vil den Internasjonale klassifikasjon av funksjon, funksjonshemming og helse (ICF) beskrives som et teoretisk rammeverk. Både

selvrapporterte måleinstrument og fysiske måleinstrument vil bli beskrevet, men med hovedfokus på fysiske måleinstrument.

Måleegenskaper vil bli presentert i et eget underkapittelet, og i det vil reliabilitet ha hovedfokus. Tidligere forskning på de fysiske testene vil bli beskrevet, før oppgavens hensikt og problemstilling blir presentert. Videre presenteres metode, resultater, diskusjon og konklusjon. Vedlegg er lagt ved til slutt.

2 TEORI

2.1 Muskel og skjelettplager

Muskel- og skjelettplager er svært vanlig i befolkningen (Mody & Brooks, 2012). I en studie av Ihlebaek et al. (2010) anga hele 80% å ha hatt slike plager i løpet av siste måned, og prevalens synes å være stabil over tid (Brage, Ihlebaek, Natvig, & Bruusgaard, 2010). De vanligste plagene er fra korsrygg, nakke og skuldre (E. Lærum et al., 2013). Ryggsmerter er den vanligste plagen og det er antatt at mellom 60-80 prosent% vil erfare ryggsmerter i livet (E. Lærum et al., 2007). Ettårs prevalens har vært rapportert til å være 48% for nakke plager og 47% for skulder plager (Ihlebaek et al., 2010).

Muskel og skjelettplager skilles ofte mellom spesifikke og uspesifikke plager. Ved spesifikke plager er det en mer klar årsakssammenheng mellom symptomer og smerter, som for eksempel skiveprolaps som fører til korsryggsmerter (E. Lærum et al., 2007). For uspesifikke muskel og skjelettplager er sammenhengen mellom symptomer og såkalte objektive funn ved undersøkelse svakere (E. Lærum et al., 2013). For om lag 80-90% av akutte korsryggplager kan man ikke si noe sikkert om hvorfor smerten oppstår (Ihlebaek et al., 2010), og uspesifikke muskel- og skjelettplager utgjør en stor andel av de som har muskel- og skjelettplager.

Muskel- og skjelettplager kan involvere anatomiske strukturer som skjelettet, ledd, muskler, sener (Brage et al., 2010) og kan være lokalisert til et bestemt ledd, eller være mer utbredt (Mody & Brooks, 2012). Utbredte smerter kan defineres til å gjelde smerte som er lokalisert til både over og under midjen, og på begge sider av kroppen, mens lokaliserte smerter er konsentrert til et mindre område av kroppen (Wolfe et al., 1990). Det har vist seg at de fleste som har muskel og skjelettplager rapporterer om smerte fra flere steder (Ihlebaek et al., 2010). Kamaleri, Natvig, Ihlebaek, & Bruusgaard (2008) fant ut at lokaliserte smerter ikke hadde særlig innvirkning på fysisk funksjon, følelser, eller daglige og sosiale aktiviteter. De fant derimot ut at det var en sterk assosiasjon mellom antall smerteområder og problemer med funksjonsevne.

Muskel- og skjelettplager kan ha varierende varighet og deles ofte inn i akutte og langvarige plager. Akutte plager kan defineres som plager med varighet under tre måneder (E. Lærum et al., 2007). Slike plager er ofte som en naturlig del av livet. De bedres ofte av seg selv og får som regel små konsekvenser for funksjon i dagligliv og arbeidsliv. Langvarige plager defineres som plager med varighet over tre måneder (E. Lærum et al., 2007). Plagene er mer komplekse enn de akutte plagene og har en multifaktoriell etiologi, som inkluderer både biologiske, psykologiske og sosiale faktorer (Main, Watson, & Sullivan, 2008, s. 25). De langvarige plagene fører ofte til langvarig sykemelding og i noen tilfeller uføretrygd (Ihlebaek & Lærum, 2004).

2.1.1 Sykefravær

En grundig kunnskapsoversikt om sammenheng mellom arbeid og muskel- og skjelettplager ble gjennomført av Statens Arbeidsmiljøinstitutt (STAMI) (Knardahl et al., 2008), og risikofaktorer kan utgjøre både, fysiske, psykiske og sosiale faktorer. Som en konsekvens av muskel- og skjelettplager blir en betydelig andel arbeidstakere sykemeldt. Det har vist seg at kvinner med muskel og skjelettplager er oftere sykemeldt enn menn (Andersen, Frydenberg, & Mæland, 2009) , og man har sett en økning i forekomst av muskel- og skjelettplager med økende alder (Ihlebaek et al., 2010). I tillegg er lavt utdanningsnivå (Andersen et al., 2009), høy smerteintensitet (Holtermann, Hansen, Burr, & Sogaard, 2010) og hardt fysisk arbeid (Lund, Labriola, Christensen, Bultmann, & Villadsen, 2006) faktorer som kan påvirke sykefravær. Sykefravær varier innen forskjellige yrkessektorer og arbeidstakere i helse-og sosialsektoren har vist seg å ha et høyere sykefravær enn andre arbeidstakere (Eriksen, Bruusgaard, & Knardahl, 2003).

Det er ulike sammenhenger for akutte plager og langvarige smerter. Arbeidsstillinger og arbeidstempo kan for eksempel ha stor betydning for å utløse plager, mens psykososiale forhold ofte har større betydning i forhold til om plagene blir langvarige (Even Lærum et al., 2013). Utberedte smerter og langvarige muskel og skjelettplager har vist seg å øke risikoen for langvarig sykefravær (Andersen et al., 2009).

Sannsynligheten for å komme tilbake til jobb synker jo lengre en arbeidstaker er borte fra jobb (E. Lærum et al., 2007) og dermed er det viktig for samfunnet å tilrettelegge med gode tiltak som kan holde folk i arbeid. I 2001 ble det inngått en intensjonavtale om et mer inkluderende arbeidsliv (IA-avtalen) mellom regjeringen og arbeidslivsorganisasjonene. Et av målene i avtalen går ut på å redusere sykefraværet og hindre at arbeidstakere faller ut av arbeidslivet (NAV, 2014c). Det kreves ofte forskjellig tilnærming til arbeidstakere som er sykemeldt, hvor de som har langvarige utbredte muskel- og skjelettplager kan trenge tettere tverrfaglig behandling og tiltak på arbeidsplassen enn de som har kortvarige plager (Skouen, 2006). I studien til Skouen (2006) viste resultatene at en høyere prosentandel returnerte til jobb når riktig behandling ble gitt til den riktige pasienten.

Resultatene fra studien til Ask, Skouen, Assmus, and Kvale (2014) viste at arbeidstakere som var fullt sykemeldt hadde redusert selvrapportert funksjon og redusert funksjon testet med fysiske tester, sammenlignet med arbeidstakere som jobbet selv om de hadde muskel- og skjelettplager eller var delvis sykemeldt. En viktig del av tilnærmingen til arbeidstakere med muskel- og skjelettplager blir å kartlegge fysiske og psykiske faktorer som påvirker funksjon og arbeidsevne.

2.2 Funksjon

Det finnes ulike begrep som beskriver funksjon, noen eksempler er; fysisk funksjon, funksjonsevne, fysisk evne, aktivitetsnivå, kapasitet, ytelse og funksjonsbegrensninger (Wittink, 2005). Idégrunnlaget for ICF er at alle kroppsfunksjoner, aktiviteter og deltagelse omfattes av paraplybetegnelsen funksjon (WHO, 2003, s. 3).

2.2.1 ICF

I 2001 ga Verdens helseorganisasjon (WHO) ut ICF, hvilket er et system som gjør det mulig å klassifisere funksjon. Den internasjonale sykdomsklassifikasjonen (ICD-10) har vært brukt fra før til å klassifisere sykdommer. Kombinasjon av diagnostisk informasjon i ICD-10 og beskrivelse av funksjon i ICF, vil på mange måter gi et bredere og mer meningsfylt bilde av personlig helse og folkehelse som grunnlag for tiltak på individuelt

nivå og i samfunnsperspektiv. Med ICF har man i tillegg fått et felles verdensspråk for funksjon som har til hensikt å lette tverrfaglig og internasjonal helsekommunikasjon (WHO, 2003, s. 3-5). Med ICF får man muligheten til bredde og systematikk i vurderingen av menneskers funksjonsevne (Pran, 2007).

I ICF blir vekselvirkningene mellom helsetilstand, funksjonsevne og miljøfaktorer fremhevet (WHO, 2003). Det er både et begrepsapparat, en klassifikasjon og et kodeverk og baseres på en biopsykososial forståelse av mennesket. Det vil si at biologiske, psykologiske og sosiale faktorer som påvirker menneskets funksjon er ivaretatt gjennom vekselvirkninger mellom de forskjellige funksjonsdimensjonene, miljøfaktorer og personlige faktorer i ICF (Pran, 2007). Det er viktig å skille mellom ICF som teoretisk eller begrepsmodell og ICF som klassifikasjons- og kodesystem. Det begrepsmessige grunnlaget for klassifikasjonen omfatter to hovedområder; (I) funksjon og funksjonshemming og (II) kontekst (WHO, 2003).

Hovedområde 1. Funksjon og funksjonshemming:

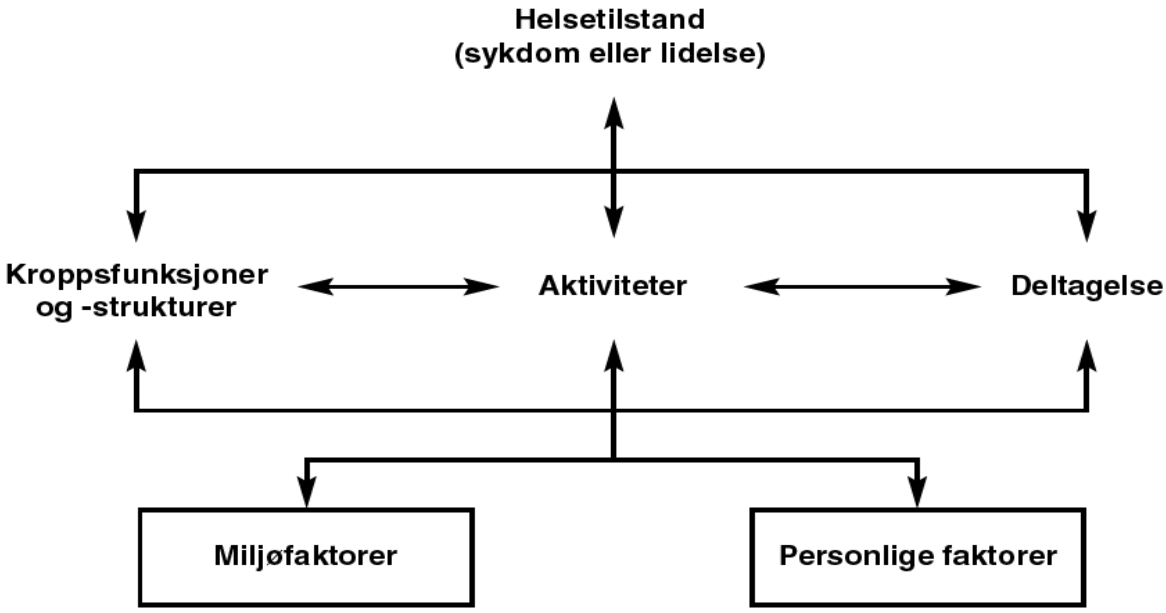
1. Kroppsfunksjoner og –strukturer (anatomiske og fysiologiske strukturer og funksjoner).
2. Aktiviteter og deltakelse (kapasitet og utførelse av oppgaver og handlinger). Avvik på et eller flere av disse områdene omtales som funksjonshemminger.

Hovedområde 2. Kontekstuelle faktorer

1. Miljøfaktorer (ytre påvirkning av funksjon og funksjonshemming)
2. Personlige faktorer (indre påvirkning av funksjon og funksjonshemming).

Kroppsfunksjoner og kroppsstrukturer sammen med aktiviteter og deltakelse beskriver hovedområdet 1; Funksjon og funksjonshemming. Miljøfaktorer er kontekstuelle faktorer som påvirker funksjonsnivået sammen med de personlige faktorene som ikke klassifiseres i ICF. Miljøfaktorer er fysiske, sosiale og holdningsmessige omgivelser som individet eksisterer innen og utfolder sitt liv i (Pran, 2007).
Figur 1 viser ICFs begrepsapparat med de ulike begrepenes innbyrdes vekselvirkning.

Et eksempel på bruk av ICF's begrepsmodell kan være at kroppsfunksjon er bevegelighet i ryggen, mens det å løfte er en typisk aktivitet. Deltakelsesperspektivet omhandler blant annet personens muligheter til deltakelse i arbeid, familie- eller fritidsaktiviteter. Miljøfaktorer kan være både hemmende og fremmende. Eksempelvis kan en sykehjemsavdeling uten hev og senk senger være en negativ miljøfaktor for de som jobber der, mens det å ha personløfter til pasienter som ikke kan reise seg opp selv vil være en positiv miljøfaktorer og kan være en fremmende faktor for funksjonsnivået til en ansatt med nedsatt bevegelighet i ryggen.



Figur 1. *Vekselvirkninger i ICF's begrepsapparat som viser interaksjon mellom de forskjellige funksjonsdimensjonene, miljøfaktorer og personlige faktorer. Figuren er hentet fra Internasjonal klassifikasjon av funksjon, funksjonshemming og helse (WHO, 2003, s. 17).*

2.3 Måleinstrument for å fange opp plager relatert til helse og funksjon

Bruk av standardiserte måleinstrumenter for å fange opp ulike nivåer av funksjon står sentralt både innen klinisk praksis samt innen helsefaglig forskning. Måleinstrumenter kan være med å danne bakgrunn for både diagnose, prognose og for å kunne vurdere

effekt av tiltak (de Vet et al., 2011, s. 1). En kombinasjon av spørreskjema og fysiske tester ser ut til å være den beste metoden for å undersøke funksjonsevne hos pasienter med muskel og skjelettplager (Wind, Gouttebarge, Kuijer, & Frings-Dresen, 2005).

2.3.1 Spørreskjema

Spørreskjema er et måleinstrument som kan fange opp hvordan pasienter erfarer muskel- og skjelettplagene sine og hvilke konsekvenser plagene får for dem (Wittink, 2005). Det finnes mange forskjellig spørreskjema om muskel- og skjelettplager (Wind et al., 2005). En fordel med spørreskjema er at de egner seg godt til å brukes både i klinikk og forskning (Waddell, 2004, s. 39) . Som regel er de raske, rimelige og enkle å administrere. Ulempene ved å bruke spørreskjema kan være at de er mer påvirket av pasientene sin psykologiske status enn fysiske tester, og kan således gi et upresist inntrykk av funksjonsnivået (Wand, Chiffelle, O'Connell, McAuley, & Desouza, 2010). Skjemaene må ha gode måleegenskaper hvis de skal benyttes, spesielt innen forskning. I den systematiske oversikten utført av Wind et al. (2005), viste tre spørreskjemaer som er utviklet for korsryggsmerter; The Pain Disability Index (Tait, Chibnall, & Krause, 1990), The Oswestry Disability Index (ODI) (Fairbank, Couper, Davies, & O'Brien, 1980) og Roland Morris Disability Questionnaire (Roland & Morris, 1983) å ha både høy reliabilitet og validitet. Andre spørreskjema viste god validitet, men dårlig reliabilitet (Wind et al., 2005).

2.3.2 Fysiske tester

I tillegg til spørreskjema bør man ha en direkte måte å måle funksjon, i form av standardiserte fysiske tester (Wittink, 2005). Fysiske tester kan deles inn i henhold til de ulike funksjonsnivåene i ICF. Skal for eksempel bevegelighet, styrke og utholdenhet undersøkes, vil slike tester defineres innenfor kroppsstruktur- og funksjonsnivå. Biering-Sørensen test er eksempel på en test innenfor dette nivået (Biering-Sorensen, 1984). Fysiske tester som innebærer å løfte, gå eller plukke noe opp fra gulvet er tester på aktivitetsnivå i henhold til ICF (Wittink, 2005). Back Performance Scale er eksempel på et testbatteri med flere deltester, som undersøker funksjon på dette aktivitetsnivået i ICF (Strand, Moe-Nilssen, & Ljunggren, 2002). Testene som utføres på aktivitetsnivå,

beskrives ofte som funksjonstester eller ”performence” tester. Ofte settes flere fysiske tester sammen som et testbatteri. Et testbatteri kan være hensiktsmessig å bruke for å vurdere funksjon med hensyn til arbeidsevne (Wind et al., 2005) og gjør det mulig å undersøke funksjon på flere nivå i ICF ved å inkludere tester på både kropsstruktur- og funkjsonsnivå samt aktivitetsnivå. For at fysiske tester skal brukes i forskning og en klinikk hverdag, bør de være lette å gjennomføre, ikke være for tidkrevende, og kreve lite avansert utstyr (Main et al., 2008) i tillegg bør de ha gode måleegenskaper (de Vet et al., 2011).

2.4 Måleegenskaper

Undersøkelse av måleegenskaper omhandler måleinstrumentets kvalitet, hvor primært vurdering av reliabilitet og validitet står sentralt (Polit & Beck, 2012, s. 739). Kvalitet på et måleinstrument generelt, må sees i sammenheng med kontekst, hvilken gruppe som studeres og hensikt (de Vet et al., 2011, s. 301). Målet til metodologiske studier er å dokumentere og forbedre reliabilitet, validitet og evne til å fange opp endring (”responsiveness”) til måleinstrument som blir brukt. Mokkink et al. (2010) påpekte at det manglet konsensus på taksonomi, begreper og definisjoner for måleegenskaper, og gjennomførte en stor internasjonal Delphi studie hvor konsensus ble oppnådd på taksonomi, begreper og definisjoner for måleegenskaper. Validitet ble definert som ”The degree to which an instrument measures the construct(s) it purports to measure” og ”responsiveness” ble definert som ”the ability of an instrument to detect change over time in the construct to be measured” (Mokkink et al., 2010). Kottner et al. (2011) har utarbeidet retningslinjer for rapportering av reliabilitet og enighet (agreement) i studier. I denne studien er det reliabilitet som blir undersøkt, og andre måleegenskaper som validitet og ”responsiveness” vil ikke bli gjort nærmere rede for.

