

Klinisk signifikans i praksis

Tomas Formo Langkaas , Sverre Urnes Johnson , Kjersti
Stenshorne , Cecilie Collin-Tiller og KariAnne R. Vrabel

Klinisk signifikans i praksis

Klinisk signifikans er ikke bare nyttig på gruppenivå. Den statistiske tilnærmingen kan også brukes for å se på endring i enkeltpasienters behandlingsforløp.

Å følge med på pasienters utvikling er en sentral del av klinisk praksis. Innenfor psykoterapifeltet i dag finnes det to hovedtilnærminger til dette. Man kan enten gjøre det (1) gjennom vanlig klinisk observasjon eller (2) ved å ta i bruk et etablert tilbakemeldingssystem som eksempelvis *Outcome Questionnaire* (OQ; Lambert, 2015) eller *Partners for Change Outcome Management System* (PCOMS; Duncan & Reese, 2015). Det er fordeler og ulemper ved begge tilnærminger. Fordelen med etablerte tilbakemeldingssystemer er at de kommer med normer og tolkningsregler som gir vitenskapelig basert kunnskapsgrunnlag for å vurdere behandlingsutvikling. Ulempen er at disse systemene er svært generelle og i begrenset grad gir informasjon om utvikling på spesifikke behandlingsmål i spesifikke saker. Her har derimot vanlig klinisk observasjon en fordel fordi man kan følge med på helt spesifikke behandlingsmål og i større grad finjustere behandlingstilnærming etter disse observasjonene. Ulempene ved klinisk observasjon er at det ofte vil være usikkerhet ved i hvor stor grad en observert endring gjenspeiler reell endring, samt risiko for å overfortolke observerte endringer til å være av større betydning enn det er faktisk grunnlag for.

Studier viser at behandlere ofte er i overkant optimistiske om behandlingsutvikling og i for liten grad fanger opp uheldig behandlingsutvikling (Hannan et al., 2005; Walfish, McAlister, O'Donnell, & Lambert, 2012). Dette har vært sentrale argumenter for å få behandlere til å bruke tilbakemeldingssystemer med tilhørende normer og feilmarginer for tolkning. Samtidig ser det ut til at det fungerer dårlig å pålegge behandlere å bruke standardisert måling. Garland, Kruse og Aarons (2003) fant at 92 % av behandlere som ble pålagt systematisk måling, valgte å ikke forholde seg til disse målingene når de tok kliniske beslutninger. Undersøkelsen deres viste videre at vanlige motforestillinger mot måling så ut til å ordne seg i tre hovedtemaer: (1) Systematisk måling oppleves som ressurskrevende, (2) det er usikkerhet rundt målingenes gyldighet og relevans, og (3) det oppleves vanskelig å tolke måleresultater på en måte som er nyttig i enkeltsaker.

Klinisk signifikans er en statistisk tilnærming for å vurdere den praktiske betydningen av endring i resultater fra enkeltpasienters behandlingsforløp

Slik sett ser det ut til at en ideell tilnærming til systematisk måling i klinisk praksis vil være et fleksibelt tilbakemeldingssystem, med mulighet for å måle ulike behandlingsmål i ulike saker, basert på vitenskapelige måleinstrumenter med normer og kjente feilmarginer, med en tilnærming til

tolkning som har klar og direkte relevans for kliniske beslutninger i enkeltsaker. I denne artikkelen presenterer vi nettopp dette som en tredje mulig tilnærming: Gitt relevante publiserte normdata kan klinisk signifikans (Jacobson & Truax, 1991) anvendes for å omgjøre et hvilket som helst måleinstrument til et vitenskapelig basert tilbakemeldingsverktøy for å følge enkeltpasienters behandlingsutvikling.

Klinisk signifikans

Klinisk signifikans (Jacobson & Truax, 1991) er en statistisk tilnærming for å vurdere den praktiske betydningen av endring i resultater fra enkeltpasienters behandlingsforløp. Ideen om klinisk signifikans oppsto som en reaksjon på at klinisk forskning i hovedsak fokuserte på statistisk signifikante gruppeforskjeller, og dermed at statistisk signifikant endring feilaktig kan tolkes som at det er påvist klinisk endring av praktisk betydning. I tillegg kan rapportering av forskjeller på gruppenivå gi inntrykk av at alle endringer går i samme retning, og dermed dekke over viktig informasjon om endringer som fordeler seg ulikt blant pasienter. Ved å fastsette den kliniske betydningen av endring hos hver enkelt pasient er det mulig både å formidle klinisk relevant informasjon om endring hos den enkelte og å formidle informasjon om hvordan disse endringene fordeler seg på gruppenivå. Jacobson og Truax anbefalte derfor å definere klinisk signifikant endring hos enkeltpasienter som kombinasjonen av (1) statistisk signifikant endring og (2) funksjonsnivå tilsvarende det som er normalt blant friske personer.

