En gennemgang af screening for depression med fokus på de psykometriske problemstillinger

Specialeafhandling

Simon Rasmussen

Laurits Skovlund Kilpinen

Vejleder: Henriette Kirkeby

Juni 2019

68,2 normalsider (163,800 anslag)

Indhold

Abstract

This master's thesis is based on four analysis sections investigating different aspects of the modelling of depression scales. The purpose of these analyses was to look into: a) the role of the prevalence of depression in validation studies, b) to estimate an optimal cut-off value for the CES-D scale, c) compare the two clinical interviews CIDI and SCID, and d) demonstrate why arguments based on factor analysis and Cronbach's Alpha might be misleading.

Methodologically, the most significant aspect of this thesis is the application of a new approach to meta-analysis, previously used in biomedicine.

We find that using this method our results differ from a previous meta-analysis looking into the same studies. Further, we found that estimates of optimal cut-offs must be interpreted in light of both the prevalence of the validation studies and the clinical interview utilized.

In the discussion we first investigated the basis for using clinical diagnoses as the gold standard for validating depression scales. We then discussed whether existing depression scales should be seen as measuring a reflective latent variable. The finding that this is not the case opened for a discussion of whether depression is to be seen as a natural kind. Further we looked into the correlation between depression and measures of invalidation. We then discussed the implications of these findings with a focus on limitations in the field's evidence base. Finally, we discussed the status of the depression diagnosis considering all these findings.

Indledning

I hvor høj grad kan psykologiske lidelser måles og hvordan kan en guldstandard for målingen af en psykologisk lidelse udfordres? Dette spørgsmål har længe interesseret os.

En test må grundlæggende kunne måle det den er udviklet til. Netop det forhold at testen har et formål, gør at der må kunne opstilles et kriterie for om dette er lykkedes. Det er tilfældet hvis testens resultater viser noget der svarer til den virkelighed vi oplever. Men virkeligheden kan være et flygtigt begreb, alt afhængig af hvad der søges målt.

Hvis man søger at udvikle en medicinsk test for hjernesvulster, kan denne sammenlignes med tidligere tests, ved at se på detektionsraten og ultimativt på overlevelsesraten for patienter. Fra et rent videnskabsteoretisk perspektiv, udgør dette derfor et simpelt problem, hvor virkeligheden lader sig undersøge direkte. Anderledes forholder det sig hvis svulsten er af psykisk karakter, fordi dette ikke på samme måde er direkte målbart. Lidelsen for en depressiv patient er ganske virkelig, men den er svært sammenlignelig med andres lidelser, fordi den optræder i en virkelighed som kun i ringe grad er delt.

Depressionsdiagnosen udgør et emne hvor vi mener alle disse overvejelser er relevante og er samtidig en ekstremt aktuel udfordring for folkesundheden. Ved depressiv lidelse prøver man ikke blot at afgøre hvorvidt patienten er depressiv, men også at måle sværhedsgraden af en eventuel depression. For depression går særligt to store skalaer igen i litteratur, fordi disse anses som guldstandarder. Dette er Hamilton Depression Rating Scale (HAM-D) og Becks Depression Inventory (BDI-ll), udviklet i henholdsvis 1960 og 1961. At guldstandarderne i et felt som psykometrien kan være knap 60 år gamle fandt vi i sig selv bemærkelsesværdigt. Dette bliver ikke mindre af at der er kommet omkring 280 skalaer til siden (Santor, Gregus & Welch, 2006).

Vores indledningsvise spørgsmål er således højaktuelt for depressionsdiagnosen idet det ikke er lykkedes for nogen af disse skalaer at overtage positionen som anerkendt guldstandard.

Vi ser en risiko for at pladsen som guldstandard kan være en fæstning der indtages af de der kommer først i historien. Denne fæstning ønskede vi oprindelig at udfordre ved et statistisk orienteret speciale rettet mod at optimere eksisterende skalaer ved at bryde med antagelsen om at alle symptomer vejer lige tungt. Det lod sig imidlertid ikke gøre, da den nødvendige data er svært tilgængelig. Vi valgte derfor at gå en anden vej og undersøger i dette speciale de psykometriske egenskaber for eksisterende depressionsskalaer, med primært fokus på diagnosticeringsnøjagtighed.

Emnet for specialet finder vi i høj grad interessant, fordi det udgør en fundamental test-teoretisk problemstilling om hvordan vi bedst kan måle psykologiske begreber. Imidlertid er dette også, som indikeret ovenfor, enormt aktuelt når vi ser på prævalensen og betydningen af depressiv lidelse.

Depression har i nyere tid overhalet iskæmisk hjertesygdom og er nu ifølge World Health Organization (WHO), den ledende årsag til invalidering på verdensbasis (WHO, 2018). De seneste estimater fra USA viser, at livstidsprævalensen for depression er omkring 16% (McKnight, 2009). Af denne grund er depression blevet kaldt ”the common cold of mental disorders” (Cuijpers, Reijnders, Karyotaki, de Wit & Ebert, 2018). Samtidig viser estimater for behandlingen af depression, at omkring halvdelen af alle med depression kommer i behandling og at omkring 60% af de der kommer i behandling ikke responderer på denne (McKnight, 2009). Det sidste vidner om en markant begrænsning i behandlingen af depression og netop spørgsmålet om behandlingseffekter har igennem årene været en åben diskussion. Forskningen i psykoterapi har afdækket en række interessante fund, herunder eksempelvis det, at erfaringsniveau for

kliniske psykologer ikke er en prædiktor for effekten af psykoterapi (Erekson, Janis, Bailey, Cattani & Pedersen, 2017). Fund som disse gør det interessant for os at se på grundlaget for dette – om skalaerne holder. Dette leder os til vores vinkling af opgaven.