2.4.1 Reliabilitet

Et viktig krav til alle måleinstrument som blir brukt i forskning og i klinisk arbeid er at de er reliable (de Vet et al., 2011, s. 96). I litteraturen beskrives reliabilitet med forskjellige termer, som for eksempel; repeterbarhet, reproduserbarhet, presisjon, variabilitet, konsistens, stabilitet og enighet (Carter, Lubinsky, & Domholdt, 2011, s.

237; de Vet et al., 2011, s. 97). Reliabilitet defineres som i hvilken grad testskårene er fri for målefeil (Carter et al., 2011, s. 237; de Vet et al., 2011, s. 97). I tillegg ble det i den internasjonale studien til Mokkink et al. (2010) enighet om at en utvidet definisjon av reliabilitet er ”the extent to which scores for patients who have not changed are the same for repeated measurments under several conditions: e.g. using different sets of items from the same multi-item measurment instrument (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by same persons (i.e raters or responders) on different occasions (intra-rater)”. Reliabilitet er en karakteristikk av måleinstrumentet brukt i en populasjon (de Vet et al., 2011, s. 102) og er påvirket av flere faktorer inkludert kilder til variabilitet, utvalg og variasjon i skår (Carter et al., 2011, s. 245). Reliabilitet undersøkes ved å finne ut forholdet mellom to mål, og størrelsen på forskjellen mellom de to målene. Dette betegnes som relativ og absolutt reliabilitet (Carter et al., 2011, s. 326).

Intertester, intratester og test-retest reliabilitet

Som det fremgår av definisjonen til Mokkink et al. (2010) kan man måle ulike former for reliabilitet, avhengig av om man sammenligner skårene fra samme tester, forskjellige testere og skårer på ulike tidspunkt. Når skårene til samme tester blir sammenlignet er det snakk om intratester reliabilitet, når skårer fra to ulike tidspunkt sammenlignes er det test-retest reliabilitet, og når resultater fra to ulike testere sammenlignes er det intertester reliabilitet.

Intertester reliabilitet

En streng definisjon av intertester reliabilitet er at det er ”consistency of performance among different raters or judges in assigning scores to the same objects or responses….[It] is determined when two or more raters judge the performance of one group of subjects at the same time” (Carter et al., 2011, s. 239). Når to testere kan skåre en test samtidig, enten ved å være tilstede under samme testseanse eller skåre samme testseanse som er tatt opp på video, kan en del målefeil bli eliminert (Portney & Watkins, 2009, s.101-102), mens når to ulike testere tester på to forskjellige tidspunkt kan forskjeller i resultatene komme av for eksempel variasjon i deltakerne sin utførelse, testerne og måleinstrumentet som blir brukt (Moe-Nilssen, Nordin, & Lundin-Olsson,

2008). For mange fysioterapeuter må en testseanse foregå på to forskjellige tidspunkt da klinikeren må tolke for eksempel berøring, trykk og/eller motstand. I disse situasjonene vil det ikke være mulig å bruke video eller at to terapeuter skårer en test samtidig.

2.4.2 Relativ reliabilitet

Relativ reliabilitet er basert på ideen om at hvis målene er reliabel, vil målene til den enkelte innen en gruppe opprettholde sin posisjon i gruppen ved repeterte målinger. Det betyr at om ikke deltakerne får akkurat samme skår ved begge testomgangene, vil man forvente at de som skårer høyest ved 1. testing vil også skåre høyest ved 2. testing, og de som skårer lavest på 1. testing vil også skåre lavest på 2. testing (Carter et al., 2011, s. 239). Relativ reliabilitet blir målt ved å beregne en korrelasjonskoeffisient.

Korrelasjonskoeffisienter

En korrelasjonskoeffisient indikerer grad av sammenheng mellom repeterte målinger. Det er flere forskjellige typer korrelasjons koeffisienter, som for eksempel, Intraclass correlation coefficient (ICC), Pearson's r, Kappa og Spearman's rho. De ulike statistiske metodene man kan bruke er avhengig av målenivå og egenskaper ved datamaterialet (Carter et al., 2011, s. 237). En korrelasjonskoeffisient på 1.0 indikerer perfekt sammenheng ved gjentatte målinger (Carter et al., 2011, s. 240). Carter et al. (2011) trekker frem en rekke faktorer som kan påvirke de ulike korrelasjonkoeffisientene og gi ulike resultat for samme data. Blant annet er korrelasjonskoeffisienten påvirket av om man gjør undersøkelsene individuelt eller på en gruppe. En del av korrelasjonskoeffisientene er dårlig til å avdekke systematiske feil, som f.eks. Pearson's r og Spearman's rho (de Vet et al., 2011, s. 110). I tillegg er korrelasjonskoeffisienten påvirket av spekteret av skår på måleskalaen. Det vil si at hvis mange deltakere skårer likt vil korrelasjonskoeffisienten blir lavere enn hvis deltakerne skårer spredt på måleskalaen. Gulv- og takeffekter er tilstede dersom 15% av deltakerne får henholdsvis laveste eller høyeste skår i et utvalg bestående av minst 50 pasienter (Terwee et al., 2007). Dette er forhold som vil påvirke korrelasjonskoeffisienten grunnet lite variasjon i skår. De forskjellige korrelasjonskoeffisientene kan gi ulike resultat for samme data, og hva som regnes som akseptabel, god eller dårlig reliabilitet er dermed vanskelig å si uten at man har mer inngående kjennskap til gruppen som blir undersøkt,

design, variasjon i skår og hvilken korrelasjonskoeffisient som er brukt (Carter et al., 2011).

Pearson's r og ICC brukes ofte for kontinuerlig skala. ICC er foretrukket fremfor Pearson's r da sistnevnte kun tar hensyn til tilfeldige feil og ikke systematiske feil (de Vet et al., 2011, s. 110). Av den grunn oppnår ofte Pearson's r en høyere korrelasjonsverdi enn ICC om det er systematiske feil til stede. Hvis det ikke er noe systematiske feil, vil $ICC_{agreement}$ og Pearson's r være lik (de Vet et al., 2011, s. 110). I utgangspunktet er parametrisk statistikk foretrukket og dette forutsetter at data er normalfordelt. Hvis ikke data er normalfordelt er ikke-parametrisk statistikk et alternativ, som for eksempel Spearman's rho (Pallant, 2013). For nominale verdier er Kappa statistikk anbefalt (Carter et al., 2011, s. 237).

Intraclass correlation coefficients (ICC)

Den anbefalte korrelasjonskoeffisienten for å undersøke reliabilitet er ICC (Carter et al., 2011; Rankin & Stokes, 1998). Det er flere ulike ICC modeller og hvilken ICC modell som egner seg er blant annet avhengig av kjennetegn ved testerne, deltakerne, samt studien sin utforming (Shrout, 1979). ICC verdier varier fra 0-1 hvor liten målefeil og stor variasjon i skår gir en reliabilitet tilnærmet 1 (de Vet et al., 2011, s. 120). $ICC_{2.1}$ betegnes som en hensiktsmessig modell for intertester design når absolutt enighet skal evalueres (Shrout, 1979). Siden systematiske forskjeller er ansett for å være en del av målefeil er $ICC_{agreement}$ (two way random effects model) foretrukket (Terwee et al., 2007). Relativ reliabilitet med korrelasjonskoeffisient gir begrenset informasjon og bør derfor suppleres med parameter for absolutt reliabilitet (Carter et al., 2011, s. 240)

2.4.3 Absolutt reliabilitet

Absolutt reliabilitet er et uttrykk for i hvilken grad en skår varierer ved gjentatte målinger og blir ofte referert til som målefeil (Carter et al., 2011, s. 240). Parameter for målefeil er av stor verdi for klinikere fordi målefeilen blir oppgitt i samme måleenhet som instrumentet (de Vet et al., 2011, s. 145). Målefeil kan uttrykkes ved Standard error of measurment (SEM) (de Vet et al., 2011, s. 122-123) eller Within-subject standard

deviation (S_w) (Bland & Altman, 1996). Bland-Altman plott er en grafisk fremstilling og et alternativt parameter for å undersøke målefeil (Bland & Altman, 1986).

Målefeil blir av Mokkink et al. (2010) definert som ”den systematiske og tilfeldige feilen i en pasient sin skår som ikke kan forklares med bakgrunn i ”virkelig” endring i fenomenet som blir målt”. Systematiske feil er forutsigbare feil som skjer i en retning hvor den ”sanne” skåren konsekvent blir overestimert eller underestimert. Per definisjon er de systematiske feilene konstante. Læringseffekt eller ”warm up” effekt er et eksempel på en systematisk feil og kan for eksempel avdekkes ved at nesten samtlige av deltakerne skårer bedre ved andre testing enn første testing. Tilfeldige målefeil skyldes tilfeldigheter og kan påvirke en deltaker sin skår på en uforutsigbar måte fra første til andre testing. Tilfeldige feil skjer på grunn av uforutsigbare faktorer som trøtthet, uoppmerksomhet, unøyaktighet eller enkle feil (Portney & Watkins, 2009, s. 92).

Det er flere faktorer som kan påvirke målefeilen, hvor målefeil blant annet kan variere med størrelsen på utvalget. Hvis man har et lite utvalg på for eksempel 10 deltagere, kan en ekstrem lav eller høy verdi (”outlier”) påvirke reliabiliteten og målefeil i stor grad (Carter et al., 2011, s. 266). En annen faktor er at målefeilen vil bli mindre om man tar flere målinger og tar gjennomsnitt av målingene (de Vet et al., 2011, s 108-110).

For at en skal kunne vite om pasienten har endret seg, er det nødvendig å vite hvor stor målefeil en kan forvente seg (Carter et al., 2011, s. 240). Smallest detectable change (SDC) er nært relatert til absolutt reliabilitet, men skiller seg fra målefeil ved at det er den minste forandringen som kan bli oppdaget i et måleinstrument, utover målefeilen (de Vet et al., 2011, s. 242). Flere navn er blitt nevnt i litteraturen for samme begrep, blant annet ”the smallest detectable difference” (SDD) (Moe-Nilssen et al., 2008) og ”minimal detectable change” (MCD) (Carter et al., 2011, s. 273).

2.5 Tidligere forskning

Vel vitende om at det er flere ulike fysiske tester som brukes til personer med muskel- og skjelettplager (Strand et al., 2011; Tveter, Dagfinrud, Moseng, & Holm, 2014), velger jeg i dette avsnittet å begrense presentasjon av tidligere forskning til de seks

testene som brukes i FAktA-prosjeket. Testene i FAktA-prosjektet er valgt ut på bakgrunn av at de er lett å gjennomføre, krever lite utstyr, lett å lære opp andre, og er aktuell for individer med forskjellige muskel- og skjelettplager. Det er et kort testbatteri og det er beregnet på at man kan bruke det i primærhelsetjeneste, eventuelt bedriftshelsetjeneste. Beskrivelse av de ulike testene er beskrevet i tabell 1.

Tabell 1. Beskrivelse av seks fysiske tester i FAktA-prosjektet og ICF nivå.

Fysiske funksjonstester	Beskrivelse av testen	Skår	ICF - Nivå
GBE-fleksibilitet (Kvale, Bunkan, Opjordsmoen, & Friis, 2012; Kvale, Ljunggren, & Johnsen, 2003)	Global kroppsundersøkelse (GBE) er ofte brukt på pasienter med langvarige muskel- og skjelettsmerter eller med psykosomatiske plager. Det er valgt ut 6 deltester fra GBE. Disse deltestene undersøker avspenningsevne og fleksibilitet og består av; *albu-slipp, lumbo-scaral fleksibilitet, hode rotasjon motstand, og motstand ved passiv hofte/knefleksjon, hofte sirkumduksjon og arm/skulder fleksjon.*	Hver enkelt deltest kan skåres på en skala fra 0 (ideell) til 7 (dårligst). Med seks deltester kan sum skår variere fra 0-42.	Kroppsfunksjon
Back Performance Scale (BPS) (Magnussen, Strand, & Lygren, 2004; Strand et al., 2002)	Back Performance Scale (BPS) består av 5 tester og er en test som er utviklet for å teste fysisk funksjon for pasienter med ryggsmerter. Den består av 5 tester; *sokketest, plukke opp test, rull-opp fra liggende, finger-tupp til gulv test* og *løfte-test av en kasse i 1 minutt fra gulv til midjehøyde.*	Testene skåres på en ordinal skala fra 0-3, hvor 0 indikerer ingen aktivitetsbegrensninger og 3 indikerer vesentlige aktivitetsbegrensninger. Sumskår på de fem testene til sammen er 0-15.	Aktivitet/ deltagelse
Høy løftetest	Høy løftetest er en modifisert test fra løftetesten i BPS. Den utføres ved å løfte en kasse (2 kg for kvinner og 3 kg for menn) fra midje til skulderhøyde og tilbake igjen med en valgfri løfteteknikk.	Antall ganger kassen blir løftet i løpet av 1 minutt blir registert.	Aktivitet/ deltagelse
Biering-Sørensen test (Biering-Sorensen, 1984; Demoulin, Vanderthommen, Duysens, & Crielaard, 2006; Keller, Hellesnes, & Brox, 2001; Latimer, Maher, Refshauge, & Colaco, 1999)	Utholdenhet av ryggekstensorene undersøkes ved at deltager ligger på magen på en benk med overkroppen utenfor benken og underkroppen fiksert til benken med tre stropper.	Deltager skåres på en skala fra 0-240 sekunder hvor lenge hun/han kan holde kroppen i en horisontal stilling.	Kroppsfunksjon
Dynamisk sit-up test (Oja & Tuxworth, 1995; Suni. J, 2009)	Undersøkelse av styrke og utholdenhet i magemusklene gjennomføres ved dynamisk sit-ups med tre nivå, hvor hvert nivå har økende krav. Deltakeren ligger på rygg med fleksjon i knærne mens føttene blir støttet av testeren.	Antall gjennomførte sit-ups blir registert (0-15)	Kroppsfunksjon
ACR-18 (Wolfe et al., 1990)	18 definerte punkter undersøkes etter kriteriene til American College of Rheumatology (ACR)	Antall smertefulle punkter telles opp.	Kroppsfunksjon

2.5.1 GFM-52

Kvale et al. (2003) har undersøkt intertester reliabilitet av 16 standardiserte bevegelsestester fra Global Fysioterapi Metode (GFM-52). Utvalget i studien bestod av 19 deltager; 10 friske og 9 pasienter sykemeldt for langvarige muskel- og skjelettplager. Intertester reliabilitet ble undersøkt mellom tre testere. Studien viste meget god intertester reliabilitet for domenet Bevegelse ($ICC_{2.1}$ 0.89). Reliabiliteten var god også for de fire sub-skalaene som inngår i domenet. $ICC_{2.1}$ viste 0.64 til 0.88, og målefeil for subskalaene viste ($S_w \leq 0.7$). Det gjøres oppmerksom på at deltester fra Global Fysioterapi metode (GFM-52) er slått sammen med deltester fra Den omfattende kroppsundersøkelsen (DOK) og nå heter Global kroppsundersøkelse (GBE) (Kvale et al., 2012).

2.5.2 Back Performance Scale (BPS)

Magnussen et al. (2004) undersøkte intertester reliabilitet av BPS på 32 pasienter med korsryggsmerter som hadde vart over 8 uker. Det var kun to deltagere fra utvalget som ikke var langtidssykemeldt, i tillegg til en deltaker som ikke var i arbeid . Det ble funnet meget god reliabilitet ($ICC_{2.1}$ 0.996) og liten målefeil S_w 0.25 av BPS. Test-retest reliabilitet ble undersøkt på 28 av deltagerne og var også høy $ICC_{2.1}$ 0.91, med målefeil S_w 1.3. I 95% av tilfellene er forskjellen mellom to målinger forventet å ligge mellom $\pm$ 3.6 poeng på BPS. Intertester reliabilitet ble undersøkt ved at to testere skåret samtidig, mens test-retest ble gjennomført med 2-3 dagers mellomrom. Konklusjonen på studien var at BPS er et reliabelt og valid måleinstrument for å måle aktivitetsbegrensninger hos pasienter med ryggplager (Magnussen et al., 2004).

I en annen studie utført av Strand et al. (2011) er test-retest reliabilitet av BPS utført på et mindre utvalg (n=9) av utvalget i en større studie. Deltagerne hadde langvarige korsryggsmerter. Test-retest ble utført før og etter et 3½ ukers rehabiliteringsopplegg, og reliabilitet ble undersøkt på deltagerne som oppga at de ikke hadde forandring på The patient Global Impression of Change (PGIC) etter rehabiliteringsopplegget var ferdig . Test-retest reliabilitet viste $ICC_{2.1}$ 0.89, målefeil uttrykt i SEM viste 1 og SDC viste 2.9.

2.5.3 Høy løftetest

Høy løftetest er en modifisert utgave av løftetesten i BPS Strand et al. (2002) og er ikke beskrevet i tidligere studier. Forskjellen på den originale versjonen og den modifiserte utgaven er at i den originale versjonen løftes en kasse fra gulv til midje og skåres på en ordinal skår. I den modifiserte utgaven løfter en deltager kassen fra midje til skulderhøyde og antall løft i løpet av 1 minutt blir registrert. I studien til Magnussen et al. (2004) har de undersøkt reliabilitet på deltesten løftetest i BPS. Siden deltestene er skåret på en ordinal skala fra 0-3 er hver deltest reliabilitet regnet ut ved hjelp av Kappa verdi. Deltesten Løftetest oppnår en kappaverdi på 1.0 for intertester reliabilitet og 0.55 for test-retest reliabilitet.