Ulike metoder for klinisk signifikans har vært foreslått (for en oversikt, se Jacobson, Roberts, Berns, & McGlinchey, 1999), men tilnærmingen til Jacobson og Truax (1991) ser i dag ut til å ha etablert seg som en standard (Lambert & Ogles, 2009). Denne tilnærmingen er basert på å besvare to spørsmål ved en målt endring hos en enkeltpasient: (1) Er endringen større enn feilmarginer for endringsmåling, og (2) er siste måling på samme nivå som kan forventes av friske personer?

Alle målinger og observasjoner har begrensninger i presisjon, og det kan være av stor betydning å kjenne til størrelsen på feilmarginene. Dersom en observert endring er mindre enn feilmarginen for målemetoden, kan dette skyldes tilfeldig variasjon som følge av begrenset presisjon ved målemetoden. En mekanisk badevekt kan for eksempel ha en feilmargin på noen hundre gram, og dersom en slik vekt viser en vektøkning på 100 gram, stoler man ikke umiddelbart på at dette dokumenterer en faktisk økning i vekt. Det samme gjelder for andre måleinstrumenter, og dersom vi

kan tallfeste feilmarginen for å påvise endring, kan vi avgjøre om en observert endring er stor nok til å anta at forskjellen representerer en faktisk endring. Dette gir mulighet for å besvare det første spørsmålet i klinisk signifikans.

Dersom vi vet at badevekten er presis nok til å påvise en økning på 100 gram, er neste spørsmål om denne endringen er av noen praktisk betydning. Det er bakgrunnen for det andre spørsmålet i klinisk signifikans. Dette avhenger naturligvis av om det er et mål å øke i vekt, og hva det faktiske vektmålet er. Utgangspunktet for psykologisk behandling er vanligvis at en pasient har et problem som betyr at vedkommende ikke fungerer som andre. En deprimert pasient kjennetegnes ofte ved mangel på initiativ og mestringsglede, og et spesifikt mål vil da være å gjenvinne initiativ og mestringsglede, noe som vil kunne føre til at pasienten fungerer på samme nivå som andre friske personer. Det vil si at pasientens målinger skifter fra å være på et klinisk nivå til å være på et normalt nivå.

La oss illustrere konkret bruk av klinisk signifikans med en fiktiv pasient som blir behandlet for depresjon og rapporterer en nedgang i depressive symptomer fra 26 til 19 målt med *Beck Depression Inventory II* (BDI-II; Beck, Steer, & Brown, 1996). Med utgangspunkt i publiserte normdata og psykometriske egenskaper kan vi anslå at måling av endring i depresjon med BDI-II har en feilmargin på $\pm 9,5$ poeng (tekniske detaljer er beskrevet i appendiks). Nedgangen på 7 poeng (fra 26 til 19) er for liten til å kunne si at det har skjedd noen målbar endring. Det er selvfølgelig fullt mulig at pasienten er blitt mindre deprimert, men endringen er i så fall for liten til å kunne fanges opp med BDI-II. Når svaret på det første spørsmålet er nei, kategoriseres endringen som *uendret*. En økning på 10 poeng eller mer ville vært kategorisert som *verre*.

Med utgangspunkt i publiserte normdata kan vi anslå at målinger under 16 på BDI-II er vanligere for ikke-deprimerte personer enn for deprimerte (tekniske detaljer er igjen beskrevet i appendiks). En annen fiktiv pasient har en nedgang på 21 poeng (fra 27 til 6). Siden denne pasienten både har en nedgang på 10 poeng eller mer og en siste måling under 16, kategoriseres dette som *frisk*. Dersom siste måling hadde vært 16 eller høyere samtidig som nedgangen hadde vært på 10 poeng eller mer, ville endringen vært kategorisert som *bedre*. Bruk av endringskategorier i klinisk signifikans er vist i tabell 1.

Tabell 1. Beslutningstabell for endringskategorier i klinisk signifikans

1. Målbar endring?	2. Siste måling innenfor område for normal funksjon?	
	Ja	Nei
Ja, til det bedre	<i>Frisk</i>	<i>Bedre</i>
Nei, innenfor feilmargin	<i>Uendret</i>	<i>Uendret</i>
Ja, til det verre	<i>Verre</i>	<i>Verre</i>

Endringskategoriene *frisk*, *bedre*, *uendret* og *verre* tilsvarer endringskategoriene *recovered*, *improved*, *unchanged* og *deteriorated* brukt i OQ-systemet (Lambert, 2015).