Vores emne her er bredt og vi har fra starten ønsket undgå at dykke ned i en enkelt psykometrisk faktor og isolere denne. Dette gør på den ene side at vi kommer omkring mange forskellige aspekter af emnet, men medfører også en risiko for at læseren vil opleve specialet som meget bredt. Dette er en risiko vi er os bevidste om. Særligt ønsker vi med denne bredde at afdække to perspektiver på emnet. De første dele af vores analyse fungerer implicit ved den antagelse at depressionskonstruktet har acceptable psykometriske egenskaber. I senere afsnit ændrer vi perspektiv til netop at undersøge i hvor høj grad dette er tilfældet. Som det vil blive tydeligt i løbet af opgaven, finder vi betydelige psykometriske begrænsninger. Uanset omfanget af disse ser vi det imidlertid som overvejende usandsynligt, at tilgangen til begrebet ændres væsentligt over de næste år, historiske tendenser taget i betragtning.

For de første tre dele af analysen udgør cut-off værdier for depressionsskalaer omdrejningspunktet. Cut-off scorer spiller en helt central rolle, fordi det kan forme den enkeltes videre liv ved hvorvidt der tilbydes behandling eller ikke, og ved det at disse anvendes i forskningen når det opgøres hvor mange der erklæres raske efter en behandling. Videre udgør brugen af cut-offs et helt konkret punkt hvor vi mener at kunne forbedre den mest udbredte tilgang. Dette gør vi ved anvendelsen af en af de nyeste tilgange til meta-analyse indenfor biomedicin.

I analysens første del ser vi på betydningen af hvor høj prævalens der er i det enkelte studies sample. Dette fører til en længere diskussion omkring hvordan et optimalt cut-off defineres.

I analysens anden del anvender vi en ny tilgang til meta-analyse og estimerer et optimalt cut-off herudfra. Dette leder til diskussion af betydningen af denne nye metodiske tilgang.

I analysens tredje del undersøger vi om der er forskel imellem studier der anvender to forskellige kliniske interviews.

Fælles for særligt de to første analysedele er at vi undersøger rent metodiske forhold i estimeringen af det optimale cut-off, mens vi ikke berører betydningen af guldstandarden. I den tredje analysedel åbner vi op for en diskussion af netop dette, i det vi diskuterer betydningen af vores sammenligning af de to interviews. Denne skepsis overfor guldstandarden former det næste afsnit hvor vi skriver om beslutningstagning og ser på potentielle problemer ved at anvende den kliniske diagnose som guldstandard for skalaer.

Herefter skifter vi i nogen grad perspektiv og ser på 30 valideringsstudier for Becks depression Scale (BDI-ll). Her fokuserer vi på hvad disse studier finder i to af de mest centrale statistiske analyser for valideringsstudier. Efter denne del går vi over til diskussion.

I diskussionsafsnittet gennemgår vi en række aspekter af hvad disse analyser betyder for berettigelsen af den nuværende tilgang til begrebet. Særligt undersøger vi om der er basis for at se depression som en refleksiv latent variabel, en model vi uddyber i teoriafsnittet, og om depression udgør en naturlig art (engelsk: natural kind). Endelig forsøger vi at sammenfatte hvad vi kan udlede af dette brede fokus, herunder en række implikationer for forskningen. Her kommer vi også tilbage til nogle af de spørgsmål vi har undret os over, behandlingseffekt etc.

Afslutningen på denne gennemgang leder os til en kort note, inden vi præsenterer opgavens problemformulering.

Denne opgave vil have en kritisk tilgang, fordi det er målet at forsøge at belyse berettigelsen af den måde depression operationaliseres. Denne vinkel på depressionsdiagnosen kan lyde meget skeptisk, særligt når vi i diskussionen diskuterer om depression bør ses som en såkaldt naturlig art. Her er det vigtigt for os at understrege at kritikken gennemgående er en metodekritik og en diskussion af implikationerne af vores fund. Det er vigtigt at slå fast at uafhængigt af hvilke psykometriske forhold der gør sig gældende for depressionsbegrebet, ændrer det intet ved hvordan patienten har det. Kritikken er således aldrig rettet mod den deprimerede, men mod det felt der har til mål at hjælpe den deprimerede – vores eget felt.

Problemformulering

Hovedspørgsmål

-Hvilke psykometriske problemstillinger er der i valideringsprocessen for depressionsskalaer og hvilken betydning har disse for skalaernes psykometriske egenskaber?

Underspørgsmål

-Hvordan kan cut-off bedst estimeres og hvordan stemmer det overens med hvad, der bliver gjort?

-Hvilke antagelser ligger der bag de klassiske valideringsanalyser og i hvilket omfang er der brud på disse i praksis?

-Hvilken konsekvenser vil det have for skalaernes psykometriske egenskaber, hvis der er brud på antagelserne?

Teori

Det historiske fundament for depressionsbegrebet

For at forstå begrebet depression er det nødvendigt at forstå, hvordan begrebet er opstået og hvorfor vi forstår depression på den måde vi gør i dag,

Historien bag depression går mange år tilbage. Forløberen for depression var en lidelse, man kaldte for melankoli som kan spores tilbage til Hippokrates (460-351 f.kr), der definerede melankoli som frygt eller kedsomhed over lang tid (Horwitz, Wakefield & Lorenzo-Luaces, 2016). Hippokrates beskrev symptomer for melankoli, der ligner de nuværende DSM-kriterier for depression, Hippokrates beskriver symptomer som: Spisevægring, modløshed, søvnløshed, irritabilitet og uro (Ban, 2014; Horwitz et al., 2016). Årsagerne til melankoli mente man var uligevægt i de fire temperamenter og primært en overvægt af sort galde. Denne humorapatologiske forklaringsmodel for depression var den primære forståelsesmodel op til 1700-tallet (Shorter, 1992).

Som følge af den empiriske og observationelle metodes fremgang i det 18. århundrede bliver den humoralpatologiske måde at forstå melankoli erstattet af teorier omkring forstyrrelser i hjernen og nærmere bestemt i nervesystemet. Forskning i hjernelæsioner, fører til idéen om at depressionssymptomer skyldes nervesygdomme (Shorter, 2013)

Hvilket betyder man bryder med tiders holistisk syn på mentale lidelser. Det bliver teoretiseret at sygdomme har en naturlig form, som alle har en ensartet manifestation (Ban, 2014)

Den tyske psykiater Emil Kraepelin sammenkoblede i denne periode depression med manilidelser, Kraepelin mente at depression og den depressive del af manilidelserne var umulige

at adskille kvalitativt og de derfor måtte have samme ætiologi og samme underliggende fysiologi (Greenberg, 2010). Det var også Kraepelin, der er skyld i at depression bliver set som en diskret lidelse frem for et kontinuum (Leite, Macedo, Borges & Santos, 2017).