Strand et al. (2011) har undersøkt test-retest reliabilitet på 9 deltagere med langvarige korsryggsmerter som var inkludert i en "responsiveness" studie. Test-retest reliabilitet ble undersøkt på deltagerne i studien som oppga at de ikke hadde forandring på The patient Global Impression of Change (PGIC) etter et 3 ½ uker langt rehabiliteringsopplegg. To fysioterapeuter var testere i studien. Løftetesten i denne studien bestod i å løfte kassen fra gulv til midje, og antall løft i løpet av 1 minutt ble registrert. Test-retest reliabilitet i studien til Strand et al. (2011) viste $ICC_{2.1}$ 0.87, og målefeil uttrykt i SEM viste 2.4.

2.5.4 Biering-Sørensen test

Ulike studier har undersøkt reliabilitet av Biering-Sørensen test og funnet varierende grad av reliabilitet (Demoulin et al., 2006). I en studie av Keller et al. (2001) ble reliabilitet undersøkt på pasienter med korsryggsmerter (n=31) og friske personer (n=31) av samme tester med 5-10 dager mellom første og andre testing. ICC viste 0.93 for pasienter med korsryggsmerter og 0.80 for de friske deltagerne. I en annen studie av Latimer et al. (1999) ble det undersøkt reliabilitet på et utvalg bestående av tre grupper hvor 23 deltakere hadde korsryggsmerter, 20 deltakere hadde hatt en episode med korsryggsmerter og 20 deltagere var asymptomatiske. Deltagerne ble testet av to forskjellige testere ved to testseanser med 15 minutters mellomrom. Intertester reliabilitet viste $ICC_{1.1}$ 0.85 og SEM 15.6 sekunder for hele gruppen. Undersøkelse av

reliabilitet for de ulike gruppene viste ICC verdier fra 0.77 til 0.88 for og SEM 11.6 til 17.5.

2.5.5 Dynamisk sit-up test

Suni et al. (1996) har undersøkt intertester reliabilitet på isometrisk sit-up mellom tre testere gruppert i tre par. Utvalget bestod av 20 friske deltagere fra to forskjellige arbeidsplasser (10 menn og 10 kvinner) og det var 6-8 dager mellom første og andre testing. Intertester reliabilitet uttrykt i ICC viste 0.79 og målefeil uttrykt i SEM 20.1. Samlet for deltagerne var det gjennomsnittlig litt lavere testskår på test 1 (34.7 sekunder) enn test 2 (41.3 sekunder).

2.5.6 American College of Rheumatology (ACR-18)

American College of Rheumatology (ACR) har definert 18 ”tenderpoints” som har akseptabel sensitivitet og spesifisitet hos en gruppe pasienter med langvarige smertetilstander (Wolfe et al., 1990). Disse 18 punktene (kalt ACR kriteriene for fibromyalgi) benyttes i utredning av pasienter med utbredte muskelplager, og en studie av Weiner, Sakamoto, Perera, and Breuer (2006) viste utmerket intertester reliabilitet (0.84) uttrykt med Pearson’s r korrelasjonskoeffisient på 30 eldre deltagere med kroniske korsryggsmerter.

3 HENSIKT OG PROBLEMSTILLING

Det er utført reliabilitetsstudier av de fleste fysiske testene som inngår i FAktA-prosjektet tidligere, men det er ikke utført reliabilitetsundersøkelse på de seks fysiske testene utført samlet som et testbatteri. I de tidligere reliabilitetsstudiene har studiepopulasjonen vært langtidssykemeldte deltagere, deltagere med forskjellige type muskel- og skjelettplager og friske deltagere, mens i denne studiepopulasjonen er deltakerne arbeidstakere med muskel- og skjelettplager som står i fare for å bli sykemeldt eller har vært fullt sykemeldt i mindre enn 4 måneder på grunn av muskel- og skjelettplager.

Hensikten med denne studien er å undersøke intertester reliabilitet mellom tre testere på seks fysiske tester som inngår i funksjonsundersøkelsen i FAktA-prosjektet.

Følgende problemstilling har blitt undersøkt:
Hvordan er intertester reliabiliteten mellom tre testere på seks fysiske tester som blir benyttet i FAktA-prosjektet?

4 METODE

Denne studien er en metodestudie med tverrsnittdesign hvor intertester reliabilitet ble undersøkt. Studien ble gjennomført fra februar 2014 til august 2014.

4.1 Deltagere

Deltagerne i foreliggende studie er inkludert i en større studie kalt FAktA-prosjektet, ved Institutt for global helse og samfunnsmedisin ved Universitetet i Bergen. Rekruttering av deltagere ble gjort fortløpende til FAktA-prosjektet fra Helse og sosialavdelingen, arbeidstakere fra barnehage og skoler samt en tverrfaglig nakke og ryggpoliklinikk i Bergen Kommune. Deltagerne tok selv kontakt og avtalte time for funksjonsundersøkelsen i FAktA-prosjektet og ble spurt om å delta i intertester reliabilitetsstudien samme dag som funksjonsundersøkelsen skulle gjennomføres. Inklusjonskriteriene var deltagere med muskel- og skjelettplager, hovedsakelig på grunn av korsryggsmerter, nakke- og skuldersmerter eller utbredte smerter. Eksklusjonskriteriene var at deltagerne hadde vært sammenhengende fullt sykemeldt i mer enn 4 måneder samt utilstrekkelig norskkunnskap.

4.2 Testere

Tre fysioterapeuter, merket A, B og C, med lang erfaring i bruk av testene, byttet på å teste deltagerne i studien. Tester A og B testet 20 deltagere, tester A og C testet 14 deltagere og tester B og C testet 14 deltagere. Intertester reliabilitet for alle deltagerne samlet er merket med ABC. Parvis intertester reliabilitet er merket med AB, AC og BC.

4.3 Trening

For å sikre at testerne skulle ha lik oppfatning av prosedyre og skåringskriterier for testene, ble det gjennomført en treningsperiode i forkant av studien. Testerne hadde i tillegg en gjennomgang midtveis i studien hvor de på nytt gikk gjennom testene for å kalibrere seg.

4.4 Klinisk prosedyre

Datainnsamling ble gjennomført i henhold til en standardisert testprotokoll og ble utført ved at hver deltager ble testet 2 ganger med ca. 30 minutters pause. Hver testing tok til sammen ca. 20 minutter. For å forebygge at pasientene ble sliten og at dette kunne påvirke resultatene, ble testene som krevde minst innsats gjort først og de mest anstrengende testene utført til slutt. Deltagerne gjennomførte Numerical Pain Rating Scale (NPRS) i forkant av første testing. Mellom første og andre testing besvarte deltagerne standardiserte spørreskjema og skjema for personlig bakgrunnsinformasjon. I forkant av andre testing fylte de på nytt ut NPRS. Hvis deltageren sin skår gikk opp mer enn et nivå mellom testseansene, ble han/hun ekskludert. Resultatene fra testene ble skrevet ned på skåringskjema (vedlegg 1), lagt i en konvolutt og oppbevart i en låst skuff. For å unngå systematiske forskjeller var det utarbeidet et skjema som beskrev hvem av testerne som skulle teste først og sist (vedlegg 2). Testerne prøvde å følge skjemaet, men av praktisk grunner lot det seg ikke gjennomføre for alle deltagerne.

4.5 Testene

4.5.1 Selvrapporterte mål

Numerical Pain Rating Scale

NPRS er et selvrapportert mål på smerteintensitet (vedlegg 4). Det er en 11-punkts skala hvor deltakeren må rangere smertene sine på en skala fra 0-10 ved å bruke hele nummer, hvor 0 representerer ”ingen smerte” og 10 representerer ”den verst tenkelige smerte” (Jensen, Karoly, & Braver, 1986).

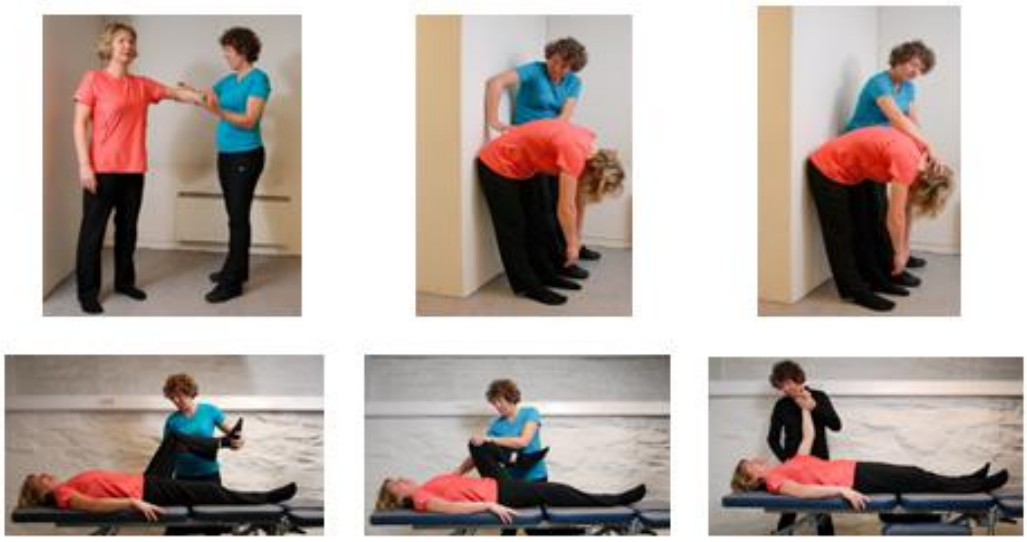
4.5.2 Fysiske tester

De fysiske testene i FAktA-prosjektet er valgt for å gi et generelt inntrykk av fysisk funksjon. Sammen med spørreskjema reflekterer de ulike dimensjoner av ICF-modellen. De seks fysiske testene som ble gjennomført var GBE - fleksibilitet (seks deltester fra GBE), BPS, høy løftetest, Biering-Sørensen test, dynamisk sit-up test og ACR-18. Det ble i tillegg utført en test for nakke/skulder mobilitet. Ut fra erfaring underveis i FAktA-prosjektet, ble denne testen ikke funnet relevant å bruke for alle

diagnosegrupper og ble derfor ekskludert. Den er dermed ikke med i intertester reliabilitetsstudien og vil ikke bli beskrevet noe videre. De seks testene er beskrevet i Tabell 1.

GBE-Fleksibilitet

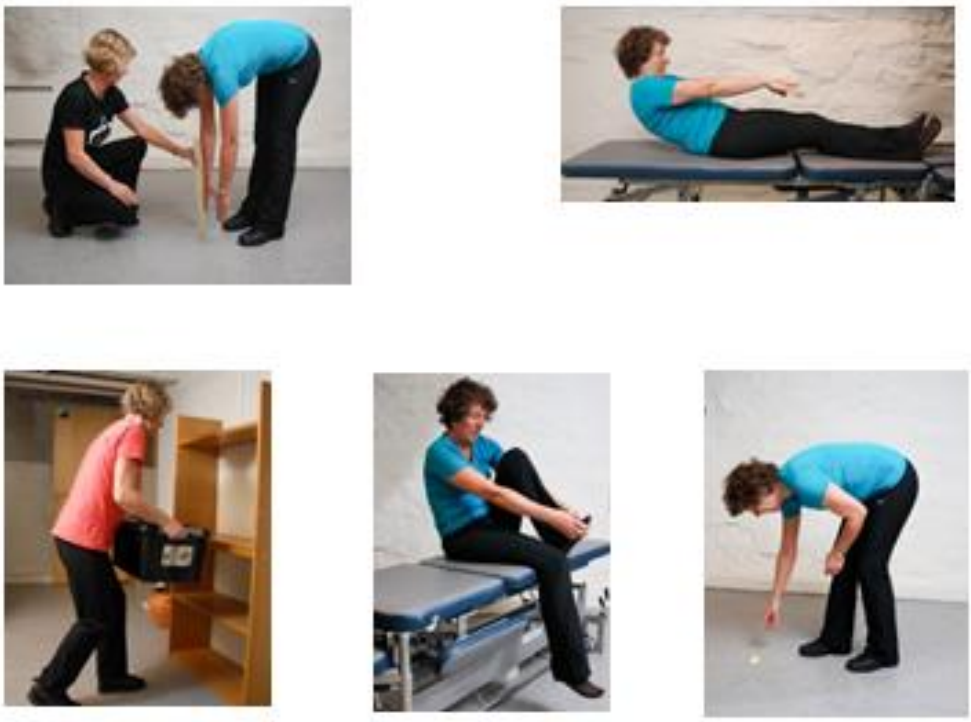
Det ble valgt ut seks deltester fra GBE (Kvale et al., 2012). Disse deltestene var; albu-slipp, lumbo-sacral fleksibilitet, hode rotasjon motstand, og motstand ved passiv hofte/knefleksjon, hofte sirkumduksjon og arm/skulder fleksjon. Hver enkelt deltest ble skåret på en skala fra 0 (ideell) til 7 (dårligst), slik at sumskår kunne variere fra 0-42.



Illustrasjon 1. *GBE-Fleksibilitet*

BPS

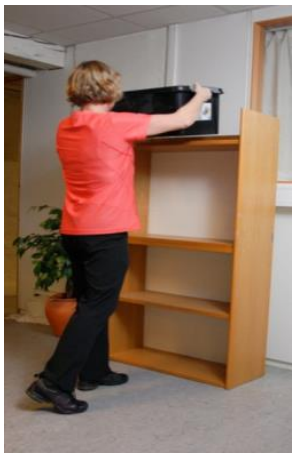
I BPS inngår de fem testene; sokketest, plukke opp test, rull-opp fra liggende, finger-tupp til gulv test og løftetest av en kasse i 1 minutt fra gulv til midjehøyde (Strand et al., 2002). Deltestene ble skåret på en ordinal skala fra 0-3, hvor 0 indikerer ingen aktivitetsbegrensninger og 3 indikerer vesentlige aktivitetsbegrensninger. Sumskår på de fem testene varierte fra 0-15.



Illustrasjon 2. *Back Performance Scale*

Høy løftetest

Høy løftetest er en modifisert test fra løftetesten i BPS og ikke beskrevet tidligere. Den ble utført ved at deltager løftet en kasse (2 kg for kvinner og 3 kg for menn) fra midje til skulderhøyde og tilbake igjen med en valgfri løfteteknikk. Antall ganger kassen ble løftet i løpet av 1 minutt ble registrert.



Illustrasjon 3. *Høy løftetest*

Biering-Sørensen test

Utholdenhet av ryggekstensorene ble undersøkt med Biering-Sørensen test (Biering-Sorensen, 1984; Demoulin et al., 2006; Keller et al., 2001; Latimer et al., 1999). Deltager lå på magen på en benk med overkroppen utenfor benken og underkroppen fiksert til benken med tre stropper. Tiden ble tatt på hvor lenge deltageren klarte å holde kroppen i en horisontal stilling fra 0-240 sekunder.

Illustrasjon 4. *Biering-Sørensen test*

Dynamisk sit-up test

Undersøkelse av styrke og utholdenhet i magemusklene ble gjennomført med en dynamisk sit-up test (Oja & Tuxworth, 1995; Suni. J, 2009). Den ble gjennomført med tre nivå, hvor hvert nivå hadde økende krav. Deltakeren lå på rygg med fleksjon i knærne mens føttene ble støttet av testeren. Antall gjennomførte sit-ups ble registrert (0-15).

Illustrasjon 5. *Dynamisk sit-up test*

ACR-18

18 definerte punkter ble undersøkt etter klassifiseringskriteriene til ACR (Wolfe et al., 1990) og antall smertefulle punkter registrert.

Illustrasjon 6. *ACR-18*

4.6 Analyse

Deskriptiv statistikk ble brukt til å beskrive utvalget i studien og normalfordeling ble undersøkt på datamaterialet ved å bruke Kolmogorov-Smirnof test (Pallant, 2013). Både ”relativ reliabilitet” og ”absolutt reliabilitet” er anbefalt når man undersøker intertester reliabilitet (Rankin & Stokes, 1998). I foreliggende studie er relativ reliabilitet kalkulert ved modellen $ICC_{2.1}$ (Shrout, 1979). Denne ICC modellen er tilsvarende til ANOVA modellen kalt ”two way random with absolute agreement i SPSS og $ICC_{2.1}$ verdien for ”single measurment” er benyttet i studien. Hvis Kolmogorov-Smirnof test viste at noen av variablene ikke var normalfordelt ble det vurdert å benytte Spearman’s rho.

ICC verdien er et utrykk for hvor god korrelasjon eller samsvar det er mellom to testinger uttrykt med en koeffisient, og sier ikke noe om størrelse på enighet. S_w er oppgitt i samme måleenhet som skåren og gir dermed en viktig opplysning med tanke på å vurdere om det er store eller små målefeil. I denne studien ble størrelsen på enighet angitt ved ”within-subject standard deviation” ($S_{w)}$. En analyse av varians (ANOVA) ble brukt for å regne ut S_w ved å ta kvadratroten av ”the within-people total mean square”. Lav S_w uttrykker en lav grad av målefeil (Bland & Altman, 1996). Forskjellen mellom skåren til en deltager utført av en tester og den sanne verdi er forventet å være mindre enn 1.96 S_w for 95% av observasjonene. For målefeil ble den tilsvarende prosentdelen av S_w relatert til totalskåren av skalaen på måleinstrumentet kalkulert.

Tolkning av S_w etter Ostelo, de Vet, Knol, and van den Brandt (2004) sin inndeling;

S_w	Grad av målefeil
≤5%	liten
>5% og ≤10%	moderat
>10% og <20%	stor
≥20%	betydelig

Forskjellen mellom to målinger for samme deltager er forventet å være mindre enn $\sqrt{2}$ x 1.96 x S_w = 2.77S_w for 95 % av de parvise observasjonene (Bland & Altman, 1996). Denne verdien som er nært relatert til absolutt reliabilitet blir definert som "den minste forandringen som kan bli oppdaget i et måleinstrument, utover målefeilen" og er referert til som SDC (de Vet et al., 2011, s. 258). Denne verdien er oppgitt i samme måleenhet som skåren og gir informasjon om en deltaker har endret seg utover målefeilen.