Ved å anvende publiserte normdata sammen med statistisk teori for å besvare disse to spørsmålene har Jacobson og Truax laget en tilnærming til klinisk signifikans som er statistisk basert, og som samtidig kategoriserer endring på en måte som er klinisk relevant i enkeltsaker. I praksis er tilnærmingen enkel å anvende: Den forutsetter kun kjennskap til to tall, minste målbare endring (± 10 på BDI-II) og grensen som angir området for normale resultater (< 16 på BDI-II). Tilnærmingen kan brukes for et hvilket som helst måleinstrument der det finnes relevante publiserte normdata. Dette åpner for å velge måleinstrumenter tilpasset spesifikke behandlingsmål i hvert enkelt behandlingsforløp og gjør dermed at spesifikk måling kan tilpasses hver enkelt pasient.

Eksempler fra praksis

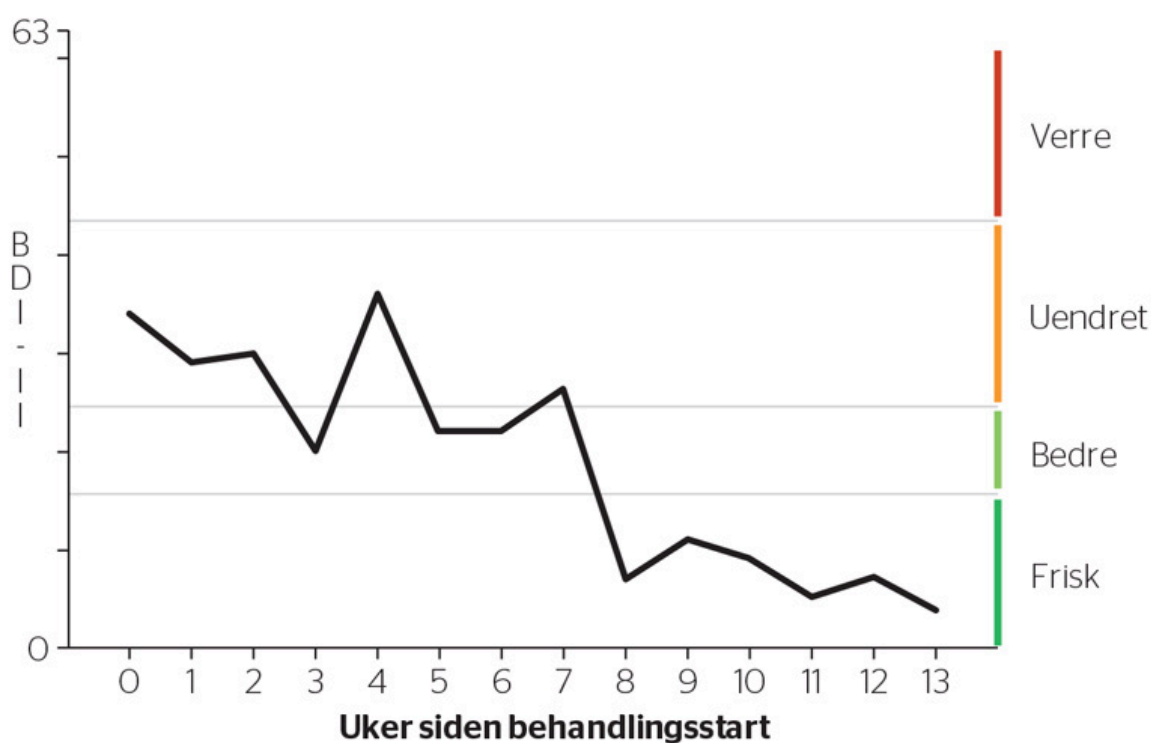
Eksempel 1 – sammensatt angst og depresjon

En pasient var innlagt på avdeling for angstlidelser ved Modum Bad med sammensatt problematikk. Ved behandlingsstart var pasienten diagnostisert med generalisert angstlidelse, sosial fobi og tilbakevendende moderat depresjon. Behandleren hadde tilgang på spesifikke måleinstrumenter for hver enkelt av disse diagnosene, og kunne dermed valgt å bruke flere måleinstrumenter samtidig for å måle utvikling på alle diagnoser. I samråd med pasienten valgte behandleren å kun bruke ett ukentlig mål på endring og anvendte BDI-II for dette.