I 1952 udkom DSM-1 i kølvandet på anden verdenskrig, som en hjælp til amerikanske psykiatere med at behandle hjemvendte amerikanske soldater med psykiske mén (Blashfield, Keeley, Flanagan & Miles, 2014).

I DSM-I var der tre såkaldte psykotiske lidelser og en af disse var affektive lidelser, hvor melankolsk depression var grupperet sammen med manilidelse. Psykoneurotisk depression var også beskrevet i DSM-I, grupperet med angstlidelser. Efter psykodynamisk tradition bliver denne lidelse set som en manifestation af en angstlidelse og en psykologisk forsvarsmekanisme mod denne (Horwitz et al., 2016). Der var ikke de store ændringer på de to diagnoser i DSM-II.

Der var stærk psykodynamisk inspiration bag DSM-I og DSM-II og det gjorde, at man i disse så symptomer som refleksioner af en underliggende psykodynamisk konflikt og at symptomer kun var meningsfulde, når man udforskede den personlige historie af hver patient. Af denne grund gjorde DSM-I og DSM-II meget lidt ud af at udvikle klassificeringssystemer. Psykiaterne fokuserede mere på underliggende mentale konflikter end på patienternes symptomer (Worboys, 2013; Shorter, 2015)

I 1970'erne kommer der mere og mere pres fra forsikringsselskaber på psykiaterne. Forsikringsselskaber er ikke tilfredse med, at så mange mennesker skal have antidepressiva og psykoterapi, og efterlyser klare retningslinjer for, hvem der skal og hvem der ikke skal have behandling. (Mayes & Horwitz, 2005; Ban, 2014)

I 1976 publicerede den amerikanske psykiater R.E Kendell en artikel, hvori han beskrev at der fandtes 12 store forskellige depressionsklassificeringssystemer, som alle var meget forskellige (Kendell, 1976).

Det fører til at American Psychiatric Association (APA) valgte en komite til at kigge på diagnosticeringskriterierne med udgangspunkt i Kendell's artikel. Oprindeligt var det meningen at DSM-III blot skulle ændre i nomenklaturen fra DSM-II, men da DSM-III udkommer i 1980 har den over natten ændret psykiatrien markant. Psykiatrien blev en disciplin, hvor diagnosticering var centralt, i modsætning til tidligere hvor diagnosticering spillede en minimal rolle (Mayes & Horwitz, 2005; Leite et al., 2017).

DSM-III inkluderer for første gang diagnosen, Major Depression Disorder (MDD). Symptomerne for lidelsen er næsten identiske med kriterierne fra en af de 12 klassificeringssystemer fra Kendell's artikel fra 1976, nemlig Feighnerkriterierne (Kendell, 1976; Ban, 2014). Feighnerkriterierne hed i Kendell's artikel St. Louis Klassifikationen og var udviklet af en gruppe psykiatere fra Washington University (Kendell, 1976).

For at opnå diagnosen depression, skal man i Feighnerkriterierne have dysforisk humør enten ved at være nedtrykt eller ved håbløshed. Derudover skulle man have 5 af følgende symptomer: tab af appetit, problemer med at sove, uro (agitation), mindsket interesse i normale aktiviteter, skyldfølelse, langsom tænkning og gentagende selvmordstanker.

Symptomerne skal være der i mindst en måned og må ikke skyldes anden psykisk eller fysisk lidelse (Feighner, Robins, Guze, Woodruff, Winokur & Munoz, 1972).

Feighnerkriterierne for depression er konstrueret ud fra 5 studier, hvoraf kun et af studierne er empirisk funderet (Cassidy, Flanagan, Spellman & Cohen, 1957).

I dette studie bliver 100 patienter, betegnet som manidepressive og 50 fysisk syge patienter, testet via en symptomcheckliste for at se hvilke symptomer de har.

I studiet opstillede de to krav for depression: Patienterne skulle mindst en gang have nævnt at deres humør havde ændret sig. Derudover skulle de også have mindst 6 af disse 10 symptomer: Langsomt tænkende, dårlig appetit, forstoppelse, insomnia, følelse af træthed, mindsket koncentration, ”over-talkativeness” eller ”press of complaints”. Det fremgår ikke hvordan denne checkliste er udarbejdet (Cassidy et al., 1957).

Feighnerkriterierne har blot fire ændringer fra Cassidy et al., (1957) studiet. Forstoppelse er droppet, skyldfølelse er tilføjet, insomnia er udvidet til søvnproblemer og vægttab er kombineret med anoreksi (Feighner et al., 1972).

Derudover tilføjede de at symptomerne skal være der i en måned, hvilket ikke var med i Cassidy et al. studiet. Det kan skyldes at de i dette studie alle var indlagt og de fleste af dem havde haft symptomerne i mere end 6 måneder (Cassidy, 1957; Shorter, 2013).

Deltagerne i studiet var meget syge noget forfatterne også var opmærksomme på, derfor skrev de: “The question immediately arises as to whether all these patients did, in fact, have manic-depressive disease. At present, one cannot go beyond saying that the patients had a psychiatric illness” (s. 1542, Cassidy et al., 1957)

Feighnergruppen mente selv at kriterierne ikke skulle ses som værende færdige, men blot skulle fungere som en start (Feighner et al., 1972). R. E. Kendell (1976) lagde ikke specielt vægt på Feighnerkriterierne og skrev ”As a research strategy it has a great deal to commend it. But no evidence has been offered to suggest that it is anything more than a convenient strategy.” (s. 24 Kendell, 1976).

Den daværende chef for APA DSM-III task force Robert Spitzer, udtalte efter DSM-III var udkommet, at den nye klassifikation af MDD var politisk drevet og at sygdomsdefinitioner i næste udgave af DSM ville blive mere videnskabelige (Leite et al., 2017).