Bland-Altman plot ble brukt som et grafisk verktøy for å illustrere absolutt reliabilitet (Bland & Altman, 1986). Systematiske forskjeller ble registrert og vist visuelt ved den grafiske fremstillingen kalt "limits of agreement", dvs. "grense for enighet". For hver deltager ble gjennomsnittsforskjellen mellom de to skåringene plottet på x-aksen og differansen på de to skåringene på y-aksen. Verdiene fra en paret t-test uttrykt i gjennomsnittsforskjell (d) og differanse på standardavvik ($SD_{difference}$) kan brukes til å regne ut "grense for enighet" med ligningen ($d \pm 1.96SD_{difference}$) (de Vet et al., 2011, s. 114).

Det er ikke enighet om hva som er en akseptabel verdi for reliabilitet, men i litteraturen er ofte 0.70 eller høyere beskrevet som akseptabel verdi for reliabilitet (de Vet et al., 2011). I denne studien er reliabilitet tolket etter Munro's beskrivende termer for styrke på korrelasjons koeffisienter (Carter et al., 2011).

Tolkning av reliabilitet, uttrykt ved ICC verdier etter Munro's beskrivende termer for styrke på korrelasjons koeffisienter (Carter et al., 2011, s. 318).

ICC verdi	Styrke på enighet
,00-,25	Liten, hvis noen korrelasjon
,26-,49	Lav korrelasjon
,50-,69	Moderat korrelasjon
,70-,89	Høy korrelasjon
,90-1,00	Svært høy korrelasjon

Analysene ble gjennomført ved bruk av SPSS (Statistical Package for Social Sciences, Chicago, IL, USA), versjon 21.

4.7 Etikk

Studien har vært gjennomført i henhold til Helsinki deklarasjonen og ble godkjent av Regional komité for medisinsk og helsefaglig forskningsetikk (REK) via FAktA-prosjektet (vedlegg 5). Det har vært frivillig å delta i studien. Deltagerne fikk skriftlig og muntlig informasjon om prosjektet i forkant av studien, før de skrev under informert samtykke (vedlegg 6). Enhver deltager har hatt mulighet for å trekke seg når som helst, uten å måtte begrunne dette. Det har ikke vært noe økonomiske fordeler eller ulemper med å være med i studien. Deltagerne måtte bruke ca. 20 minutter ekstra (i forhold til den vanlige funksjonsvurderingen i FAktA-prosjektet) ved å være med på reliabilitetsstudien.

5 RESULTATER

I alt deltok 52 arbeidstakere i intertester reliabilitetsstudien. To pasienter gjennomførte ikke andre testing; én på grunn av hodepine, og den andre på grunn av rapportering om økt smerte fra 3 til 7 på NPRS. I tillegg ble ytterligere to deltagere ekskludert under plotting av data, på grunn av manglende NPRS skjema fra første og andre testing. To deltagere manglet NPRS skjema fra andre testing, men ble inkludert i studien av testerne. I disse to tilfellene ble skåren fra første testing brukt som ”least measured carried forward”.

5.1 Demografi

Oversikt over demografiske variabler er gitt i tabell 2. Totalt 48 deltagere ble inkludert i studien, 38 kvinner (79%) og 10 menn (21%). Deltagerne var mellom 23-65 år, med en gjennomsnittsalder på 46 år. Deltagerne som hadde utbredte smerter utgjorde den største gruppen med 21 deltakere. Ni deltagere var klassifisert med korsryggsmerter og 18 var klassifisert med nakke og skulder smerter. Ti deltagere var sykemeldt og 38 deltagere var ikke sykemeldt.

Tabell 2. Bakgrunnsdata for utvalget (n=48)

Bakgrunnsvariabler	
Kjønn n (%)	
Kvinner	38 (79)
Menn	10 (21)
Alder i år: gjennomsnitt (min-max)	46 (23-65)
Hovedplage n (%)	
Nakke- og skuldersmerte	18 (37)
Korsryggsmerter	9 (19)
Utbredte smerter	21 (44)
Jobbstatus n (%)	
Ikke sykemeldt	38 (79)
Sykemeldt	10 (21)
NPRS: gjennomsnitt	3.9
BMI (n=43) gjennomsnitt (max-min)	25 (21-33)

Resultatene fra Kolmogorov-Smirnof test viste at noen av variablene ikke var normalfordelt. ICC verdien til disse variablene ble sammenlignet med det ikke parametriske alternativet Spearman's rho for å få et inntrykk av systematisk drift.

5.2 Resultat intertester reliabilitet

5.2.1 Intertester reliabilitet av GBE-Fleksibilitet

Relativ intertester reliabilitet (ABC) for 48 deltagere viste høy reliabilitet ($ICC_{2.1}$ 0.88) (tabell 3 og figur 2). Absolutt intertester reliabilitet (ABC) for 48 deltagere er gitt ved S_w og er uttrykt som målefeil. Ut i fra ligningen $\sqrt{}$7.34, ble S_w regnet ut til 2.71 (tabell 3). Forskjellen mellom én deltaker sin skår og den "virkelige" måleverdi er forventet å være mindre enn 5.31 poeng for 95% av skårene (±1.96 x 2.71). Den minste påviselige endring uttrykt som SDC mellom to målinger fra samme deltaker er regnet ut ved $2.77S_w$ og var i dette tilfellet 7.50.

Bland-Altman plot for intertester reliabilitet (ABC) for 48 deltagere er vist i figur 3. Den viser at gjennomsnittlig forskjell (d) i intertester enighet fra første til andre test var på 1.06 poeng. Målefeil uttrykt som "grense for enighet" er regnet ut ved d ± 1.96 x SD av forskjellen = 1.06 ± 1.96 x 3.72 = 8.36/-6.23. Det ble registrert 2 deltagere "outliere" som ligger utenfor "grense for enighet" som har ført til større målefeil (figur 3). Totalt 95.8 % var innenfor "grense for enighet". Gjennomsnittskår var litt lavere for andre enn første test (tabell 3). Bland-Altman plot viser at et flertall av deltagerne har skåret bedre ved andre enn første test (figur 3).

Parvis intertester reliabilitet (AB – AC – BC) viste høy korrelasjon ($ICC_{2.1} \geq 0.86$). Målefeil varierte fra S_w 2.55 til 2.98. Resultatene er presentert i tabell 4.

Tabell 3. Intertester reliabilitet og målefeil mellom tre testere (ABC) for seks fysiske tester n=48.

Fysisk test	**Test1***	**Test2***	**$ICC_{2.1}$¶[1]**	**S_w[2]**	**$SDC_{95\%}$**
GBE-Fleksibilitet	16.94	15.88	0.88 (0.80-0.93)	2.71	7.50
Back Performance Scale	3.69	3.63	0.94 (0.90-0.97)	0.88	2.16
Høy løftetest	14.73	16.25	0.88 (0.63-0.95)	1.84	5.1
Biering-Sørensen	79.33	72.27	0.80 (0.69-0.88)	21.67	60
Dynamisk sit-up test	11.44	11.92	0.80 (0.67-0.88)	2.01	5.57
ACR-18	6.21	6.75	0.82 (0.71-0.90)	2.04	5.65

* Verdien er gjennomsnittsverdi

¶ Verdien er confidence intervall

ICC[1] = Intraclass correlation coefficient

S_w[2] = Within-subject standard deviation

SDC[3] = Smallest detectable change

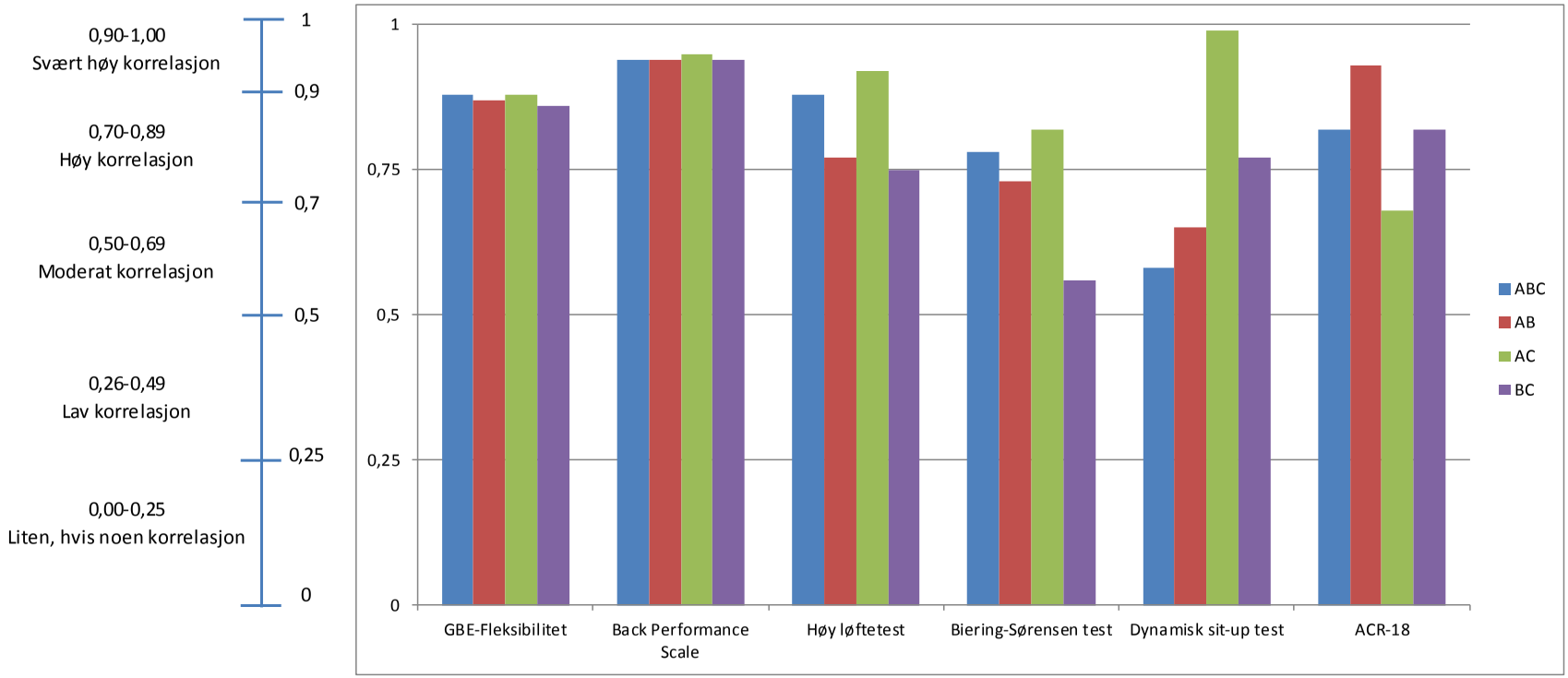
Tabell 4. Parvis intertester reliabilitet og målefeil for seks fysiske tester AB (n=20) AC (n=14) og BC (n=14).

	AB			**AC**			**BC**		
Fysisk test	ICC[1]	CI[2]	S_w[3]	ICC[1]	CI[2]	S_w[3]	ICC[1]	CI[2]	S_w[3]
GBE-Fleksibilitet	0.87	0.70-0.95	2.62	0.88	0.61-0.96	2.55	0.86	0.62-0.95	2.98
Back Performance Scale	0.94	0.83-0.97	0.67	0.95	0.85-0.98	0.84	0.94	0.84-0.98	1.15
Høy løftetest	0.77	0.40-0.91	2.04	0.93	0.55-0.98	1.60	0.92	0.72-0.97	1.75
Biering-Sørensen	0.74	0.46-0.89	23.49	0.82	0.52-0.94	21.32	0.85	0.59-0.95	19.17
Dynamisk sit-up test	0.65	0.27-0.84	2.25	0.99	0.96-0.97	0.5	0.78	0.46-0.92	2.53
ACR-18	0.93	0.83-0.97	1.29	0.68	0.26-0.88	2.67	0.82	0.56-0.94	2.19

ICC[1]:

CI[2] = Confidence interval

S_w[3]= Within-subject standard deviation



***Figur 2**. Relativ intertester reliabilitet vist for seks fysiske tester; GBE-Fleksibilitet, Back Performance Scale (BPS), høy løftetest, Biering-Sørensen test (B-S), dynamisk sit-up test og ACR-18 for 48 deltagere (ABC) og parvis (AB – AC – BC).*

5.2.2 Intertester reliabilitet av BPS

Relativ intertester reliabilitet (ABC) for 48 deltagere viste svært høy korrelasjon ($ICC_{2.1}$ 0.94) (tabell 3 og figur 2). Variablene var ikke normalfordelt og $ICC_{2.1}$ ble sammenlignet med Spearman's rho. Forskjell mellom $ICC_{2.1}$ og Spearman's rho viste en minimal forskjell; (0.94-0.92 = 0.02), og det ble derfor valgt å bruke $ICC_{2.1.}$ Absolutt intertester reliabilitet (ABC) for 48 deltagere viste moderat grad av målefeil S_w 0.88 (5.87% av måleskalaen) (tabell 3).

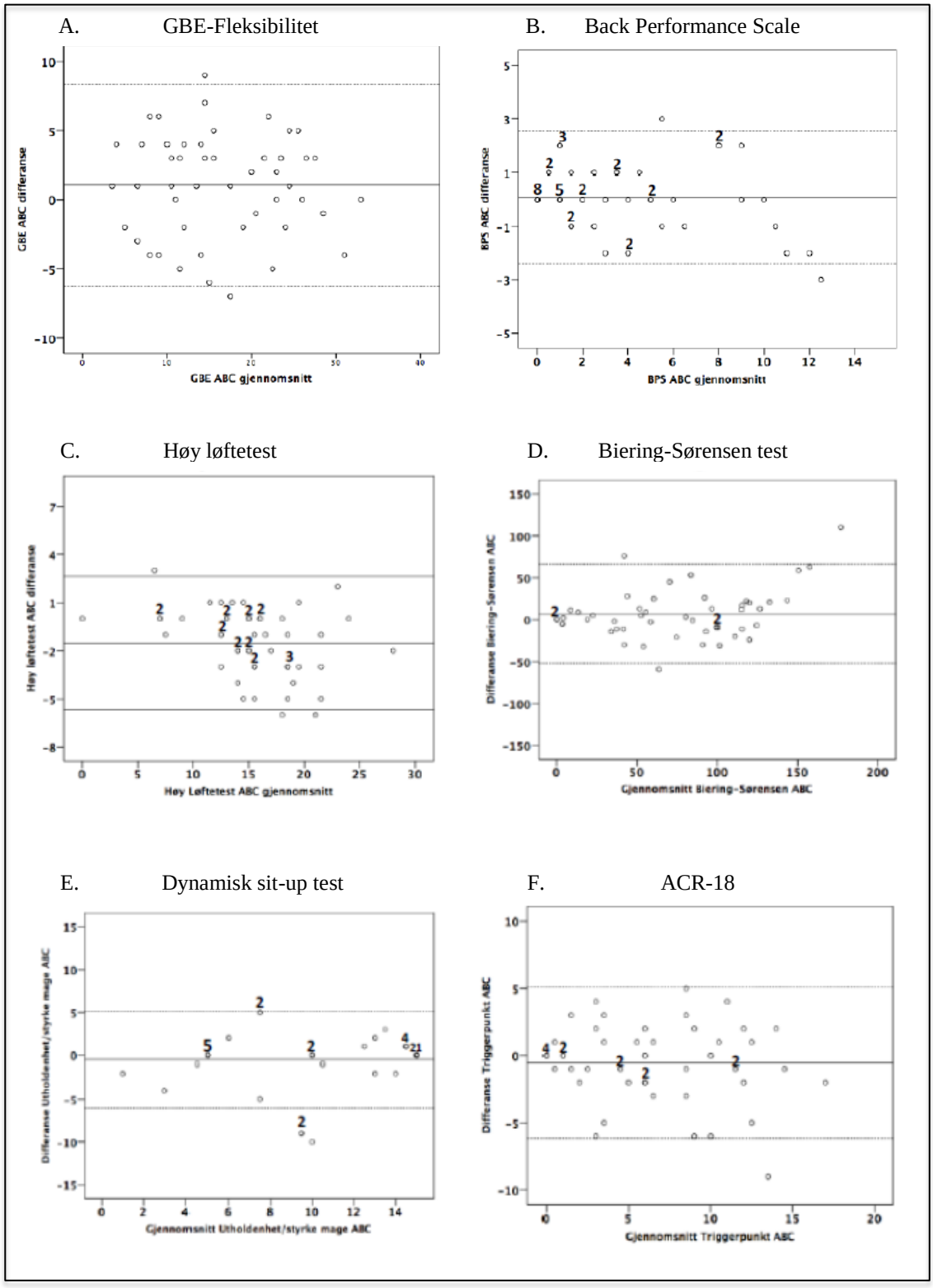
Bland-Altman plot for intertester reliabilitet (ABC) for 48 deltagere er vist i figur 3. Gjennomsnittlig forskjell i intertester enighet var på 0.06 poeng og "grense for enighet" viste 2.53/-2.41. Plottet viste 2 "outliere" som ligger utenfor "grense for enighet". Totalt 95.8 % var innenfor "grense for enighet". Et overtall av deltagerne har skåret på nedre del av skalaen.

Parvis intertester reliabilitet (AB – AC – BC) viste svært høy korrelasjon ($ICC_{2.1} \geq 0.86$) og liten grad til moderat grad av målefeil (S_w 0.67 til 1.15). Resultatene er presentert i tabell 4. Variabler for AB og BC var ikke normalfordelt, men forskjell mellom ICC og Spearman's rho viste liten forskjell ≤0.02.

5.2.3 Intertester reliabilitet av høy løftetest

Relativ intertester reliabilitet (ABC) for 48 deltagere viste høy korrelasjon $ICC_{2.1}$ 0.88 (tabell 3 og figur 2). Målefeil viste S_w 1.84. Bland-Altman plot er grafisk fremstilt i figur 3. Gjennomsnittlig forskjell i intertester enighet var på -1.52 poeng. "Grense for enighet" viste 2.65/-5.69. Bland-Altman plot viste 3 "outliere" og totalt 90 % av deltakerne var innenfor "grense for enighet". Et overtall av deltagerne skåret høyere ved andre testing enn første testing (figur 3).