Figur 1 viser pasientens ukentlige målinger gjennom behandlingsforløpet. Pasienten rapporterer liten endring de to første ukene, i uke 3 ser det ut til å ha vært en målbar bedring, men i uke 4 er målingen tilbake innenfor området for ingen målbar endring – innenfor feilmarginene for endringsmåling. Behandleren konkluderer med at det ikke har skjedd noen målbar endring på dette

tidspunktet. Fra tidligere pasienter som har blitt friske i løpet av behandling, har behandleren lagt merke til at målbar endring oftest skjer før uke 5. Behandleren tolker den manglende endringen som et dårlig tegn og tar med seg videoopptak fra behandlingen til veiledning. På opptaket fremkommer ulike former for unnvikelser som tidligere ikke er tilstrekkelig tematisert. Unnvikelsen og den manglende endringen drøftes med pasienten i påfølgende behandlingstime, og de gjør sammen noen endringer i behandlingstilnærmingen. Etter dette skjer det også vedvarende målbar endring i resultatene. Resultatene går kraftig ned i uke 8 og befinner seg innenfor området hvor pasienten anses som *frisk* i henhold til endringskategorier for klinisk signifikans. Resultatene holder seg på dette nivået videre i behandlingsforløpet, og målingene samsvarer med behandlerens og pasientens oppfatninger av forløpet.

FIGUR 1. Klinisk signifikans anvendt på behandlingsutvikling målt med *Beck Depression Inventory II* (BDI-II) i eksempel 1.

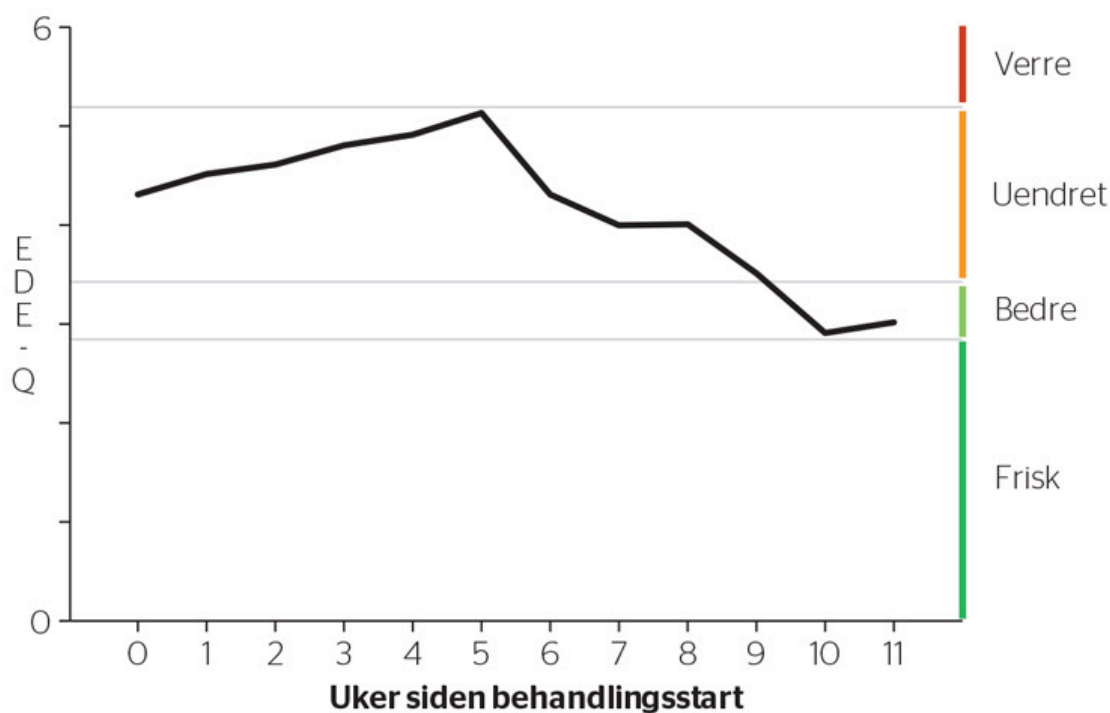


Eksempel 2 – spiseforstyrrelse

En pasient ble innlagt på avdeling for spiseforstyrrelser med diagnosene bulimi og moderat depresjon. Pasientens depressive symptomer var på innleggelsestidspunktet vanskelig å skille fra en vedvarende sorgreaksjon etter å ha mistet et nært familiemedlem 1 ½ år før innleggelsen. I samråd med pasienten ble behandlingsmål å arbeide med å normalisere egen vekt, spise regelmessige normale måltider

uten bruk av kompenserende strategier og å tåle og akseptere egen kropp. Behandleren valgte *Eating Disorder Examination Questionnaire* (EDE-Q; Fairburn & Beglin, 1994) som ukentlig mål for å følge behandlingsutvikling hos pasienten. Figur 2 viser pasientens målinger gjennom behandlingsforløpet.