Efterfølgende er sorg-kriteriet blev fjernet i DSM-5, ellers er kriterierne for depression næsten identiske med dem fra DSM-3 og dermed næsten identiske med kriterierne fra Cassidy et al. (1957) (Shorter, 2015; Ban, 2014).

Det fører os til begrebet depression i dag, som beskrevet har begrebet ikke ændret sig markant fra 1980. Nu vil vi forlade den historiske vinkel og beskrive hvordan depression bliver forstået i dag og hvordan vi vil definere det i vores opgave.

Konstruktet depression

Depression (engelsk: major depression) er defineret i den europæiske diagnosticeringsmanual International Statistical Classification of Diseases and Related Health Problems-10 (ICD-10) og i den amerikanske The Diagnostic and Statistical Manual of Mental Disorders (DSM-5).

Når vi i dette projekt omtaler depression, er det med henvisning til diagnosen depression, som i DSM-5 manualen hedder Major Depressive Disorder.

De to definitioner i hhv. ICD og DSM er stærkt ensartede. Forskellen i definitionen af diagnosen for ICD-10 og DSM-5 er meget begrænset (Leita et al., 2017). Dette skyldes at ICD-10 diagnosen for depression er formuleret med fokus på at ligne DSM-5 diagnosen (Bech et al., 2005).

Indenfor forskning i depression anvendes stort set udelukkende definitionen fra DSM-5 (Mitchell

& Coyne, 2010). Det gør DSM-5 særligt relevant for dette speciale og vi vil derfor som udgangspunkt referere til DSM-5 og ikke til ICD-10.

Depressionsskalaer

Diagnosticeringskriterierne for depression i både DSM og ICD håndterer symptomer som værende enten tilstede eller ikke-tilstede (World Health Organization, 1992; American Psychiatric Association, 2013). Det besværliggør muligheden for at måle sværhedsgraden af depression udelukkende ved hjælp af diagnosticeringskriterierne. Derfor bruges depressionsskalaer ofte i forskningen til at kvantificere depressiv lidelse (Mitchell & Coyne, 2010).

Der eksisterer i dag over 280 forskellige depressionsskalaer (Santor, Gregus & Welch, 2006). Det skyldes blandt andet mange måder at konceptualisere depression på, fra kognitivt, til psykodynamisk, udviklingsmæssigt, og til de mere somatiske måder (Mitchell & Coyne, 2010). Størstedelen af interventionsstudier benytter sig af de samme seks depressionsskalaer. De to mest anvendte, Beck og Hamilton, bliver brugt i 40 % af alle interventionsstudier (Santor et al., 2006). Det høje antal er også en afspejling af at flere og flere antidepressiva er udviklet løbende (Faravelli, Albanesi & Poli, 1986).

Der findes grundlæggende to slags depressionsskalaer interview-baserede skalaer (IBS) og selvvurderingsskalaer (SVS) (Bech et al., 2005). Som betegnelsen lægger op til, er det i interview-baserede skalaer en kliniker, som udfylder skalaen og i selvvurderingsskalaer er det personen, der testes, som udfylder skalaen. Det er svært at vurdere om, hvilken slags skala, der er mest valid, der er fund der favoriserer SVS (Trivedi et al., 2004) og ligeledes er der fund, der favoriserer IBS (Edwards, Lambert, Moran, McCully, Smith & Glade Ellingson, 1984; Prusoff,

Klerman & Paykel, 1972). Et stort studie fra Uher et al. (2007) fandt at der ikke var forskel på SVS og IBS, men derimod mellem skalaerne.

Der er dog forskelle i resultaterne mellem IBS og SVS. Der er studier, som tyder på en generel tendens til at patienter vurderer symptomer højere end klinikere, når de selv skal rate dem (Faravelli et al., 1996; Trivedi et al., 2004). Carroll et al. (1981) lavede en selvvurderingsskala med nøjagtigt de samme items som HAM-D, og fandt en korrelation på 0,8 mellem denne og den originale HAM-D.

Ligesom der er stor spredning i måden skalaerne konceptualiserer depression, er konstrueringsprocessen af skalaerne også meget forskellig.

De to mest anvendte depressionsskalaer, Hamilton Rating Scale for Depression (HAM-D) (Hamilton, 1960) og Becks Depression Inventory (BDI) (Beck, Ward, Mendelson, Mock & Erbaugh, 1961) blev lavet med et års mellemrum i henholdsvis 1960 og 1961 (Hamilton, 1960; Beck et al., 1961), altså ca. 20 år før MDD blev en diagnose.

HAM-D blev lavet af englænderen Max Hamilton og ideen bag var at lave en, interview-baseret skala, der kunne måle sværhedsgraden og symptomerne af depression, men for patienter, der allerede har diagnosen. Man manglede en skala til at måle effekten af antidepressiva (Hamilton, 1960). Som beskrevet tidligere i teoriafsnittet, var depressionsdiagnosen noget anderledes dengang. Items i HAM-D er udvalgt ud fra Max Hamiltons egen kliniske erfaring og medicinsk litteratur (Hamilton, 1960; Bech & Coppen, 1990). BDI blev udviklet af den amerikanske, daværende psykoanalytiker Aaron T. Beck. Denne skala blev udviklet for at måle de adfærdsmæssige manifestationer af depression (Beck et al. 1961). Items blev genereret ud fra Becks systematiske observationer af deprimerede patienter, der undergik psykoanalytisk terapi hos Beck (Beck et al., 1961). Beck og Hamilton så meget forskelligt på depression. Hvor

Hamilton så det som en lidelse med mange somatiske symptomer, så Beck det som en lidelse med affektive-, kognitive-, motivations- og adfærdskomponenter. Modsat Hamilton, lagde Beck ikke vægt på somatiske symptomer (Bech & Coppen, 1990).