Parvis intertester reliabilitet (AB – AC – BC) viste høy korrelasjon for tester AB ($ICC_{2.1}$ 0.77) og svært høy korrelasjon for AC ($ICC_{2.1}$ 0.93) og BC ($ICC_{2.1}$ 0.92). Tester AC viste lavest målefeil med S_w 1.60 og AB hadde høyest målefeil med S_w 2.04. Resultatene er presentert i tabell 4.



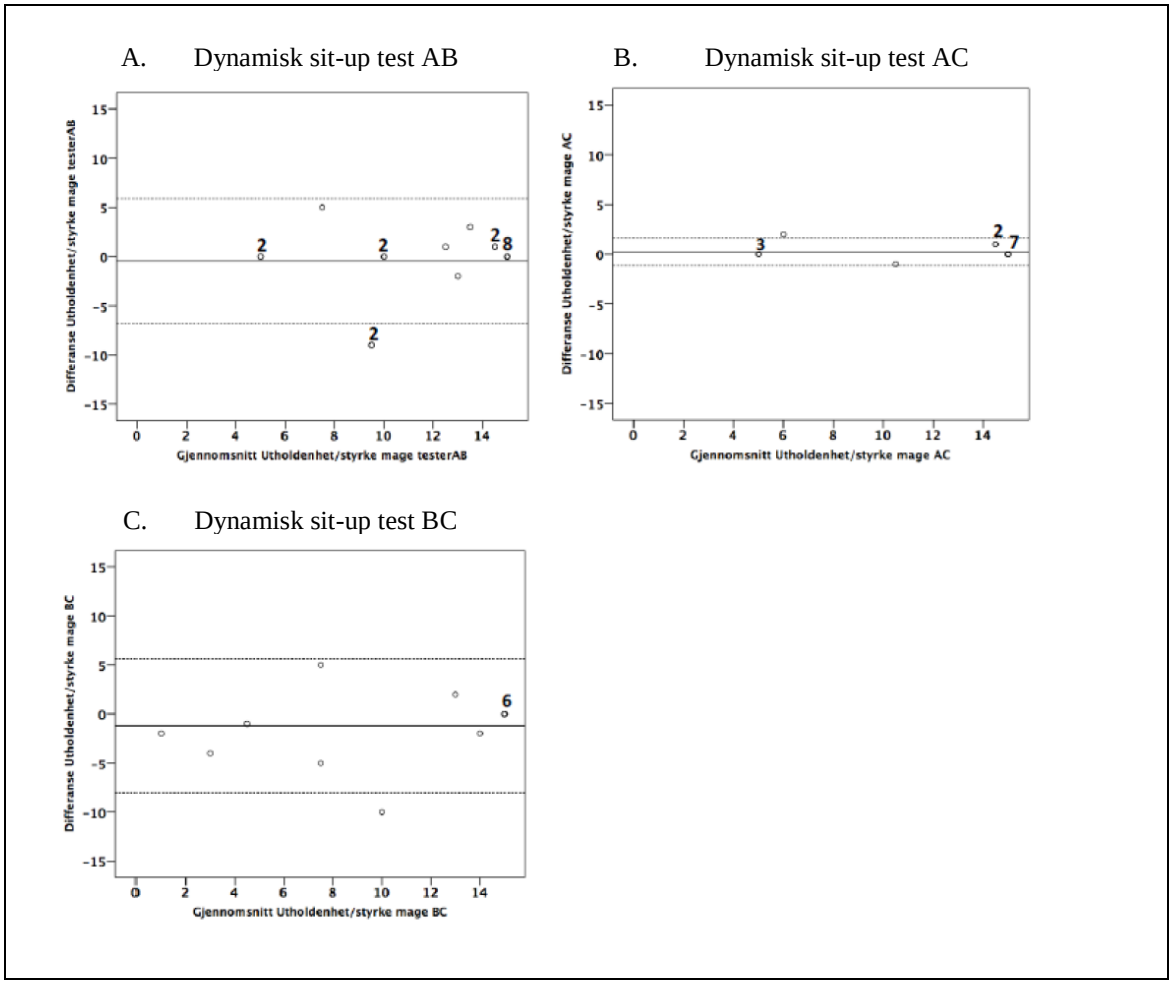
Figur 3. *Bland-Altman plot for 6 fysiske tester n= 48. Testere ABC. Gjennomsnittsskår for test 1 og test 2 vises på x-aksen og differanse på gjennomsnittskår mellom test 1 og test 2 vises på y-aksen. Hel linje beskriver gjennomsnittsforskjell. Stiplet linje over og*

under beskriver grense for enighet. Hvert datapunkt er en deltager, tall over et datapunkt beskriver flere deltagere.

5.2.4 Intertester reliabilitet av Biering-Sørensen test

Relativ intertester reliabilitet (ABC) for 48 deltagere viste høy korrelasjon ($ICC_{2.1}$ 0.80). (tabell 3 og figur 2) og moderat grad av målefeil S_w 21.67 (9.02% av måleskalen). Forskjell mellom $ICC_{2.1}$ og Spearman's rho viste forskjell på 0.08. Bland-Altman plot for intertester reliabilitet (ABC) for 48 deltagere er vist i figur 3. Gjennomsnittlig forskjell i intertester enighet var på 7.06 og målefeil oppgitt i "grense for enighet" = 66.11/-51.99. Plottet viste 3 "outliere" og totalt 93.75 % var innenfor "grense for enighet".

Parvis intertester reliabilitet (AB – AC – BC) viste høy korrelasjon ($ICC_{2.1}$ 0.74 til 0.85) og moderat grad av målefeil S_w 19.17 til 23.49. Resultatene er presentert i tabell 4. Forskjell mellom $ICC_{2.1}$ og Spearman's rho viste forskjell på 0.12 for AB og 0.07 for AC.



Figur 4. *Bland-Altman plot dynamisk sit-up test AB (n=20), AC (n=14) og BC (n=14). Gjennomsnittsskår for test 1 og test 2 vises på x-aksen og differanse på gjennomsnittskår mellom test 1 og test 2 vises på y-aksen. Hel linje beskriver gjennomsnittsforskjell. Stiplet linje over og under beskriver grense for enighet. Hvert datapunkt er en deltager, tall over et datapunkt beskriver flere deltagere.*

5.2.5 Intertester reliabilitet av dynamisk sit-up test

Relativ intertester reliabilitet (ABC) for 48 deltagere viste høy korrelasjon ($ICC_{2.1}$ 0.80) (tabell 3 og figur 2). Variablene var ikke normalfordelt, men det ble funnet en svært liten forskjell mellom ICC og Spearman's rho (0.01).

Absolutt intertester reliabilitet (ABC) for 48 deltagere viste stor grad av målefeil med S_w 2.01 (13.4% av måleskalaen). Bland-Altman plot for 48 deltagere er vist i figur 3.

Gjennomsnittlig forskjell i intertester enighet var på -0.48. Målefeil uttrykt ved grense for enighet viste 5.09/-6.05. Bland-Altman plot viste 3 ”outliere”.

Parvis intertester reliabilitet (AB – AC – BC) viste moderat til svært høy korrelasjon. Tester AB viste moderat korrelasjon $ICC_{2.1}$ 0.65 med stor grad av målefeil S_w 2.25 (15% av måleskalaen). BC viste høy korrelasjon $ICC_{2.1}$ 0.78 med stor grad av målefeil S_w 2.25 (16.87% av måleskalaen) og tester AC viste svært høy korrelasjon $ICC_{2.1}$ 0.99 med liten grad av målefeil S_w 0.5 (3.33% av måleskalaen). Resultatene er presentert i tabell 4. Bland-Altman plot for parvis intertester reliabilitet er vist i figur 4 og viser forskjell i gjennomsnittlig forskjell og ”grense for enighet”.

5.2.6 Intertester reliabilitet av ACR-18

Relativ intertester reliabilitet (ABC) for 48 deltagere viste høy korrelasjon ($ICC_{2.1}$ 0.82) og stor grad av målefeil S_w 2.04 (11.33% av måleskalaen). Resultatene er presentert i tabell 3 og figur 2. ”Grense for enighet” viste 5.08/-6.16 og gjennomsnittlig forskjell i intertester enighet var på -0.54 (figur 3). Plottet viste en ”outlier” som lå utenfor ”grense for enighet”. Totalt 97.92 % var innenfor ”grense for enighet”. En del av deltagerne hadde en høyere skår ved andre enn første test.

Parvis intertester reliabilitet (AB – AC – BC) er undersøkt og viste moderat til svært høy korrelasjon ($ICC_{2.1}$ 0.68 til 0.93). AB viste moderat grad av målefeil S_w 1.29 (7.17% av måleskalaen), AC viste stor grad av målefeil S_w 2.67 (14.83% av måleskalaen) og BC viste stor grad av målefeil S_w 2.19 (12.17% av måleskalaen). Resultatene er presentert i tabell 4.

6 DISKUSJON

6.1 Diskusjon av resultater

En viktig egenskap til alle måleinstrument som blir brukt i forskning og i klinisk arbeid er at de er reliable (de Vet et al., 2011, s. 96). Reliabilitet på et måleinstrument må alltid sees i sammenheng med kontekst, hvilken gruppe som studeres og hensikt (de Vet et al., 2011, s. 301). Formålet med denne studien var å undersøke intertester reliabilitet på seks fysiske tester som inngår i funksjonsundersøkelsen i FAktA prosjektet. Dette er første reliabilitetsstudien som tester alle disse seks testene gjennomført som et testbatteri.

Resultatene viste at:

- Intertester reliabilitet ble funnet til å være høy til svært høy for alle de seks testene utført på 48 deltagere av tre forskjellige testere A, B og C ($ICC_{2.1}$ 0.80 til 0.94). GBE-Fleksibilitet, BPS, Biering-Sørensen test og høy løftetest viste moderat grad av målefeil, mens dynamisk sit-up test og ACR-18 viste stor grad av målefeil.
- Parvis intertester reliabilitet AB, AC og BC viste moderat til svært høy reliabilitet for de seks testene ($ICC_{2.1}$ 0.65-0.99) og testene viste liten til stor grad av målefeil.

BPS viste høyest korrelasjon av de 6 testene med $ICC_{2.1}$ 0.94, og dynamisk sit-up test samt Biering-Sørensen test viste lavest korrelasjon av de 6 testene med $ICC_{2.1}$ 0.80 for intertester reliabilitet undersøkt mellom 3 testere (ABC). For parvis intertester reliabilitet viste dynamisk sit-up test (AB) lavest $ICC_{2.1}$ 0.65 og dynamisk sit-up test (AC) oppnådde høyest $ICC_{2.1}$ 0.99. Testene viste liten til stor grad av målefeil, hvor dynamisk sit-up test (AC) viste liten grad av målefeil med S_w 0.5, mens dynamisk sit-up test (BC) hadde stor grad av målefeil med S_w 2.53.

6.1.1 Intertester reliabilitet av GBE-Fleksibilitet

Relativ reliabilitet viste høy korrelasjon for intertester reliabilitet utført av tre testere (ABC) på 48 deltagere ($ICC_{2.1}$ 0.88) og moderat grad av målefeil (S_w 2.71). Det var bedre gjennomsnittskår for andre enn første testseanse. Dette kan også visuelt sees på Bland-Altman plot ved gjennomsnittlig forskjell (d) som er regnet ut til å være 1.06.

de Vet et al. (2011, s. 114) beskriver at gjennomsnittlig forskjell er et utrykk for systematisk feil, og grense for enighet kan betegnes som tilfeldig feil. Det kan diskuteres om den systematiske feilen er på grunn av læringseffekt eller at deltagerne har fått ”warm up” effekt til andre testseanse. I Bland-Altman plottet vises det to ”outliere”. Den ene har testet dårligere på andre testseanse, mens den andre har testet bedre. For deltageren som har testet bedre, er det notert på testskjemaet at deltageren kjente seg mye ledigere ved andregangs testing. For denne deltageren er det da snakk om en ”warm up” effekt, og det er mulig at dette også gjelder for andre i gruppen som har skåret bedre ved andre testseanse. En annen faktor som kan påvirke systematiske feil, er at en av testerne jevnt over tester dårligere eller bedre enn den andre. Siden det er 3 testere og alle har byttet på å teste først og sist, er det grunn til å tro at gjennomsnittsforskjellen ikke har blitt påvirket av testerne i noen stor grad.

Høy korrelasjon for intertester reliabilitet er i tråd med en tidligere studie utført på domenet ”bevegelse” i GFM-52 utført av Kvale et al. (2003). Den oppnådde intertester reliabilitet mellom tre testere (ABC) på sumskår (16 tester) ICC $_{2.1}$ 0.89. Seks av testene fra GFM-52 hører til under GBE-fleksibilitet i FAktA-prosjektet. Det blir dermed vanskelig å kunne sammenligne studiene, siden Kvale et al. (2003) har testet reliabilitet på mange flere deltester samtidig. I tillegg er studien til Kvale et al. (2003) utført på 10 friske personer og 9 pasienter sykemeldt med langvarige muskel og skjelettplager. Dette fører til at de har testet reliabilitet på en annen populasjon og Rankin and Stokes (1998) har påpekt at man ikke kan sammenligne resultatene i reliabilitetsstudier, uten at man har forholdsvis identisk utvalg.

Målefeil er i vår studie gitt ved S_w og Bland-Altman plot. S_w for intertester reliabilitet mellom tre testere var S_w 2.71. I Kvale et al. (2003) sin studie var S_w 1.6 for modifisert utgave (16 tester). Det ble brukt en annen måleskala enn i foreliggende studie, hvor maks skår for hver test var 2.3 og total skår for 16 tester var 36.8 i studien til Kvale et al. (2003). Siden S_w verdi blir oppgitt i testens måleskala, blir det vanskeligere å sammenligne skåren for disse to studiene. En måte man kan sammenligne er å regne ut S_w i prosent av måleskalaen. Foreliggende studie viste moderat grad av målefeil med S_w

2.71 som utgjør 6.45% av måleskalaen, mens studien til Kvale et al. (2003) viste S_w 1.6 som utgjør 4.35% av måleskalaen og tilsvarer liten grad av målefeil.

6.1.2 Intertester reliabilitet av BPS

Intertester reliabilitet i vår studie viste litt lavere korrelasjon $ICC_{2.1}$ 0.94 enn i studien til Magnussen et al. (2004), som viste $ICC_{2.1}$ 0.996. I studien til Magnussen et al. (2004) var det lavere målefeil for intertester reliabilitet S_w 0.25, enn i vår studie S_w 0.88.

Det er grunn til å tro at intertester reliabiliteten i vår studie var noe lavere enn i Magnussen et al. (2004) sin intertester studie fordi testingen i vår studie ble gjort på to forskjellige tidspunkt med ca. 30 minutters pause og en kan da forvente større variasjon hos deltagerne enn om det bare var en testseanse (Portney & Watkins, 2009, s. 102). ICC verdien for intertester reliabilitet i vår studie er mer i samsvar med resultatene fra test-retest i studien til Magnussen et al. (2004), som viste $ICC_{2.1}$ 0.91 og S_w 1.3. I studien til Magnussen et al. (2004) rapporterte 61% av pasienten at tilstanden hadde forandret seg mellom test og retest, mens 39 var uforandret. Det er viktig at reliabilitets testing foregår på et utvalg som det ikke har skjedd en endring hos (de Vet et al., 2011, s. 125). En viss variabilitet i stivhet og smerter fra dag til dag er ansett som normalt hos pasienter med muskel- og skjelettplager, men det kan være en viktig kilde til variasjon i testskåren om deltagerne har forandret seg. I studier hvor måleegenskaper blir undersøkt er det viktig å vite hvor mye variasjon i målinger som er på grunn av denne type variabilitet i tilstanden. For å indikere en forverring eller forbedring av tilstanden, må variabiliteten i en skår overstige slik naturlig forekommende variabilitet (Magnussen et al., 2004). I henhold til studien til Magnussen et al. (2004), bør en forandring i BPS skår hos en pasient endres ± 3.6 poeng for å kunne hevde at den fysiske ytelsen faktisk har endret seg. I denne intertester reliabilitetsstudien utført av 3 testere (ABC) på 48 deltagere, var SDC 2.47, og en vil da forvente at en deltager må skåre en positiv eller negativ endring på over 3 poeng for å vite at det har skjedd en endring som er utover målefeilen. Det var lavere målefeil på BPS undersøkt i intertester reliabilitetsstudie i FAktA-prosjektet sammenliknet med test-retest reliabilitet undersøkt på BPS av Magnussen et al. (2004). En grunn kan være at deltagerne i studien til Magnussen et al.

(2004) hadde større variasjon på test-retest siden testingen ble gjennomført med 2-3 dagers mellomrom.

I en annen studie utført av Strand et al. (2011) ble test-retest reliabilitet av BPS undersøkt og viste en $ICC_{2.1}$ på 0.89 og moderat grad av målefeil som viste SEM 1. ICC verdien for test-retest er noe lavere enn i vår studie og målefeil litt høyere. Studien er gjort på et lite utvalg (n=9) av utvalget i en større studie hvor de har undersøkt ”responsiveness”. Lite utvalg og forskjellig design gjør at det blir vanskelig å sammenligne resultatene. Utvalget i studien til Strand et al. (2011) er vesentlig mindre enn det som er anbefalt av de Vet et al. (2011) med 50 deltagere, slik at man kan være mer kritisk til resultatene i den studien enn i foreliggende studie hvor det var 48 deltagere.