FIGUR 2. Klinisk signifikans anvendt på behandlingsutvikling målt med *Eating Disorder Examination Questionnaire* (EDE-Q) i eksempel 2.



Behandlingsmålene viste seg raskt å bli vanskelige for pasienten. Pasienten opplevde aktivisering av sterke forsinkede sorgreaksjoner, strevde med å spise og startet samtidig med trening av tvangspregget karakter for å håndtere de vanskelige tankene og følelsene som ble aktivert. Personalet rundt pasienten rapporterte at pasienten viste tegn til forverring. Endringene som ble observert i de innledende ukene var derimot mindre enn feilmarginene for endringsmåling og ble av behandleren ikke tolket som reell forverring. Samtidig var det tydelig at pasienten ikke viste tegn til bedring. Behandleren valgte å drøfte med pasienten om hun fortsatt ønsket å arbeide med behandlingsmålene og å ta opp mulige unnvikelsesstrategier som så ut til å opprettholde tilstanden fremfor å endre den. De neste ukene fremkommer det en gradvis nedgang i måleresultater, men først i uke 10 er endringen større enn feilmargin for endringsmåling og tolkes som tegn på reell bedring.

Eksempel 3 – klinisk signifikans på gruppenivå

På avdeling for angstlidelser ble det opprettet et nytt behandlingstilbud for posttraumatisk stresslidelse (PTSD), og behandlerne brukte *PTSD Symptom Scale, Self-Report* (PSS-SR; Foa, Riggs, Dancu & Rothbaum, 1993) som ukentlig mål for å følge behandlingsutviklingen hos hver enkelt pasient. Ved behandlingsslutt registrerte behandlerne hvilken endringskategori pasienten endte opp i, og kunne dermed se hvordan behandlingsresultatene fordelte seg. Etter å ha behandlet 40 pasienter, ble 35 % av pasientene vurdert som friske, mens 5 % ble vurdert som målbart verre ved behandlingsslutt.

Avdelingen hadde høyere ambisjonsnivå enn de foreliggende resultatene og iverksatte derfor en rekke endringer i organisering av behandlingen, blant annet å skjerme mer tid for pasientene og behandlerne til å arbeide traumefokusert. Etter at 72 nye pasienter hadde vært gjennom behandling med endret organisering, ble det foretatt en ny gjennomgang. Behandlingsresultatene er oppgitt i tabell 2 og viste seg å ha blitt bedre.

Tabell 2. Fordeling av klinisk signifikant endring

Endringskategori	Opprinne
Frisk	
Bedre	
Uendret	
Verre	
Mangler data	

Fordeling av klinisk signifikant endring hos 112 pasienter som ble behandlet med en eller annen form for psykologisk behandling organisert på to ulike måter. Behandlingen ble evaluert ved hjelp av *Perceived Stress Scale, Self-Report* (PSS-SR), normdata for beregning av feilmargen er hentet fra Cashman, Jaycox og Perry (1997).

I prinsippet er det lite som skal til for å ta klinisk signifikans i bruk dersom kvantitative måleinstrumenter allerede anvendes i behandling. Ved å undersøke litteraturen kan man finne normdata som kreves for (a) å beregne feilmarginen for endring

og (b) å beregne hvilke resultater som ligger innenfor et normalt funksjonsnivå.

Tabell 3 viser tall for å anvende klinisk signifikans for et utvalg av spørreskjemaer som i skrivende stund er tilgjengelige via nettsidene til enten Helsebiblioteket.no (www.helsebiblioteket.no) eller Norsk Forening for Kognitiv Terapi (www.kognitiv.no).