Af andre skalaer, der ofte bliver brugt, kan Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1977) og Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery & Åsberg, 1979) også nævnes. Ligesom med HAM-D og BDI er der stor forskel på hvordan disse skalaer er udviklet. MADRS blev lavet til at måle ændringerne af behandling med antidepressiva bedre end HAM-D. Skalaen blev konstrueret ved at 106 patienter tog en 67 items test, hvor de 17 mest hyppige depressionsitems blev udvalgt ud fra et arbitrært cut-off på 70 %. Derefter tog 64 patienter, der var i antidepressiv behandling en test med de 17 items, og så blev de 10 items som var bedst til at måle effekten af behandlingen udvalgt (Montgomery & Åsberg, 1979).

CES-D blev lavet til brug i epidemiologiske studier. Her blev items valgt, fra en stor samlet pulje af items fra andre depressionsskalaer (Radloff, 1977), og primært med fokus på affektive og somatiske aspekter af depression. En tredjedel af alle items i CES-D fremgår ikke i andre af de mest anvendte depressionsskalaer (Fried, 2015).

Af de fem mest bruge depressionsskalaer, er fire af dem fra før 1980. Der er derfor også flere af dem, som i varierende grad, er blevet revideret i tidens løb. BDI blev revideret til BDI-II i 1996 (Beck, Steer, Ball & Ranieri, 1996), for at skalaen skulle passe bedre med DSM. Hvorimod HAM-D sidst blev revideret i 1967, hvor item'et "agitation" blev ændret, så man kunne få 0-4 point stedet for 0-2 point (Hamilton, 1967).

Der er stor heterogenitet i items mellem skalaerne. De 7 mest brugte depressionsskalaer (heriblandt HAM-D, BDI, CES-D og MADRS) indeholder 52 forskellige symptomer hvor 21

items er unikke til en skala, hvilket betyder at der var 21 items, der kun fremgik i en skala (Fried, 2016). Det varierer også hvor mange af symptomerne fra DSM, skalaerne inkluderer. HAM-D adskiller sig signifikant fra DSM og inkluderer kun fire af ni depressionskriterier fra DSM (Bech & Coppen, 2010), hvorimod CES-D har syv af ni DSM-symptomer (Okun, Stein, Baumna & Johnson, 1996).

Validitet

Validitet kan defineres som at en test er valid, når den måler det den er konstrueret til at måle (Nunnally & Bernstein, 1994).

Cronbach & Meehl's (1955) udlægning af validitetsbegrebet fremhæves ofte, med fokus på deres inddeling af validitet i fire undertyper: begrebsvaliditet (eng: construct validity), konvergerende validitet (eng: convergent validity), forudsigelses validitet (eng: predictive validity) og samtidighedsvaliditet (eng: concurrent validity).

Denne klassiske opdeling er at finde i de fleste lærebøger om psykometri. Senere har definitionen ved opdeling mødt modstand. Her har særligt Borsboom, Mellembergh & Heerden (2004) leveret en markant kritik af den oprindelige definition og betegnet udviklingen af denne som særligt negativ i validitetsbegrebets historie. De argumenterer for at begrebet validitet dækker over noget langt simplere end hvad både Cronbach og Meehl og den teoretiske litteratur generelt angiver. De henviser til at Kelley allerede i 1927 angav, at en test er valid hvis den måler hvad den skal måle. Uddybende opstiller de selv to kriterier for hvorvidt en test er valid: a) en test er valid til målingen af et begreb, hvis det begreb eksisterer, og b) der er en kausal sammenhæng så ændringer i begrebet forårsager ændringer i testen (Borsboom et al., 2004).

Begrebsteoretiske modeller

Vi vil i denne opgave beskæftige os med to måder at anskue depression på, hvilket også er de mest anvendte måder at anskue et sådant begreb. Det er henholdsvis som en formativ latent variabel og som en refleksiv latent variabel.

Den refleksive latente variabel

Ved en refleksiv latent variabel forstås en variabel, der ikke er direkte observerbar og som lader sig reflektere/afspejle/manifestere i andre variabler (Edwards & Bagozzi, 2000). Refleksive latente variabler operationaliseres derfor ved måling af en række items der fungerer som proxy indikatorer for den latente variabel (Bollen & Lennox, 1991; Edwards, 2011). Det forhold at den latente variabel ikke observeres, leder til teststøj (eng: test error) relateret til hver proxy indikator.

At de målte items ses som proxy indikatorer afspejler at der ved refleksive latente variabler er en klar antagelse om kausalitetsforholdet. Der antages en kausal envejssammenhæng, hvor de enkelte items er forårsaget af den latente variabel.
Af dette følger det at manipulering af den latente variabel, såfremt muligt, vil afspejles i de målte proxy indikatorerne. Modsat vil manipulering af disse ikke indvirke på den latente variabel (Edwards & Bagozzi, 2000).

I praksis vil observerbare symptomer kunne hjælpe med at identificere en sygdom, når denne ses som en refleksiv latent variabel. Ved en sådan situation følger det af definitionen, at symptombehandling ikke vil ændre noget i sygdommens tilstedeværelse og sværhedsgrad. I en

praktisk kontekst kan målet være enten at afgøre tilstedeværelsen af den latente variabel eller at afgøre et kvantitativt spørgsmål relateret til den latente variabel.

Hvor den kausale sammenhæng er entydigt angivet i litteraturen, står det mindre klart hvad det at et begreb ses som en refleksiv latent variabel, stiller af krav til selve begrebets natur.

Fried (2017) skriver at det er indforstået i den refleksive latent variabel model at begrebet man undersøger er en naturlig art.

Det betyder at det er en urokkelig enhed, der eksisterer uafhængigt af hvorvidt den bliver anerkendt. At arter er naturlige, betyder at de er grupperet efter en struktur, der reflekterer strukturen af den naturlige verden, frem for menneskelige interesser og handlinger (Fried, 2017).

Naturlige arter har indre træk, der bestemmer en naturlig sæt ”kind”-medlemmer. Et eksempel fra naturvidenskaben er grundstoffet Zink, der har 30 protoner, og alt med 30 protoner er Zink.

Det er sygdommen, der forårsager indikatorerne. Der er ingen kausal forbindelse mellem symptomerne og ingen kausal forbindelse fra symptomerne til sygdommen. De er passive indikatorer. Videre er samtlige symptomer, hvilket i en skala kontekst svarer til items, alle udvalgt som alternative og indbyrdes komplimentære proxy mål for det underliggende begreb.