For variablene som ikke var normalfordelt var differansen mellom $ICC_{2.1}$ og Spearman rho ≤0.02 og viser på lik linje med Bland-Altman plot at systematisk drift var liten, visualisert gjennom gjennomsnitt av forskjellen. Gjennomsnitt av forskjellen var på ≥ 0.50 poeng for ABC, AB, AC og BC. Da det var mange i vår studie som hadde skåret toppskår er det sannsynlig at det har påvirket normalfordelingen. Ved å ha enda større spredning i skår hos utvalget, for eksempel med et mer heterogent utvalg, kan man oppnå en enda bedre reliabilitet (Carter et al., 2011, s. 245). Siden over 15 % har skåret 0 på både første og andre test, er det snakk om en gulveffekt (Terwee et al., 2007). Det kan påvirke reliabiliteten og man kan si at den øvre delen av skalaen er dårligere undersøkt for reliabilitet.

6.1.3 Intertester reliabilitet av høy løftetest

Høy løftetest viste høy korrelasjon for intertester reliabilitet undersøkt på 48 deltagere av testere ABC ($ICC_{2.1}$ 0.88) og moderat grad av målefeil S_w 1.84. Det var flere deltagere som skåret høyere ved andre enn første testing og gjennomsnittsforskjell i intertester reliabilitet regnet ut ved paret t-test viste -1.52 poeng. Denne verdien viser at det er systematisk drift i datamaterialet. Det kan ha en sammenheng med at deltakerne har hatt en læringseffekt, eller en ”warm up” effekt. Bland-Altman plot viser tre ”outliere”, hvor to av ”outlierne” tester bedre og en tester dårligere. Deltageren som

testet dårligere hadde nakke og skulder problem. Det kan tenkes at høy løftetest kan være mer krevende hvis man har et spesifikt skulderproblem og at smerter kan ha påvirket gjennomføringen. Det er vanskelig å spekulere i grunnen, da det ikke er notert noe spesielt for dette.

I studien til Magnussen et al. (2004) har de undersøkt løftetest med Kappa som er en annen korrelasjonskoeffisient brukt ved ordinale variabler. Flere har beskrevet at resultatene er avhengig av hvilken korrelasjonskoeffisient som brukes (Carter et al., 2011). For intertester reliabilitet oppnådde de Kappa på 1.0 som viser perfekt reliabilitet. For Kappa er det ikke mulig å oppgi målefeil, og gjør at man må være mer kritisk til denne metoden enn ICC som er brukt i vår studie. Høy løftetest i FAktA-prosjektet oppnådde omtrent samme ICC verdi som løftetest i studien til Strand et al. (2011), men i vår studie viste det lavere målefeil S_w 1.84 og lavere SDC 5.1. Siden både utvalg, design og hvordan man gjennomførte testen var forskjellig kan dette ha påvirket resultatene.

6.1.4 Intertester reliabilitet av Biering-Sørensen test

I studien til Keller et al. (2001) viste ICC 0.93 for gruppen med korsryggsmerter og var høyere enn i foreliggende studie som viste $ICC_{2.1}$ 0.80. For friske deltagere oppnådde man samme verdi ICC 0.80 i Keller et al. (2001) sin studie og foreliggende studie. Reliabilitet er undersøkt med en tester i studien til Keller et al. (2001) og Streiner, Cairney, and Norman (2015) har påpekt at det er lettere for å få høyere reliabilitet ved intratester reliabilitet enn intertester reliabilitet. ”Grense for enighet” viser litt lavere verdier i studien til Keller et al. (2001) for deltagere med korsryggsmerter enn i foreliggende studie, men for friske deltagere er ”grense for enighet” noe høyere enn i vår studie. Det er ikke oppgitt hvilken ICC verdi som er brukt i studien til Keller et al. (2001) og det er flere faktorer som gjør at studiene er forskjellig designet, blant annet tid mellom testing og utvalg, slik at dette har påvirket resultatene.

I studien til Latimer et al. (1999) viste $ICC_{1.1}$ 0.85 og SEM 15.6 sekunder for et utvalg bestående av 23 deltakere med korsryggsmerter, 20 deltakere som tidligere har hatt en episode med korsryggsmerter og 20 deltagere som var asymptomatisk. I foreliggende

studie var $ICC_{2.1}$ 0.80 som er noe lavere, samt at målefeil S_w 21.67 sekunder var noe høyere i foreliggende studie i forhold til Latimer et al. (1999) sin studie. Det er flere faktorer som gjør at studiene er vanskelig å sammenligne på grunn av forskjellig utvalg og design. I studien til Latimer et al. (1999) rapporterte de at en deltager ble ekskludert fra gruppen på grunn av mer en 9.5 SD større en gjennomsnittet. Analyser gjort med denne deltakeren inkludert viste en $ICC_{1.1}$ 0.82 og SEM 36.8 sekunder. Det er omdiskutert i litteraturen om hvorvidt man kan ekskludere en "outlier" (de Vet et al., 2011, s. 103). I vår studie er det 3 "outliere" som har påvirket reliabiliteten. En deltager skåret 232 sekunder på første testing, og 122 på andre testing. Pallant (2013) har påpekt at det er viktig å gå tilbake i datamaterialet å sjekke at en "outlier" er reell. I dette tilfellet er "outlieren" sjekket, og skåren var reell. Det er usikkert hvorfor pasienten skåret så mye lavere på andre testing. En faktor kan være fysisk trøtthet. Det er ikke gjort analyser uten "outlierne" i foreliggende studie, men en antar at en da ville fått høyere ICC verdi og lavere målefeil.

6.1.5 Intertester reliabilitet av dynamisk sit-up test

Relativ intertester reliabilitet (ABC) for 48 deltagere viste høy korrelasjon ($ICC_{2.1}$ 0.80) og stor grad av målefeil S_w 2.01. Variablene var ikke normalfordelt, men det ble funnet svært liten forskjell mellom ICC og Spearman (0.01). I Bland-Altman plottet kan man se at den gjennomsnittlige forskjellen i intertester enighet var liten mellom første og andre testing (-0.48), men 21 deltagere skåret toppskår og det antyder en klar takeffekt (Terwee et al., 2007). Siden så mange deltagere har skåret i øvre del av skalaen vil dette ha påvirket reliabiliteten i tillegg til at man kan si at reliabiliteten for nedre del av skalaen er mindre undersøkt.

For parvis intertester reliabilitet så man tydelige forskjeller mellom de ulike testparene i ICC verdi og målefeil uttrykt i S_w (tabell 4) samt Bland-Altman plot (figur 4). Testpar AC har høyest reliabilitet og minst målefeil. I Bland-Altman plottet ser man at deltagerne ligger nærme 0-linjen som antyder at det var svært liten gjennomsnittsforskjell mellom første og andre testing, og at testerne hadde stor grad av enighet. Testpar AB hadde lavest ICC-verdi, men noe lavere målefeil enn BC. Det illustrerer at en lavere ICC-verdi kan gi mindre målefeil enn en variabel som viser

høyere ICC-verdi. ICC er avhengig av spredning i skår, og siden BC hadde større spredning i skår enn AB, kan dette være en årsak som resulterte i en høyere ICC verdi for BC (de Vet et al., 2011, s. 101-102). Både AB og BC hadde begge ”outliere”. I et lite utvalg kan ”outliere” påvirke reliabilitet i stor grad (Carter et al., 2011, s. 266). For deltagerne som ligger utenfor ”grense for enighet” kan man si at det har skjedd en endring som er over det man kan forvente som målefeil. Det er notert i testskjemaet at deltageren som ligger utenfor ”grense for enighet” i Bland-Altman plottet til BC, plutselig klarte å bryte en barriere på mageøvelsen. Dette førte til at skåren økte fra 5 repetisjoner på første testing til 15 repetisjoner på andre testing. På den måten vet man at det har skjedd en endring hos deltageren og at målefeilen ikke ligger hos testerne som har skåret forskjellig.

6.1.6 Intertester reliabilitet av ACR-18

Intertester reliabilitet for 48 deltagere viste høy korrelasjon $ICC_{2.1}$ 0.82 og stor grad av målefeil S_w 2.04. En studie av Weiner et al. (2006) viste intertester reliabilitet (0.84) uttrykt med Pearson’s korrelasjonskoeffisient på 30 eldre deltagere med kroniske korsryggsmerter. Man kan være mer kritisk til resultatene i studien til Weiner et al. (2006) da Pearson’s r er regnet som en mindre kritisk korrelasjonskoeffisient enn ICC fordi den ikke tar hensyn til systematiske feil (de Vet et al., 2011, s. 110). I tillegg er det 30 deltagere i studien til Weiner et al. (2006), noe som er vesentlig mindre enn de Vet et al. (2011) sine anbefalinger på 50 deltagere .

6.2 Diskusjon av metode

Intern validitet kan beskrives som i hvilken grad resultatene er gyldige for det utvalget og det fenomenet som er undersøkt uten at ukontrollerte, utenforliggende faktorer er ansvarlig for resultatene (Polit & Beck, 2012). Når man designer en forskningsstudie, bør man nøye vurdere hver faktor som kan være en trussel for intern validitet i henhold til det designet som er skissert. Man kan enten prøve å kontrollere faktorene for å minimere trusselen for intern validitet, eller akseptere trusselen som en uunngåelig feil for designet. Det er ingen forskningsdesign som er helt perfekt, og intern validitet kan i enkelte tilfeller bli en kompromiss for ekstern valid (Carter et al., 2011, s. 76). I

foreliggende studie er det flere faktorer som kan påvirke den interne validiteten, blant annet design og utvalg.

6.2.1 Design

Tverrsnittstudier egner seg godt til å beskrive status på et fenomen eller beskrive forholdet mellom fenomen på et gitt tidspunkt (Polit & Beck, 2012, s. 184). I denne intertester reliabilitetsstudien er to testseanser utført av to forskjellige individuelle testere med ca. 30 minutters pause imellom. Det var ikke mulig å bruke video eller at testerne testet samtidig i foreliggende studie, siden en del av testene innebærer at hver tester individuelt må vurdere egenskaper som fleksibilitet, bevegelighet og triggerpunkter med bruk av egne hender for å kunne skåre. Målefeil i denne studien kan skyldes varians hos deltager, tester og/eller testen som er brukt (Moe-Nilssen et al., 2008). Når testerne kan teste samtidig eller vurdere samme video av testene, kan man sikre at variabiliteten ikke skyldes variabilitet hos deltagerne eller testerne sin utførelse (Portney & Watkins, 2009, s. 101).

Det er ikke noe standard kriterier med hensyn til tidsintervall mellom testseansene. Det hevdes at det må være en viss tid mellom dem for å minimere påvirkning av læringseffekten, men på den andre siden kan for lang tid forandre deltageren sin tilstand (de Vet et al., 2011, s. 125). I foreliggende studie var det relativt kort tid mellom testseansene, slik at dette kan ha bidratt til læringseffekt, fysisk trøtthet, eller at deltagerne fikk en ”warm up” effekt og ble mer bevegelig fra første til andre testing. Alle disse faktorene kan ha påvirket reliabiliteten (Portney & Watkins, 2009, s. 109).

I studien til Latimer et al. (1999) mente en ekspert at det var nok med 15 min mellom hver testing av Biering-Sørensen test for å fysisk kunne hente seg inn igjen. I denne studien var det et helt testbatteri som ble undersøkt, og det er naturlig å tenke seg at deltagerne trengte noe lenger tid til å fysisk hente seg inn igjen. Hadde det gått enda lengre tid mellom testseansene kunne man, imidlertid ha risikert at deltagerne ikke ville delta i studien fordi testingen samlet sett ville tatt for lang tid. Det ansees heller ikke som etisk riktig å la deltagerne bruke mer tid enn det som strengt tatt var nødvendig.

De seks fysiske testene er satt sammen som et testbatteri i denne studien og er valgt på bakgrunn av at de er standardiserte og at de er egnet til bruk i klinikken. En del av de tidligere reliabilitetsstudiene som er utført på testene er bare testet på en av testene om gangen, som i for eksempel studien til Magnussen et al. (2004) som kun undersøkte BPS. I studien til Strand et al. (2011) er det på den andre siden testet en rekke tester samtidig, blant annet BPS, Løftetest og Biering-Sørensen test. Man kan forvente at det er fysiske mer krevende å gjøre flere tester enn bare en test. Det kan være en faktor som har påvirket reliabiliteten i foreliggende studie.

Deltagerne som gikk opp mer enn et nivå eller mer i NPRS ble ekskludert for at ikke økt smerte skulle påvirke resultatet (Farrar, Young, LaMoreaux, Werth, & Poole, 2001). I andre studier har de spurt pasientene om de føler en endring siden forrige testing, som i studiene til Magnussen et al. (2004) og Strand et al. (2011). Når man tester reliabilitet er det generelt viktig at det ikke har skjedd en endring (de Vet et al., 2011, s. 125).

Det er flere forhold som mest sannsynlig kunne forbedret reliabiliteten i studiet. Hvis man tar flere mål og tar gjennomsnitt av målingene, vil reliabiliteten øke og målefeil bli mindre (de Vet et al., 2011, s. 244). I denne studien var det ikke aktuelt, da testene til vanlig gjennomføres kun en gang, og at gjennomsnittet av flere testresultat ville kunne gi kunstig høy reliabilitet. Hvis man skal forbedre reliabiliteten til et måleinstrument kan det på en annen side være aktuelt å gjøre dette. Det kan være hensiktsmessig å gjennomføre testene en gang før man tester reliabilitet, slik at deltageren er kjent med testene og at man eliminerer læringseffekt eller ”warm up” effekt (Portney & Watkins, 2009, s. 109). Det var heller ikke aktuelt å gjøre i denne studien, da det ville blitt uhensiktsmessig stor belastning på deltagerne om de skulle gjennomført testene en gang til. Med video-opptak ville det eventuelt vært lettere å skåret flere ganger samme hendelse.

Intratester reliabilitet gir ofte en høyere korrelasjonskoeffisient enn intertester reliabilitet (Streiner et al., 2015). Det er en styrke i studien at det var 3 forskjellige testere. En annen styrke med studien er at testerne ikke hadde mulighet for å se hva den andre testeren hadde skåret, og var dermed blindet.

6.2.2 Testprosedyre

Datainnsamling ble gjennomført i henhold til en standardisert testprotokoll. Testprosedyren som var utviklet for de fysiske testene i FAktA prosjektet ble brukt, i tillegg til at testene ble utført i samme rom, med samme utstyr og med tilnærmet like forhold. Instruksjon og måten testerne instruerte deltagerne på kan ha variert fra tester til tester, selv om de hadde øvd sammen. Carter et al. (2011) diskuterer at grad av standardisering må speile hva resultatene skal brukes til, for eksempel i klinisk sammenheng eller i forskning. I denne studien var det av interesse å teste intertester reliabilitet mellom tre testere i FAktA-prosjektet. Testerne hadde en treningsperiode hvor de gikk gjennom prosedyre og skåringskriterier før intertester reliabilitetsstudien. De hadde også en samling underveis i intertester reliabilitetsstudiet hvor de gikk gjennom testene på nytt for å kalibrere seg. Testerne fortalte at det kunne vært nyttig å ha samlingen hvor de kalibrerte seg litt tidligere. De påpekte i tillegg at det var vanskelig å vite hvor punktene for triggerpunktene skulle testes, selv om det var tegnet inn på en figur med beskrivelse i testprosedyren. Til tross for dette har triggerpunkttesten oppnådd høy reliabilitet, men med stor grad av målefeil. Carter et al. (2011) beskriver form for ekstremt høy standardisering hvor f.eks. punkter blir tegnet inn, slik at testerne har samme utgangspunkt og at det eliminerer sjanse for målefeil. Det ansees som ikke hensiktsmessig i dette tilfellet da, standardisering ut over hvordan testprosedyren utføres generelt i FAktA-prosjektet kunne gitt kunstig høy reliabilitet. Hvis testbatteriet skal brukes i klinikken senere av andre testere ansees det heller ikke som hensiktsmessig med ekstremt høy standardisering, da det vil ta for mye tid og ressurser.

Det ble ikke utført pilottesting før intertester reliabilitetsstudien. Portney and Watkins (2009, s. 108) anbefaler å gjennomføre pilottesting i forkant av studien for å finne måter å eventuelt forbedre reliabiliteten. Systematisk feil er blant annet feil som det er mulig å rette opp i. Hvis det viser seg at den ene testeren systematisk tester bedre enn den andre, at det er for kort tid mellom testseansen slik at deltakeren får læringseffekt, eller at deltakerne tester dårligere på andre testseanse på grunn av fysisk trøtthet er dette faktorer som kan justeres. I tillegg kan pilottest avdekke en del praktiske utfordringer.

Dette kan sees som en svakhet at ikke ble gjort, men en valgte istedenfor å øve på testene i forkant av studien.

Testerne hadde et oppsett på hvem som skulle være tester 1 og hvem som skulle være tester 2 for å unngå systematiske feil. Av praktiske grunner lot det seg ikke gjøre for alle deltagerne. Det er ikke noe trend på at det er store forskjeller mellom testerne, så mest sannsynlig har ikke dette hatt noe stor påvirkning.

6.2.3 Utvalg

Det ble rekruttert 52 deltaker til reliabilitetsstudien fra FAktA-prosjektet etter anbefalinger av de Vet et al. (2011), om å ha rundt 50 deltagere. Fire deltakere ble ekskludert og 48 deltakere ble inkludert i studien. Det å ha 48 deltakere regnes som en styrke for studien.

Utvalget i intertester reliabilitetsstudien består av arbeidstakere med muskel- og skjelettplager. Man ser at utvalget i foreliggende studie har noe forskjellig sammensetning i demografi sammenlignet med utvalget i studien til Ask et al. (2014), hvor 250 deltagere fra FAktA- prosjektet ble inkludert. I studien til Ask et al. (2014) var det en større andel som var sykemeldt og en større andel kvinner enn i foreliggende studie. Det var flere som var sykemeldt i studien til Ask et al. (2014), og korsryggsmerter var rapportert som hovedplage til den største andelen av deltagerne i motsetning til foreliggende studie hvor største andelen hadde utbredte muskel- og skjelettplager. Studien til Ask et al. (2014) representerer første fase og FAktA-prosjektet fortsatte å inkludere deltagere etter den studien ble gjennomført. Siden utvalget påvirker reliabiliteten er det nødvendig å vite bakgrunnsvariabler for det aktuelle utvalget. Hvis intertester reliabilitet hadde blitt undersøkt på et annet utvalg i FAktA-prosjektet kunne resultatene blitt annerledes.