Om å ta klinisk signifikans i bruk i praksis

Tabell 3. Tall for å anvende klinisk signifikans for et utvalg av tilgjengelige spørreskjemaer

Instrument	Endringsmåling av	Klinisk signifikans	
		Minste målbare endring	Standardavvik
PHQ-9 ^a	Depresjon	± 6	1,5
GAD-7 ^b	Generalisert angstlidelse	± 4	1,5
OCI-R ^d	Tvangslidelse	± 17	1,5
BSQ ^e	Panikklidelse	± 0,86	1,5
MI ^f (alene)	Agorafobi	± 0,67	1,5
EDE-Q ^g	Spiseforstyrrelse	± 0,90	1,5

PHQ-9 = Patient Health Questionnaire-9, GAD-7 = Generalized Anxiety Disorder-7, OCI-R = Obsessive Compulsive Inventory-Revised, BSQ = Brief Symptom Questionnaire, MI = Mobility Inventory (merk at det er tall for skalaen *unngåelse alene*), m_i = gjennomsnitt ikke-klinisk normgruppe, s_i = standardavvik ikke-klinisk normgruppe, r = reliabilitet. ^aKroenke, Spitzer og Williams (2001), ^bSpitzer, Kroenke og Williams (2002), ^cSpitzer, Kroenke og Williams (2002), ^dFoa og Tolin (1998), ^eSpitzer, Kroenke, Bress, Caputo, Bright og Gallagher (1984), ^fChambless et al. (1985), ^gFairburn og Beggs (1990).

Begrensninger ved bruk av klinisk signifikans Vi har i egen praksis valgt å plote inn grafer ved hjelp av regneark (MS Excel) der vi også kunne vise grensene for de enkelte endringskategoriene (som

vist i figur 1 og 2), noe som har gjort det enkelt å gi informativ tilbakemelding til pasientene våre. Vi kjenner så langt ikke til annen programvare som gjør det enkelt å vurdere behandlingsutvikling med bruk av klinisk signifikans utenom eksisterende programvare for OQ-systemet og PCOMS, som er bundet til disse systemenes måleinstrumenter.

Som med all annen ukontrollert observasjon av endring kan ikke klinisk signifikans brukes for å påvise effekt av behandling. Dette er en generell begrensning ved all observasjon og måling av endring i vanlig praksis og er ikke spesifikt for denne tilnærmingen. Tilnærmingen kan derimot brukes for å anslå hvilke pasienter som oppnår såpass stor endring at de fungerer innenfor det som kan regnes som normalt. En utfordring er at det eksisterer ulike måter å beregne klinisk signifikans på (for en oversikt, se Atkins, Bedics, McGlinchey & Beauchaine, 2005, og Jacobson, Roberts, Berns, & McGlinchey, 1999). Dette gjør det vanskelig å sammenligne egne resultater med behandlingsresultater fra publiserte studier, med mindre det klart fremgår at klinisk signifikans er beregnet på helt samme måte.

I likhet med mange andre statistiske tilnærminger tar metoden til Jacobson og Truax (1991) utgangspunkt i at fordelingene av både målefeil, kliniske resultater og ikke-kliniske resultater kan beskrives som normalfordelte. Ved større avvik fra denne forutsetningen er det derfor teoretisk mulig at metoden kan gi misvisende resultater, men det er så langt vi kjenner til uvisst i hvilken grad dette ville representere et faktisk problem.

Ideelt sett bør det anvendes norske eller skandinaviske normdata, dersom dette finnes

En annen begrensning gjelder tilgang på normdata. For enkelte instrumenter finnes det ikke relevante publiserte normdata for å kunne anvende klinisk signifikans. Jacobson og Truax (1991) advarer tydelig mot å bruke egne data som normdata. Ved tilgang på begrensede publiserte normdata, for eksempel kun kliniske normgrupper, beskriver Jacobson og Truax (1991) alternative måter å angi funksjonsnivå på. I praksis vil dette føre til helt andre resultater enn når alle relevante normdata foreligger, og vår løsning på dette har vært å kun anvende klinisk signifikans på måleinstrumenter der det er mulig å finne alle relevante normdata i forskningslitteraturen.

Ideelt sett bør det anvendes norske eller skandinaviske normdata, dersom dette finnes. En vanlig begrensning er nemlig at dette ikke finnes, og da er alternativet å anvende utenlandske normdata i håp om at de likevel kan gi et representativt bilde av normgruppene. Flere av eksemplene i denne artikkelen anvender utenlandske normdata, med de forbehold dette medfører.