Disse forhold betyder at de enkelte variabler eller proxier forventes at korrelere positivt via den latente variabel (Bollen & Lennox, 1991; Diamantopoulos & Siguaw, 2006)

Den formative latente variabel

Ved en formativ latent variabel forstås en latent variabel som defineres ud fra de variabler, der indgår i operationaliseringen af variablen (Coltman, Devinney, Midgley & Venaik,

2008; Fried, 2017). Kausaliteten er derfor modsat af hvad der er gældende for den reflektive latente variabel.

Det følger af den definitionsmæssige antagelse om kausaliteten, at målingen af den latente variabel vil afspejle hvilke variabler der måles. Modsat den refleksive latent variabel model, er det latente begreb i den formative latent variabel model summen af indikatorerne, og derfor i sig selv socialt konstrueret. Dette forhold gør at de enkelte variabler/items i en formativ model ikke ubetinget forventes at korrelere positivt (Bollen & Lennox, 1991).

Modsat tilfældet ved en refleksiv variabel, afspejler de enkelte variabler/items ved en formativ variabel forskellige aspekter af begrebet (Bollen & Lennox, 1991; Diamantopoulos & Siguaw, 2006). Hvor der ved den refleksive variabel model tales om 'nyttig overflødighed' (eng: useful redundancy) (DeVellis, 2006), angiver flere kilder at der for formative modeller bør søges en minimering af overlap således at hvert item/proxyvariabel måler et særskilt aspekt af begrebet (Edwards, 2011).

Man kan ikke manipulere en formativ latent variabel uden at manipulere variabler, der udgør den latente variabel. Hvis en sygdom betragtes som en formativ variabel, vil der derfor ikke være tale om at manipulere den latente variabel, som ved en refleksiv variabel, men derimod at ændre de enkelte målte variabler, hvorved summen af disse vil ændre sig. På samme måde vil man i forskning skulle undersøge de forskellige symptomer og hvilke problemer de giver. Det giver heller ikke mening at forsøge at finde korrelationer mellem den latente variabel og biomarkører eller lignende (Diamantopoulos & Siguaw, 2006; Edwards, 2011)

Det kan f.eks. være penge, som kun eksisterer i en social konstruktion. Det vil sige social kind er begreber, der er produceret og ikke fundet. Begrebet eksisterer ikke, hvis dem der har konstrueret begrebet ikke findes.

Depression, bliver af de fleste teoretikere set som en refleksiv latent variabel, det betyder at man kan benytte sig af klassisk testteori til at undersøge reliabiliteten af begrebet. Hvilket også er kutymen (Fried, 2015).

Klassisk testteori

Reliabilitet henviser til hvor konsistent et mål er over tid, imellem forskellige test sekvenser (Nunnaly & Bernstein, 1994). Det er vigtigt indenfor psykometrien, fordi man ofte ønsker at måle begreber som forventes at være stabile over tid, såsom personlighedstræk eller intelligens. Reliabilitet hænger også sammen med reproducerbarhed. For at en skala kan være brugbar, skal den være konsistent på den måde at den producerer mere eller mindre det samme resultat for den samme testperson, hver gang den bliver brugt (Gregory, 2011).

Til at måle reliabiliteten kan man som skrevet benytte sig af klassisk test teori, som er en konventionel kvantitativ tilgang til at teste reliabiliteten af en skala, baseret på skalaens items (Cappelleri, Lundy & Hays, 2014).

Klassisk testteori er baseret på fem antagelser. Dette grundlag muliggør udledningen af en række sammenhænge og omskrivninger, der kan bruges til at tolke reliabiliteten af en test.

Den grundlæggende antagelse i klassisk testteori er axiomet:

$$X=T+E \qquad (1.1)$$

(Crocker & Algina, 1986; Nunnally & Bernstein, 1994). Som beskrevet ovenfor betyder det at den observerede variabel (X) består af true score (T) og tilfældig støj (eng: random error) (E).

Definitionen på true score er den gennemsnitlige score ved gentagen sampling uden træningseffekt eller anden ændring i testpersonen (Slocum, 2005). Det er den forventede score og ikke den bedste eller højest mulige score (Irwing, Booth & Hughes, 2018). True score er ikke en score opnået på en dag, hvor testpersonen yder den højest eller lavest muligt opnåelige personlige score, men scoren for en gennemsnitlig dag. True scoren er en hypotetisk score, som ikke observeres i praksis (Nunnally & Bernstein, 1994).

For at kunne forstå klassisk testteori er det nødvendigt at have en forståelse for begrebet tilfældig støj. Tilfældig støj er en stokastisk variabel, variansen af den observerede variabel, der ikke er forårsaget af den underliggende variabel - altså differencen mellem den observerede variabel og den teoretiske true score: E = X - T. Af denne grund bliver støj også kaldt den "resterende variabel" (eng: residual variable) (Slocum, 2005).

Et teoretisk eksempel på udregning af støj, kunne være hvis vi kender en testdeltagers true score i en IQ-test, det kan være T = 40, og giver personen den samme test tre gange X = 35, 45, 42, er E for de tre test -5, +5, +2. Det er sammenhængen mellem E, T og X.

Det at støj betragtes som tilfældigt betyder, at alt andet varians end den true score, antages at have lige stor sandsynlighed for at forhøje eller formindske den observerede score for hver item (DeVellis, 2006). Den forventede værdi af støj er altså lig nul. Støj for de observerede variabler er indbyrdes uafhængig, så støjen for en variabel vil ikke indvirke på støjen for de andre variabler (DeVellis, 2006).

Dette fører til tre andre antagelser om støj i klassisk testteori:

$$p(E_1E_2) = 0 \qquad (1.2)$$

Støjen E mellem to test (E1 og E2) er ikke korreleret.

$$p(T_1E_2)=0 \quad (1.3)$$

Korrelationen mellem true score for en test 1 er ikke korreleret med støjen for en test 2.

$$P(T_2E_1)=0 \quad (1.4)$$

True score for test 2 er ikke korreleret med støj for test 1. Disse antagelser betyder at støjscorer i klassisk testteori er additive og ikke multiplikative (Irwing et al., 2018).