Andre reliabilitetsstudier som har testet reliabilitet på de fysiske testene har variert i demografi med tanke på jobbstatus, hovedplage, osv. Rankin and Stokes (1998) har hevdet at det å sammenligne reliabilitet resultater mellom forskjellige studier ikke er

mulig foruten at utvalget i hver studie er helt identisk. Dermed er det viktig å ta hensyn til utvalget når man skal sammenligne resultatene i denne studien med tidligere studier.

Reliabilitetsstudier er avhengig av en heterogen gruppe for å oppnå god reliabilitet slik at hele måleskalaen benyttes (Streiner et al., 2015). Studier hvor de har inkludert både friske og syke har større mulighet for å få bedre reliabilitet på grunn av at det som regel blir større variasjon i skår. Streiner et al. (2015) påpeker at det ikke er legitimert å inkludere friske personer for å oppnå bedre reliabilitet, hvis hensikten er å bare bruke måleinstrumentet på en populasjon med pasienter i etterkant. Det ansees som en styrke i foreliggende studie at det kun er inkludert deltagere med muskel- og skjelettplager og at det ble oppnådd høy til svært høy reliabilitet.

6.2.4 Analyse

Kottner et al, 2011 har påpekt at det ofte er manglende rapportering om hvilken statistikk som er brukt i reliabilitetsstudier. Siden forskjellige type statistikk kan gi ulike resultat for samme data, er det svært viktig at dette blir opplyst om (Kottner et al., 2011). I vår studie har vi brukt $ICC_{2.1}$ med ”absolute agreement” som inkluderer systematiske feil med ”single measure”. Det er en styrke at relativ reliabilitet er undersøkt med $ICC_{2.1}$ i den foreliggende studien, framfor for eksempel Pearson’s r som er mindre kritisk. (de Vet et al., 2011, s. 110). Som regel vil en ”average measure” i SPSS ha en høyere verdi enn en ”single measure” siden ”average measure” i SPSS tar høyde for at det blir tatt flere målinger (de Vet et al., 2011, s. 108). I FAktA-prosjektet blir kun en måling utført og ”average measure” kunne dermed ha gitt kunstig høy reliabilitet.

Det ansees som en styrke i studien at absolutt reliabilitet er uttrykt i S_w og at Bland-Altman plot er kalkulert. I tillegg er $SDC_{95\%}$, som er nært relatert til målefeil oppgitt i foreliggende studie. Når man undersøker absolutt reliabilitet kan det være nyttig å få et inntrykk av hvor stor den systematiske feilen og tilfeldige feilen er. Mens S_w er anvendt i studien til å si noe om hvor stor en endring må være for at man skal være sikker på at endringen er høyere enn målefeilen, kan man visuelt se i Bland-Altman plottet hvor stor den systematiske og tilfeldige feilen er (de Vet et al., 2011, s. 114). Rankin and Stokes

(1998) påpekte i sin studie at verken ICC eller Bland-Altman plot gir nok informasjon alene og at det er anbefalt at begge metodene brukes. I foreliggende studie er både ICC og Bland-Altman plot undersøkt som er en styrke for studien.

Det er varierende forslag i litteraturen i henhold til grad av reliabilitet, og hva som for eksempel indikerer ”dårlig”, ”middels” eller ”perfekt” reliabilitet (Kottner et al., 2011). I foreliggende studie er Munro’s inndeling for styrke på korrelasjons koeffisienter anvendt (Carter et al., 2011, s. 318). Det er viktig å ta hensyn til at resultatene kunne blitt annerledes med en annen inndeling. Portney and Watkins (2009, s. 97) har påpekt at det er forskeren sin forpliktelse å rettferdiggjøre hvilket reliabilitetsnivå som er akseptabelt basert på måleinstrumentet som blir testet. Det ansees som en relevant inndeling i henhold til grad av reliabilitet i foreliggende studie med hensyn til tidligere studier som er gjort.

Parametriske metoder er generelt fortrukket innenfor statistikk, siden de gir en generelt mer kompleks analyse, enn ikke-parametrisk statistikk (Altman & Bland, 2009). Parametrisk statistikk lager beregninger med bakgrunn for at de antar at datamaterialet er normalfordelt. Valg av parametrisk og ikke parametrisk statistikk hevdes å være relatert til størrelse på utvalget, da normalfordeling er viktigere for mindre utvalg (Altman & Bland, 2009). Kolmogorv-smirnof test ble utført i intertester reliabilitetsstudien for å undersøke normalfordeling. Pallant (2013, s. 214) har beskrevet at med et stort nok utvalg (30+) vil det mest sannsynlig ikke forårsake problemer med å bruke parametrisk statistikk, selv om ikke data er normalfordelt. Resultatene fra Kolmogorov-Smirnof test viste at noen av variablene ikke var normalfordelt. Siden utvalget i intertester reliabilitetsstudien består av 48 deltagere vil det slik Pallant (2013) har hevdet mest sannsynlig ikke ha forårsaket problemer at det ble brukt parametrisk statistikk. I foreliggende studie ble det dermed valgt å bruke $ICC_{2.1}$ verdien, men $ICC_{2.1}$ verdien ble i tillegg sammenlignet med det ikke parametriske alternativet Spearman’s rho. Forskjellen mellom ICC og Spearman’s rho viste seg å være liten. Resultatene fra parvis reliabilitet i foreliggende studie er mer usikker enn resultatene for hele gruppen, siden parvis intertester reliabilitet ble undersøkt på et mindre utvalg en det som er anbefalt (de Vet et al., 2011).

6.2.5 Etiske betraktninger

For å ivareta etisk forsvarlighet er det en styrke at studien har vært gjennomført i henhold til Helsinki deklarasjonen, og at forskningsprosjektet har blitt godkjent av Regional komité for medisinsk og helsefaglig forskningsetikk (REK). Det er viktig at deltakere blir informert om forskningsprosjektet før de skriver under informert samtykke (Carter et al., 2011). I denne studien fikk deltakerne skriftlig og muntlig informasjon i forkant av studien. Det var ikke forbundet noe helserisiko med å være med i foreliggende studie, men deltagerne ble informert om at de fysiske testene i FAktA-prosjektet kan for enkelte medføre noe stølhet i etterkant. Hvis deltageren fikk økte smerter tilsvarende mer enn ett nivå på NPRS på første testseanse ble deltageren ekskludert fra studien. Det var to deltagere som ble ekskludert fra studien på grunn av økte smerter, hvor en av deltagerne oppga spesifikt at det var på grunn av økt hodepine. To deltagere ansees som en liten del av utvalget i denne studien, men det er viktig at eventuelle fysiske og psykiske påkjenninger alltid må vurderes med hensyn til etisk forsvarlighet. Deltakerne kunne når som helst trekke seg fra studien (Carter et al., 2011).

6.2.6 Forskerrollen

Som forsker er det viktig å designe et etisk forsvarlig forskningsprosjekt, i tillegg til at resultatene fra forskningsprosjektet må kunne gi ny viten og tilføre samfunnet noe. Man har alltid en form for førforståelse som forsker, og det har vært viktig for meg å ta hensyn og være bevisst på dette. Jeg hadde liten erfaring som forsker fra tidligere av og kjente lite til de fysiske testene som ble anvendt i FAktA-prosjektet. Det er viktig at man tilegner seg nok bakgrunnskunnskap for å ha mulighet til å gjennomføre studien på en tilfredsstillende måte. Jeg var med på gjennomgang av testene sammen med testerne ved en anledning for å få bedre kjennskap til gjennomføring av testene. Det kan være en svakhet at jeg ikke kjente til de fysiske testene bedre før jeg designet innhenting av datamaterialet til studien. Det var tre fysioterapeuter tilknyttet FAktA-prosjektet som gjennomførte datainnsamling. To av de tre testerne har fungert som veiledere på masteroppgaven. De har både hatt god kjennskap til prosjektet og lang erfaring med testene fra før, og har gitt mange nyttige tilbakemeldinger underveis i prosessen, både i planleggingsfasen, datainnsamling og ved analyse av datamaterialet. Den tredje testeren

har også kommet med nyttige innspill i prosessen. Min rolle som forsker har vært å designe studiet, plotte data, analysere datamaterialet og presentere resultatene i masteroppgaven.

6.2.7 Ekstern validitet

Et sentralt spørsmål med hensyn til ekstern validitet er om resultatene kan generaliseres til andre mennesker, settinger og situasjoner enn det som er presentert i den aktuelle studien (Polit & Beck, 2012). Portney and Watkins (2009) hevder at reliabilitet kun er gjeldende for den populasjonen og i den konteksten man undersøker reliabilitet, og av den grunn kan ikke resultatene i foreliggende studie umiddelbart generaliseres til en annen populasjon enn de som har liknende karakteristika til deltakerne i dette utvalget. Resultatene fra foreliggende studie kan generaliseres til en populasjon med arbeidstakere som har muskel-og skjelettplager, hvor flesteparten er kvinner og kortidssykemeldt eller står i fare for å bli sykemeldt. Arbeidstakerne var inkludert fra Helse og sosial avdelingen, barnehager og en ryggklinikk i Bergen kommune. Utvalget i foreliggende studie stemmer godt med at det er høyest forekomst av muskel og skjelettplager hos kvinner og helsearbeidere ellers i samfunnet (Eriksen et al., 2003; Ihlebaek et al., 2010). Det må understrekes at for å kunne generalisere funnene i et utvalg med like karakteristika er det i tillegg avhengig av lignende design.

I foreliggende studie var testerne tre fysioterapeuter med lang erfaring med bruk av testene. Resultatene kan med større sikkerhet overføres til andre testere, enn om det bare var en tester (Portney & Watkins, 2009, s. 102). Testerne i denne studien hadde gjennomført en treningsperiode i forkant av studien for å sikre at testerne hadde lik oppfatning av prosedyre og skåringskriterier. Hvis testbatteriet skal anvendes i primærhelsetjenesten, eventuelt bedriftshelsetjenesten vil testere der ha et annet utgangspunkt for vurderingene sine enn det testerne i denne studien hadde. Det må en viss opplæring til på testene for at det skal ha overføringsverdi til andre og det bør være en forutsetning med god opplæring og at man setter seg godt inn i testbatteriet før det tas i bruk.

6.2.8 Anbefalinger for framtidige studier:

Det anbefales i framtidige studier å undersøke hvor mye opplæring som er nødvendig for at nye testere skal oppnå en god reliabilitet ved bruk av testbatteriet. I første omgang kan det være aktuelt i primærhelsetjenesten og eventuelt i bedriftshelsetjenesten. Reliabilitet må i denne sammenheng testes i den konteksten og på den populasjonen som er relevant at testbatteriet skal brukes på. Dynamisk sit-up test og ACR-18 bør undersøkes nærmere når det gjelder målefeil.

7 KONKLUSJON

I foreliggende studie ble intertester reliabilitet mellom 3 testere undersøkt på seks fysiske tester. Populasjonen bestod av arbeidstakere som var kortidssykemeldte på grunn av muskel- og skjelettplager eller var i jobb tross plagene. Studien viste høy til svært høy intertester reliabilitet ($ICC_{2.1}$ 0.80 til 0.94) for de seks fysiske testene undersøkt på 48 deltagere. Testene viste moderat til stor grad av målefeil uttrykt i S_w. Biering-Sørensen test og dynamisk sit-up test viste lavest reliabilitet av de seks testene, hvor Biering-Sørensen test hadde moderat grad av målefeil, mens dynamisk sit-up test hadde stor grad av målefeil. BPS viste høyest reliabilitet med lavest målefeil. For parvis intertester reliabilitet viste resultatene moderat til svært høy korrelasjon $ICC_{2.1}$ 0.65 til 0.99 og liten til stor grad av målefeil. For noen av testene viste Bland-Altman plot en liten systematisk feil uttrykt i gjennomsnittsforskjell mellom første og andre test. De systematiske feilene er sannsynligvis utrykk for læringseffekt, "warm up" effekt, eller fysisk trøtthet. Resultatene i denne studien er i henhold til lignende studier som er gjort tidligere, men siden reliabilitet er avhengig av populasjon og kontekst er det vanskelig å konkludere basert på sammenligninger. Det er første gang alle disse seks testene er utført samtidig som et testbatteri, og det kan ha påvirket reliabiliteten. Basert på resultatene i denne studien var intertester reliabilitet mellom testere i FAktA-prosjektet god for alle de seks fysiske testene når de ble testet som et testbatteri, men dynamisk sit-up test og ACR-18 viste stor grad av målefeil og bør undersøkes nærmere.

8 REFERANSER

Altman, D. G., & Bland, J. M. (2009). Parametric v non-parametric methods for data analysis. *BMJ, 338*, a3167. doi: 10.1136/bmj.a3167

Andersen, I., Frydenberg, H., & Mæland, J. H. (2009). Muskel- og skjelettplager og fremtidig sykefravær. *Tidsskr Nor Legeforen, 129*(12), 1210-1213.

Ask, T., Skouen, J. S., Assmus, J., & Kvale, A. (2014). Self-Reported and Tested Function in Health Care Workers with Musculoskeletal Disorders on Full, Partial or Not on Sick Leave. *J Occup Rehabil*. doi: 10.1007/s10926-014-9557-y

Biering-Sorensen, F. (1984). Physical measurements as risk indicators for low-back trouble over a one-year period. *Spine (Phila Pa 1976), 9*(2), 106-119.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1*(8476), 307-310.

Bland, J. M., & Altman, D. G. (1996). Measurement error. *BMJ, 313*(7059), 744.

Brage, S., Ihlebaek, C., Natvig, B., & Bruusgaard, D. (2010). [Musculoskeletal disorders as causes of sick leave and disability benefits]. *Tidsskr Nor Laegeforen, 130*(23), 2369-2370. doi: 10.4045/tidsskr.10.0236

Carter, R. E., Lubinsky, J., & Domholdt, E. (2011). *Rehabilitation research: principles and applications*. St. Louis, Miss.: Elsevier Saunders.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in Medicine*. Cambridge: Cambridge University Press.

Demoulin, C., Vanderthommen, M., Duysens, C., & Crielaard, J. M. (2006). Spinal muscle evaluation using the Sorensen test: a critical appraisal of the literature. *Joint Bone Spine, 73*(1), 43-50. doi: 10.1016/j.jbspin.2004.08.002

Eriksen, W., Bruusgaard, D., & Knardahl, S. (2003). Work factors as predictors of sickness absence: a three month prospective study of nurses' aides. *Occup Environ Med, 60*(4), 271-278.

Fairbank, J. C., Couper, J., Davies, J. B., & O'Brien, J. P. (1980). The Oswestry low back pain disability questionnaire. *Physiotherapy, 66*(8), 271-273.

Farrar, J. T., Young, J. P., Jr., LaMoreaux, L., Werth, J. L., & Poole, R. M. (2001). Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain, 94*(2), 149-158.

Holtermann, A., Hansen, J. V., Burr, H., & Sogaard, K. (2010). Prognostic factors for long-term sickness absence among employees with neck-shoulder and low-back pain. *Scand J Work Environ Health, 36*(1), 34-41.

Ihlebaek, C., Brage, S., Natvig, B., & Bruusgaard, D. (2010). [Occurrence of musculoskeletal disorders in Norway]. *Tidsskr Nor Laegeforen, 130*(23), 2365-2368. doi: 10.4045/tidsskr.09.0802

Ihlebaek, C., & Lærum, E. (2004). Plager flest - koster mest - muskel-skjelettlidelser i Norge. Rapport nr. 1. Hentet fra: http://www.formi.no/images/uploads/pdf/rapport_sept_04.pdf

Jensen, M. P., Karoly, P., & Braver, S. (1986). The measurement of clinical pain intensity: a comparison of six methods. *Pain, 27*(1), 117-126.

Kamaleri, Y., Natvig, B., Ihlebaek, C. M., & Bruusgaard, D. (2008). Localized or widespread musculoskeletal pain: does it matter? *Pain, 138*(1), 41-46. doi: 10.1016/j.pain.2007.11.002

Keller, A., Hellesnes, J., & Brox, J. I. (2001). Reliability of the isokinetic trunk extensor test, Biering-Sorensen test, and Astrand bicycle test: assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. *Spine (Phila Pa 1976), 26*(7), 771-777.

Knardahl, S., Veiersted, B., Medbø, J. I., Matre, D., Jensen, J., Strøm, V., & al., e. (2008). Arbeid som årsak til muskelskjelettskader: kunnskapsstatus 2008. STAMI-rapport årg. 9. nr. 22. from http://www.arbeidstilsynet.no/binfil/download2.php?tid=103322

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hrobjartsson, A., . . . Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol, 64*(1), 96-106. doi: 10.1016/j.jclinepi.2010.03.002

Kvale, A., Bunkan, B. H., Opjordsmoen, S., & Friis, S. (2012). Development of the movement domain in the global body examination. *Physiother Theory Pract, 28*(1), 41-49. doi: 10.3109/09593985.2011.561419

Kvale, A., Ljunggren, A. E., & Johnsen, T. B. (2003). Examination of movement in patients with long-lasting musculoskeletal pain: reliability and validity. *Physiother Res Int, 8*(1), 36-52.

Latimer, J., Maher, C. G., Refshauge, K., & Colaco, I. (1999). The reliability and validity of the Biering-Sorensen test in asymptomatic subjects and subjects reporting current or previous nonspecific low back pain. *Spine (Phila Pa 1976), 24*(20), 2085-2089; discussion 2090.