Konklusjon

I denne artikkelen har vi vist hvordan systematisk måling kan gjennomføres på en måte som er forankret i kliniske behov for å måle spesifikke endringsmål i hvert enkelt behandlingsforløp. Bruk av klinisk signifikans tar hensyn til presisjonsnivå i målinger og minsker dermed risikoen for å tolke tilfeldig variasjon og målefeil som endring. Samtidig er klinisk signifikans forankret i empiriske grunnlagsdata om hva som tilsvarer normalt funksjonsnivå, noe som gjør det lett å vurdere klinisk betydning av en endring. En slik tilnærming er fleksibel og imøtegår vanlige motforestillinger behandlere har mot måling; behandlere får kontroll over valg av måleinstrument for hver enkelt sak og tilgang på tolkning av endring som er enkel å anvende direkte for kliniske beslutninger.

Appendiks – Beregning av klinisk signifikans

For å angi feilmargin for endringsmåling tar Jacobson og Truax (1991) utgangspunkt i sannsynlighetsfordelingen for differanseskårer gitt ingen forskjell. Ved å beregne standardmålefeilen for differanseskårer kan en målt endring omregnes til en standardskåre (Z-skåre) i denne fordelingen. Jacobson og Truax omtaler skalaen for disse skårene som *reliable change index* (RCI) og fastsetter bruk av grensene til et tosidig 95 % konfidensintervall som grenser for å vurdere en målt endring som mer enn tilfeldig. Det vil si at RCI-skårer som er mer enn 1,96 standardskårer fra null, regnes som målbar endring. I praksis er omregningen til RCI-skårer unødvendig. Siden grensene for målbar endring er fastsatt til $\pm 1,96$ RCI-skårer, kan man i stedet regne om hva 1,96 RCI-skårer tilsvarer på måleinstrumentets opprinnelige skala, og dermed finne den minste endringen som regnes som målbar endring. Minste målbare endring beregnes som

$$d = 1,96SE_{diff} = 1,96s_k \sqrt{2(1-r)},$$

der SE_{diff} er standardmålefeilen for differanseskårer¹, s_k er standardavvik i den kliniske normgruppen, og r er måleinstrumentets reliabilitet. Av formelen kan man se at økning i standardavviket vil bidra til å øke størrelsen på minste målbare endring, og at økning i reliabilitet vil bidra til å redusere den. Lambert og Ogles (2009) anbefaler å anvende intern konsistens som reliabilitetsmål (*Cronbachs alfa*) der hvor dette foreligger.

For BDI-II kan vi beregne minste målbare endring med utgangspunkt i grunnlagsdata fra Beck, Steer og Brown (1996), der $s_k = 12,15$ og $r = 0,92$ for måling av depresjon. Dette gir $d = 1,96 \times 12,15 \times (2 \times (1-0,92))^{0,5} = 9,5256$. Siden endringsskårer på BDI-II er begrenset til heltall, kan vi angi minste målbare endring som 10.

For å angi hvilke resultater som vurderes som mer sannsynlige å tilhøre den ikke-kliniske normgruppen enn den kliniske, beregner Jacobson og Truax (1991) punktet som skiller den kliniske normgruppen fra den ikke-kliniske normgruppen, nærmere bestemt det punktet mellom normgruppens gjennomsnitt der avstanden i standardskårer er den samme. Dette punktet beregnes som

$$C = \frac{m_i \times s_k + s_i \times m_k}{s_k + s_i},$$

der m_i og s_i er gjennomsnitt og standardavvik i den normale normgruppen og m_k og s_k er gjennomsnitt og standardavvik i den kliniske normgruppen.

For BDI-II kan vi beregne dette punktet med utgangspunkt i kliniske normdata fra Beck, Steer og Brown (1996), og ikke-kliniske normdata fra Aasen (2001), der $m_i = 8,12$, $s_i = 7,51$, $m_k = 26,57$ og $s_k = 12,15$. Dette gir $(8,12 \times 12,15 + 7,51 \times 26,57) / (12,15 + 7,51) = 15,17$. Siden målinger på BDI-II er begrenset til heltall, kan vi angi at resultater under 16 vurderes som mer sannsynlige å tilhøre den ikke-kliniske normgruppen enn den kliniske.