Den femte og sidste antagelse i klassisk testteori er

$$e(x) = T \quad (1.5)$$

Den forventede score (e(x)) er lig med true score (T). Hvis en testperson bliver testet gentagne gange, vil scoren konvergere mod true score.

Den helt grundlæggende ide med klassisk testteori er at hvis man har en test, hvor alle antagelser er mødt, er man i stand til at udvinde en masse afledninger omkring denne test. De mest essentielle afledninger er:

$$P^2_{xt}= \sigma^2_t/ \sigma^2_x \quad (1.6)$$

Kvadratroden af populationskorrelationen mellem de observerede scores og true scores er lig med variansen af true scores divideret med variansen af de observerede scores. P^2_{xt} kan forstås som reliabiliteten af en test, derfor kan reliabiliteten altså udregnes som variansen af true scores divideret med variansen af de observerede scores. Det vil vi uddybe i teoriafsnittet om reliabilitet.

$$P^2_{xt}=1-(\sigma^2E/\ \sigma^2_x) \qquad (1.7)$$

Kvadratroden af populationskorrelationen mellem de observerde scores og true scores i en test er lig med em minus variansen af støj scores divideret med variansen af de opnåede scores.

$$P^2_{XT}=1-\sigma^2E/\sigma^2_X \qquad (1.8)$$

Kvadratroden af populationskorrelationen mellem de observerede scores og true scores på en test er lig med en minus variansen af testen divideret med variansen af de observerede scores. En tests reliabilitet er lig med 1 minus rationen af støj scores til variansen af den observeredes scores varians.

$$o^2_{tx}=K^2o^2_{ty} \qquad (1.9)$$

Variansen af true scores på test X er kvadratroden af antal ”units” (K) gange variansen af true scores på en kortere test Y. Det betyder at variansen af de observerede scores stiger med K. Det betyder desto længere testen er, jo højere er reliabiliteten.

Der er mange flere afledninger, men de ovennævnte er de mest essentielle for vores opgave.

Reliabilitet

Vi vil i dette afsnit primært fokusere på intern reliabilitet det vi kalder testscore reliabiliteten

Som beskrevet i det foregående teoriafsnit, er en af afledninger fra klassisk testteori, at testscore reliabiliteten $p_{xx'}$ er lig med kvadratkorrelationen mellem scoren X og T og kan skrives med P^2_{XT}.

Testscore reliabiliteten kan også defineres som korrelationen mellem to sæt af uafhængige test-scorer afledt af to versioner af den samme test, det der hedder paralleltest (Nunnally & Bernstein, 1994).

Da det sjældent er muligt at have to sæt af paralleltest scorer, kan kovariansen mellem items, det man også betegner inter-item kovarians, bruges til at måle reliabiliteten (Van der Ark, van der Palm & Sjitsma 2011). Alle items fungerer som paralleltest i klassisk testeori, hvilket betyder at de er udskiftelige med hinanden (DeVellis, 2006).

En måde, der ofte bliver anvendt til at måle reliabilitet via inter-item kovarians, er Cronbach’s Alpha (Hogan, Benjamin & Brezinski, 2000). Cronbach’s Alpha udregnes ved at

korrelere scoren for hver item med total scoren for hver observation og så sammenholde det med variansen for hver individuel item score (Cronbach, 1951).

Cronbach's Alpha (a) bliver udregnet således:

$$a = (k/k\text{-}1)(1-(S^{k}_{i=1} s^2_{yi})/s^2_x) \qquad (2.1)$$

hvor *k* som før er antal items i skalaen, s^2_{yi} er variansen for item *i,* og s^2_x er variansen for den observerede totalscore (Cronbach, 1951).

Cronbach's Alpha er et estimat af en populations interne konsistens, når en række antagelser er opfyldt. De antagelser er: skalaen har essentiel tau-ækvivalens, støjen fra items er uafhængig og skalaen er endimensionel. Tau-ækvivalens betegner det forhold at hvert item bidrager ligeligt til totalskalascoren (Green & Yang, 2009; Sjitsma, 2009). Endimensionalitet refererer til eksistensen af en latent variabel i dataen (Cho & Kim, 2015). Når disse antagelser ikke er mødt, vil anvendelse af Cronbach's Alpha være uberettiget og derfor misvisende (Green & Yang, 2009; Sjitsma, 2009).

Hvis vi udtrykker Alpha baseret på gennemsnittet af den observerede score kovarians for alle items (k), får vi følgende:

$$a = k/(k\text{-}1)((S_i{}'S_j s_{ij})/s^2_x) = k^2 Mean(s_{ij})/s^2_x \qquad (2.2)$$

Alpha ligningen kan skrives sådan da $s^2_x = S_i S_j s_{ij} = S_i s^2_i + S_i{}' S_j s_{ij}$ hvor $s^2_i \, {}^o s^2_{xi}$ og $s_{ij} \, {}^o \, s_{xixj}$.

Hvis items overholder antagelsen om essentiel tau-ækvivalens, det vil sige hvis $s_{xixj} = s^2_t/k^2$ for alle items *i* og *j,* så bliver tælleren i ligningen ($k^2 Mean(s_{ij})$) lig med den sande true score varians (s^2_t). Det vil sige at Alpha er lige med $P_{xx'}$ når der er essentielle tau-ækvivalens (Cho & Kim,, 2015).

Hvis ikke der er tau-ækvivalens, vil Alpha være mindre end den sande reliabilitet ($P_{xx'}$) og derfor være et lower-bound for reliabiliteten altså: $P_{xx'} = a < 1$ (Cho & Kim, 2015; McNeish, 2017).

Hvis item støj korrelerer positivt vil Cronbach's Alpha overestimere reliabiliteten (McNeish, 2017). Grunden til dette er, at variansen af de observerede item scorer (s^2_{xi}) ikke bliver påvirket af korrelationen mellem item støj, men det bliver inter-item kovariansen (s_{xixj}). Inter-item kovariansen ændrer sig fra den normale inter-item kovarians s^2_t/k^2 til s^2_t/k^2+s_{eiej}. Deraf følger det, at hvis item støj korrelerer positivt, vil Alpha stige, men reliabiliteten falde og $a > p_{xx'}$.