Lund, T., Labriola, M., Christensen, K. B., Bultmann, U., & Villadsen, E. (2006). Physical work environment risk factors for long term sickness absence: prospective findings among a cohort of 5357 employees in Denmark. *BMJ, 332*(7539), 449-452. doi: 10.1136/bmj.38731.622975.3A

Lærum, E., Brage, S., Ihlebæk, C., Johnsen, K., Natvig, B., & Aas, E. (2013). Et muskel- og skjelettregnskap. Forekomst og kostnader knyttet til skader, sykdommer og plager i muskel- og skjelettsystemet *MST rapport 1/2013*.

Lærum, E., Brox, J. I., Storheim, K., Espeland, A., Haldorsen, E., Munch-Ellingsen, J., & al., e. (2007). Nasjonale kliniske retningslinjer. Korsryggsmerter: med og uten nerverotaffeksjon. Hentet fra: http://www.formi.no/Helsepersonell/id/kliniske_retningslinjer/

Magnussen, L., Strand, L. I., & Lygren, H. (2004). Reliability and validity of the back performance scale: observing activity limitation in patients with back pain. *Spine (Phila Pa 1976), 29*(8), 903-907.

Main, C. J., Watson, P. J., & Sullivan, M. J. L. (2008). *Pain management : practical applications of the biopsychosocial perspective in clinical and occupational settings* (2nd ed.). Edinburgh: Churchill Livingstone.

Mody, G. M., & Brooks, P. M. (2012). Improving musculoskeletal health: global issues. *Best Pract Res Clin Rheumatol, 26*(2), 237-249. doi: 10.1016/j.berh.2012.03.002

Moe-Nilssen, R., Nordin, E., & Lundin-Olsson, L. (2008). Criteria for evaluation of measurement properties of clinical balance measures for use in fall prevention studies. *J Eval Clin Pract, 14*(2), 236-240. doi: 10.1111/j.1365-2753.2007.00839.x

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol, 63*(7), 737-745. doi: 10.1016/j.jclinepi.2010.02.006

NAV. (2014a). Legemeldte sykefraværstilfeller 4 kv 2005-2014. Hentet 06.05.15 fra: https://http://www.nav.no/no/NAV+og+samfunn/Statistikk/Sykefravar+-+statistikk/Sykefravar

NAV. (2014c). Samarbeidsavtale om et mer inkluderende arbeidsliv. Hentet 06.05.15 fra: https://http://www.nav.no/no/Bedrift/Inkluderende+arbeidsliv/Relatert+informasjon/IA-samarbeidsavtale+2014-2018.353477.cms

Oja, P., & Tuxworth, B. (1995). *Eurofit for adults : assessment of health-related fitness*. Tampere: Council of Europe, Committee for Development of Sport.

Ostelo, R. W., de Vet, H. C., Knol, D. L., & van den Brandt, P. A. (2004). 24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. *J Clin Epidemiol, 57*(3), 268-276. doi: 10.1016/j.jclinepi.2003.09.005

Pallant, J. (2013). *SPSS survival manual / a step by step guide to data analysis using IBM SPSS* (5th ed.). Maidenhead: McGraw-Hill.

Polit, D. F., & Beck, C. T. (2012). *Nursing research: generating and assessing evidence for nursing practice*. Philadelphia, Pa.: Wolters Kluwer Health.

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research : applications to practice* (3rd ed.). Upper Saddle River: Pearson Prentice Hall.

Pran, F. (2007). ICF - et felles språk for funksjon. *Fysioterapeuten*(7), 24-26.

Rankin, G., & Stokes, M. (1998). Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil, 12*(3), 187-199.

Roland, M., & Morris, R. (1983). A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine (Phila Pa 1976), 8*(2), 141-144.

Shrout, P. E., Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Physological Bulletin, 86*(2), 420-428.

Skouen, J. S., Kvåle, A. (2006). Different outcomes in subgroups of patients with long-term musculoskeletal pain. *Norsk Epidemiologi, 16*(2), 127-135.

Strand, L. I., Anderson, B., Lygren, H., Skouen, J. S., Ostelo, R., & Magnussen, L. H. (2011). Responsiveness to change of 10 physical tests used for patients with back pain. *Phys Ther, 91*(3), 404-415. doi: 10.2522/ptj.20100016

Strand, L. I., Moe-Nilssen, R., & Ljunggren, A. E. (2002). Back Performance Scale for the assessment of mobility-related activities in people with back pain. *Phys Ther, 82*(12), 1213-1223.

Streiner, D. L., Cairney, J., & Norman, G. R. (2015). *Health measurement scales : a practical guide to their development and use* (5th ed.). Oxford: Oxford University Press.

Suni, J. H., Oja, P., Laukkanen, R. T., Miilunpalo, S. I., Pasanen, M. E., Vuori, I. M., . . . Bos, K. (1996). Health-related fitness test battery for adults: aspects of reliability. *Arch Phys Med Rehabil, 77*(4), 399-405.

Suni. J, H. p., Rinne. M. (2009). Fitness for Health: The ALPHA-FIT Test Battery for Adults Aged 18-69, Tester´s Manual. Hentet fra: http://www.ukkinstituutti.fi/filebank/500-ALPHA_FIT_Testers_Manual.pdf

Tait, R. C., Chibnall, J. T., & Krause, S. (1990). The Pain Disability Index: psychometric properties. *Pain, 40*(2), 171-182.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., . . . de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol, 60*(1), 34-42. doi: 10.1016/j.jclinepi.2006.03.012

Tveter, A. T., Dagfinrud, H., Moseng, T., & Holm, I. (2014). Measuring health-related physical fitness in physiotherapy practice: reliability, validity, and feasibility of clinical field tests and a patient-reported measure. *J Orthop Sports Phys Ther, 44*(3), 206-216. doi: 10.2519/jospt.2014.5042

Waddell, G. (2004). *The back pain revolution* (2nd ed.). Edinburgh: Churchill Livingstone.

Wand, B. M., Chiffelle, L. A., O'Connell, N. E., McAuley, J. H., & Desouza, L. H. (2010). Self-reported assessment of disability and performance-based assessment of disability are influenced by different patient characteristics in acute low back pain. *Eur Spine J, 19*(4), 633-640. doi: 10.1007/s00586-009-1180-9

Weiner, D. K., Sakamoto, S., Perera, S., & Breuer, P. (2006). Chronic low back pain in older adults: prevalence, reliability, and validity of physical examination

findings. *J Am Geriatr Soc, 54*(1), 11-20. doi: 10.1111/j.1532-5415.2005.00534.x

WHO. (2003). ICF. Internasjonal klassifikasjon av funksjon, funksjonshemming og helse. Oslo: Sosial- og helsedirektoratet.

Wind, H., Gouttebarge, V., Kuijer, P. P., & Frings-Dresen, M. H. (2005). Assessment of functional capacity of the musculoskeletal system in the context of work, daily living, and sport: a systematic review. *J Occup Rehabil, 15*(2), 253-272.

Wittink, H. (2005). Functional capacity testing in patients with chronic pain. *Clin J Pain, 21*(3), 197-199.

Wolfe, F., Smythe, H. A., Yunus, M. B., Bennett, R. M., Bombardier, C., Goldenberg, D. L., . . . et al. (1990). The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. *Arthritis Rheum, 33*(2), 160-172.

Vedlegg 1

FUNKSJONSTESTING

ID: **Tester:** **1.testing ☐** **2.testing ☐**

Navn_____

Dato_____ Klokken. _____ Høyde_____ Vekt_____ Alder:____

Test	**Score**	**Test**	**Score**	
4. A. GBE a. Albu-slipp		**3. B. Mobilitet og styrke** a. Skulder/nakke mobilitet 5=ingen restr, 3=middels restr, 1= mye restr/lav	Hø	Ve
5. b. Lumbo-sacral flexibilitet		**13.** b. Styrke mage 11-15= God, 6-10=Middels, 0-5=Lav		
6. c. Hode rotasjon motstand		**14.** c. Ryggstyrke (B-S) 2-4 min= God, 1-2 min= Middels <1 min = Lav		
9. f. Hofte/kne fleksjon		**1. D. Back Performance Scale** a. Plukk-opp test		
10. d. Hofte sirkumduksjon		**2.** b. Finger-tupp til gulv		
11. e. Arm/skulder fleksjon		**7.** c. Sokke-test		
SUM GBE **0-6 = God** **7-24 = Middels** **25-42 = Lav**		**12.** d. Rull-opp fra liggende		
8. C. ACR – 18 pkt. 0-5 = God 6-10 = Middels 11-18 = Lav		**15.** e. Løftetest rygg (gulv-midje) 4 kg kvinner, 5 kg menn: >15x = God (0) >10-15x = Middels (1) 0-10x = Lav (2) Kan ikke løfte = 3		
16. E. Løftetest arm (midje-skulder) 2 kg kvinner, 3 kg menn >15x = God >10-15 = Middels, Kan ikke løfte/0-10 x = Lav	**Antall løft =**	**SUM BPS** **0-5 = God** **6-10 = Middels,** **11-15= Lav/dårlig**		

Vedlegg 2

Liste over test rekkefølge inter-tester reliabilitet:

Dato	Deltager	ID	Tester 1	Klokken	Tester 2	Klokken
	1		A		B	
	2		B		A	
	3		B		C	
	4		C		B	
	5		A		C	
	6		C		A	
	7		A		B	
	8		B		A	
	9		B		C	
	10		C		B	
	11		A		C	
	12		C		A	
	13		A		B	
	14		B		A	
	15		B		C	
	16		C		B	
	17		A		C	
	18		C		A	
	19		A		B	
	20		B		A	
	21		B		C	
	22		C		B	
	23		A		C	
	24		C		A	
	25		A		B	
	26		B		A	
	27		B		C	
	28		C		B	
	29		A		C	
	30		C		A	

Dato	Deltager	ID	Tester 1	Klokken	Tester 2	Klokken
	31		A		B	
	32		B		A	
	33		B		C	
	34		C		B	
	35		A		C	
	36		C		A	
	37		A		B	
	38		B		A	
	39		B		C	
	40		C		B	
	41		A		C	
	42		C		A	
	43		A		B	
	44		B		A	
	45		B		C	
	46		C		B	
	47		A		C	
	48		C		A	
	49		A		B	
	50		B		A	
	51		B		C	
	52		C		B	
	53		A		C	
	54		C		A	
	55		A		B	
	56		B		A	
	57		B		C	
	58		C		B	
	59		A		C	
	60		C		A	

Vedlegg 3

Gjennomføring av inter-tester reliabilitet på funksjonstesting i FAktA-prosjektet.

Tester 1:

1. Pasient skriver under informert samtykke.
2. Pasient fyller ut NPRS rett før 1. Testing
3. 1. Testing gjennomføres.
4. Pasient venter 30 min.
5. Tester 1 gir NPRS skjema fra 1. Testing til tester 2.
6. Tester 1 tar kopi av funksjonstest skjema, legger det i konvolutt som limes igjen. Pasient ID blir skrevet utenpå konvolutten og legges i låst skuff på Tove Ask sitt kontor. Samtykkeerklæring legges i plastmappe merket med ”samtykkerklæring” i samme skuff.

Tester 2:

1. Pasient fyller ut NPRS skjema rett før 2. Testing
2. Hvis pasient har gått opp 2 nivåer fra 1. Testing til 2. Testing ekskluderes pasienten fra reliabilitetsstudiet.
3. Testing gjennomføres.
4. Funksjonstest skjema legges i konvolutt sammen med NPRS skjema fra 1.testing og 2. Testing. Pasient ID skrives utenpå konvolutten og legges i låst skuff på Tove Ask sitt kontor.

Merknad:

- Pasient ID nummeret som brukes i reliabilitetsstudien er ID nummer i FAktA prosjektet + deltager nummer i reliabilitetsstudien.
  Eks: 270-1, 271-2, 272-3

Vedlegg 4

ID: **1.testing ☐** **2.testing ☐**

Hvor sterke smerter har du akkurat nå?

Hvordan vil du rangere plagene/smertene dine på en skala fra 0 til 10, der 0 er ingen smerte og 10 er verst tenkelig smerte?
(*Sett ring rundt det tallet på skalaen som passer best.)*

0	1	2	3	4	5	6	7	8	9	10

Ingen smerte Verst tenkelig smerte

Vedlegg 5

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:
REK Vest	Trine Anikken Larsen	55978497	05.02.20
			Deres da
			22.01.20
			Vår refera

Alice Kvåle
Kalfarveien 31

2011/2264 Muskel-skjelettplager - Funksjon, aktivitet og arbeid

Forskningsansvarlig: Universitetet i Bergen
Prosjektleder: Alice Kvåle

Vi viser til søknad om prosjektendring datert 22.01.2014 for ovennevnte forskni
behandlet av leder for REK Vest på fullmakt, med hjemmel i helseforskningslov

Vurdering

Omsøkt endring

Det skal inkluderes nye medarbeidere i prosjektet.

Forskergruppen ønsker videre å foreta følgende endringer i prosjektet:

1) Vurdere inter-tester reliabilitet av de fysiske testene som alle deltakerne gjenn
reliabilitet vil bli testet på 30 nye fortløpende deltakere.
2) Gi tilbud om deltakelse i prosjektet til ansatte i Byråd for barnehage og skole
og skjelettplager.
3) At fastleger i Bergen kommune kan gi aktuelle pasienter med langvarige musk
informasjon om studien.
4) At behandlerteamet ved nakke- og ryggpoliklinikken ved Haukeland universit
aktuelle pasienter med langvarige muskel- og skjelettplager om deltakelse i prosj

Vurdering

Lise Krohn-Hansen og Tove Dragesund inkluderes som nye medarbeidere i pros

I forhold til inter-tester reliabilitetsstudien, ønsker forskergruppen at de tre tester
funksjonsundersøkelsen i FAktA og som alle roterer på å utføre førstegangsunde
forhold til inter-tester reliabilitet. Forskergruppen mener det er viktig å undersøk
kommer til samme konklusjon ved undersøkelse av samme deltaker.

Begrunnelsen for å rekruttere deltakere i FAktA-prosjektet er at forskergruppen l
til funksjonstesting og i behandlingsstudiene for pasienter med korsryggplager o
muskelplager.

REK Vest har ingen innvendinger til omsøkte endringer.

Besøksadresse:
Armuer Hansens Hus (AHH), Tverrfløy Nord, 2 etasje. Rom 281. Haukelandsveien 28...

Telefon: 55975000
E-post: rek-vest@uib.no
Web: http://helseforskning.etikkom.no/

Vedtak
REK Vest godkjenner prosjektendringen i samsvar med forelagt søknad.

Klageadgang
Du kan klage på komiteens vedtak, jf. forvaltningsloven § 28 flg. Klagen sendes til REK Vest. Klagefristen er tre uker fra du mottar dette brevet. Dersom vedtaket opprettholdes av REK Vest, sendes klagen videre til Den nasjonale forskningsetiske komité for medisin og helsefag for endelig vurdering.

Med vennlig hilsen

Ansgar Berg
Prof. Dr.med
Komitéleder

Trine Anikken Larsen
førstekonsulent

Kopi til: *postmottak@uib.no*

Vedlegg 6

Forespørsel om deltakelse i forskningsprosjekt
”Funksjon, aktivitet og arbeid (FAktA-prosjektet) – Inter-tester reliabilitet mellom terapeuter ved funksjonstesting”

Bakgrunn og hensikt

I forskningsprosjektet ”Funksjon, aktivitet og arbeid (FAktA-prosjektet)” blir det utført en funksjonsvurdering av arbeidstakere med muskel- og skjelettplager. Dette er et spørsmål til deg om å delta i en forskningsstudie for å undersøke hvorvidt de fysiske testene som blir brukt i FAktA-prosjektet er pålitelig når testene blir utført av forskjellige fysioterapeuter. Studien i inter-tester reliabilitet gjennomføres som en del av masterstudium i fysioterapivitenskap ved Universitetet i Bergen.

Hva innebærer studien?

Deltakelse i studien innebærer at du utfører de fysiske testene to ganger. Det vil si at du utfører testene først en gang, får oppsummering av undersøkelsen og råd, og utfører de fysiske testene en gang til. I forkant av førstegangstesting og andregangstesting fyller du ut et skjema om smerte. Undersøkelsen består av ulike fysiske tester som blant annet tester; ledighet, avspenningsevne, bevegelighet, ømhet i muskulatur og styrke. Førstegangstesting gjennomfører du allerede som en del av den undersøkelsen som blir gjort gjennom FAktA-prosjektet. Andregangstesting blir gjort av en annen fysioterapeut 30 min etterpå og på samme rom.

Mulige fordeler og ulemper

Det tar ca. 20 minutter ekstra å gjennomføre andregangstestingen. Man kan ta pause når en selv ønsker og man kan når som helst trekke seg. De fysiske testene kan for enkelte medføre noe stølhet i etterkant. Det er ikke forbundet noe helserisiko ved å delta i studien.

Hva skjer med informasjonen om deg?

Informasjon som registreres om deg skal kun brukes slik som beskrevet i hensikten med studien. Alle opplysningene vil bli behandlet uten navn og fødselsnummer eller andre direkte gjenkjennende opplysninger. En kode knytter deg til dine opplysninger gjennom en navneliste. Det er kun autorisert personell knyttet til prosjektet som har adgang til navnelisten og som kan finne tilbake til deg. Det vil ikke være mulig å identifisere deg i resultatene av studien når disse publiseres.

Frivillig deltakelse

Det er frivillig å delta i studien. Du kan når som helst uten å oppgi noen grunn trekke ditt samtykke til å delta i studien. Dersom du ønsker å delta, undertegner du samtykke
erklæringen på neste side. Om du nå sier ja til å delta, kan du senere trekke tilbake ditt samtykke uten at det vil få noen konsekvenser for deg. Dersom du senere ønsker å trekke deg eller har spørsmål til studien kan du kontakte navn: Lise Krohn-Hansen, mob: 92626771, Lise.Krohn-Hansen@student.uib.no

Samtykke til deltakelse i reliabilitets-studien i FAktA-prosjektet ved Universitetet i Bergen

Jeg er villig til å delta i studien

(Signert av prosjektdeltaker, dato)

Jeg bekrefter å ha gitt informasjon om studien

(Signert, rolle i studien, dato)