Referanser

- Atkins, D.C., Bedics, J.D., McGlinchey, J.B., & Beauchaine, T.P. (2005). Assessing clinical significance: Does it matter which method we use?. *Journal of consulting and clinical psychology, 73*(5), 982–989. <http://dx.doi.org/10.1037/0022-006X.73.5.982>
- Beck, A.T., Steer, R.A., & Brown, K. (1996). *Beck Depression Inventory. Version 2*. San Antonio: The Psychological Corporation, Harcourt Brace & Co.
- Chambless, D., Caputo, G., Bright, P., & Gallagher, R. (1984). Assessment of fear of fear in agoraphobics: The Body Sensations Questionnaire and the Agoraphobic Cognitions Questionnaire. *Journal of Consulting and Clinical Psychology, 52*, 1090–1097. <http://dx.doi.org/10.1037/0022-006X.52.6.1090>
- Chambless D.L., Caputo G.C., Jasin S.E., Gracely, E.J., & Williams, C. (1985). The mobility inventory for agoraphobia. *Behaviour Research and Therapy, 23*(1), 35–44. [http://dx.doi.org/10.1016/0005-7967\(85\)90140-8](http://dx.doi.org/10.1016/0005-7967(85)90140-8)
- Duncan, B.L., & Reese, R.J. (2015). The Partners for Change Outcome Management System (PCOMS) revisiting the client's frame of reference. *Psychotherapy, 52*(4), 391–401. <http://dx.doi.org/10.1002/jclp.20111>
- Fairburn, C.G., & Beglin, S.J. (1994) Assessment of eating disorders: Interview or self-report questionnaire? *International Journal of Eating Disorders, 16*, 363–370.
- Foa, E., Cashman, L., Jaycox, L., & Perry, K. (1997). The validation of a self-report measure of PTSD: The Posttraumatic Diagnostic Scale. *Psychological Assessment, 9*, 445–451. <http://dx.doi.org/10.1037/1040-3590.9.4.445>
- Foa, E.B., Huppert, J.D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. *Psychological Assessment, 14*, 485–496. <http://dx.doi.org/10.1037/1040-3590.14.4.485>
- Foa, E.B., Riggs, D.S., Dancu, C.V., & Rothbaum, B.O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of Traumatic Stress, 6*, 459–473. <http://dx.doi.org/10.1002/jts.2490060405>
- Garland, A.F., Kruse, M., & Aarons, G.A. (2003). Clinicians and outcome measurement: What's the use? *The Journal of Behavioral Health Services & Research, 30*(4), 393–405 <http://dx.doi.org/10.1007/BF02287427>

- Hannan, C., Lambert, M.J., Harmon, C., Nielsen, S.L., Smart, D.W., Shimokawa, K., & Sutton, S.W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*(2), 155–163. <http://dx.doi.org/10.1002/jclp.20108>
- Jacobson, N.S., Roberts, L.J., Berns, S.B., & McGlinchey, J.B. (1999). Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*(3), 300–307. <http://dx.doi.org/10.1037/0022-006X.67.3.300>
- Jacobson, N.S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. <http://dx.doi.org/10.1037/0022-006X.59.1.12>
- Kroenke, K., Spitzer, R.L., & Williams, J.B. (2001). The phq-9. *Journal of General Internal Medicine*, *16*(9), 606–613. <http://dx.doi.org/10.1046/j.1525-1497.2001.016009606.x>
- Lambert, M.J. (2015). Progress feedback and the OQ-system: The past and the future. *Psychotherapy*, *52*(4), 381–390. <http://dx.doi.org/10.1037/pst0000027>
- Lambert, M.J., & Ogles, B.M. (2009). Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data. *Psychotherapy Research*, *19*(4–5), 493–501. <http://dx.doi.org/10.1080/10503300902849483>
- Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P.Y. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general ation. *Medical Care*, *46*(3), 266–274. <http://dx.doi.org/10.1097/MLR.0b013e318160d093>
- Rø, Ø., Reas, D.L., & Lask, B. (2010). Norms for the Eating Disorder Examination Questionnaire among female university students in Norway. *Nordic Journal of Psychiatry*, *64*(6), 428–432. <http://dx.doi.org/10.3109/08039481003797235>
- Rø, Ø., Reas, D.L., & Stedal, K. (2015). Eating Disorder Examination Questionnaire (EDE-Q) in Norwegian Adults: Discrimination between Female Controls and Eating Disorder Patients. *European Eating Disorders Review*, *23*(5), 408–412. <http://dx.doi.org/10.1002/erv.2372>
- Spitzer, R.L., Kroenke, K., Williams, J.B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. <http://dx.doi.org/10.1001/archinte.166.10.1092>

- Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M.J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports, 110*(2), 639–644. [http://dx.doi.org/10.2466/02.07.17.PR0.110.2.639–644](http://dx.doi.org/10.2466/02.07.17.PR0.110.2.639-644)
- Aasen, H. (2001). An empirical investigation of depression symptoms: norms, psychometric characteristics and factor structure of the Beck Depression Inventory II. Bergen: Det psykologiske fakultet, Universitetet i Bergen. Hentet fra <https://bora.uib.no/handle/1956/1773>