Brugen af Cronbach's Alpha som et mål for den interne konsistens har før mødt kritik (Sjitsma, 2009). Den interne konsistens er forholdet mellem items, nærmere bestemt den gennemsnitlige inter-item korrelation (Yang & Green, 2009; Sjitsma, 2009). Cronbach's Alpha bliver, som tidligere nævnt udregnet via kovarians mellem items og er dermed sensitiv overfor den interne konsistens, hvis man definerer denne som den gennemsnitlige inter-item korrelation. Cronbach's Alpha er dog også meget sensibel overfor antal items, hvilket ligningen også viser.

Faktormodellen og faktoranalyse

Som beskrevet tidligere er en vigtig antagelse bag den refleksive variabel model éndimensionalitet. Dimensionalitet refererer til strukturen af et specifikt begreb (Pett, Lackey & Sullivan, 2003) og ved éndimensionalitet forstås, at et begreb/en test kun har en dominant latent variabel eller fænomen. Det vil sige at en skala er éndimensionel når den systematiske forskel i varians i items kun er forårsaget af en varians kilde (Ziegler & Hagemann, 2015).

Vi vil i denne opgave bruge endimensionalitet synonymt med tilstedeværelsen af én faktorstruktur, som andre også før har gjort (Kumar & Dillon, 1987). Vi er dog opmærksomme

på at det ikke er uden kontrovers (Edwards, 2011; Temme & Diamantopoulos, 2016). Spørgsmålet omkring, i hvilket omfang en multifaktoriel struktur kan være endimensionel som følge af en såkaldt "higher-order model" eller hvorvidt en faktorstruktur med en faktor, der er associeret med meget varians og flere faktorer, der er associeret med lidt, kan tolkes som værende endimensionel, ligger uden for omfanget af denne opgave.

Den mest anvendte måde at teste dimensionaliteten af en test på er med faktoranalyse. Det er en analyse der bygger på en teoretisk model, der hedder fællesfaktormodellen.

Fællesfaktormodellen er en direkte udledning klassisk testteori og bygger på det centrale teorem derfra: $X = T + E$ (Crocker & Algina, 1986).

Det essentielle I faktormodellen er at hver persons score på hver observerede variabel bliver komprimeret til en fællesfaktorscore, η vægtet af en faktorloading for det item, λ, plus tilfældig støj ε.

$$y = \lambda\eta + \varepsilon \tag{3.1}$$

Individuelle forskelle η bestemmer fællesvariansen blandt de observerede variable og derfor kan faktormodellen bruges til at separere fællesvariansen (η) (som typisk bliver tilskrevet det teoretiske begreb man vil måle) fra den unikke støj (ε) (det der typisk bliver tilskrevet tilfældig målingsstøj, på samme måde som forklaret i klassisk testteori).

Dette betyder at en testpersons værdi af den ikke-observerbare variabel kan generaliseres til en score, faktorscoren, da det i klassisk testteori antages at alle true scores for de observerede variabler, måler den samme latente variabel og derfor har et perfekt lineært forhold.

Ligesom i klassisk testteori, ligger der flere antagelser bag fællesfaktormodellen (Crocker & Algina, 1986; Nunnally & Bernstein, 1994).

Fællesfaktormodellen antager at faktorerne udviser lineære effekter på de observerede variabler. Fællesfaktormodellen antager at hver observerede variabel, er en vægtet lineær kombination af faktorene, der måles, plus en unik faktor.

Derudover antager fællesfaktormodellen lokal uafhængighed blandt de observerede variabler. Det betyder at de observerede variabler ikke korrelerer, når man korrigerer for faktoren.

Når antagelserne bag fællesfaktormodellen er overholdt, har den store fordele. Da modellen er i stand til at separere begrebs-relateret varians fra fejl-varians, giver modellen mulighed for at udregne ikke-biasede (eng: unbiased) estimater af forholdet mellem begreber Til at undersøge faktorstrukturen bruger man faktoranalyse. Faktoranalyse er en statistisk metode, som bruges til at opsummere og beskrive data. Faktoranalyse fungerer ved at inddele et større sæt observerede variabler i mindre klynger, såkaldte faktorer.

En faktor vil således bestå af variabler, der er korreleret med hinanden, men som er næsten uafhængige af andre klynger af variabler. Ideen bag faktoranalyse er at faktorer viser noget om den underliggende sammenhæng, som fører til de observerede korrelationer mellem variablerne.

Faktoranalyse bygger altså på matrixalgebra og matematikken bag faktoranalyse er noget kompliceret, så vi vil ikke gå mere i detaljer med det her. Det mest essentielle at forstå er at man komprimerer den originale korrelationsmatrice til en egenværdimatrice (eng: eigenvalue matrix).

En egenværdimatrice er en matrice med en vektor for hver item i en given skala. Disse egenværdivektorer kan hver forklare en vis procentdel af den samlede varians og det er disse egenværdier man i faktoranalyse bruger til at vurdere, hvor mange faktorer man skal udvinde fra sine data.

Den nok mest brugte metode til dette er Kaiser-Guttman kriteriet, der siger man skal beholder alle faktorer med egenværdi på over 1. Det betyder at hvis man har en egenværdirmatrice med 5 items, så skal en faktor være associeret med 20 % af variancen for at beholde den (100/5) (Kaiser, 1960).

De faktorer, der bliver udvundet, er defineret ved deres korrelationer med de originale variabler. Disse korrelationer bliver kaldt ”factor loadings” og viser hvilket forhold de forskellige variabler har til de udvundne faktorer (Crocker & Algina, 1986).

Der er to overordnede typer faktoranalyse: udforskende faktoranalyse (eng: exploratory factor analysis) og bekræftende faktoranalyse (eng: confirmatory factor analysis).

Udforskende faktoranalyse bliver brugt til at simplificere et større sæt data, for at finde de vigtigste variabler. Her undersøges hvor meget af variansen fællesfaktorer kan forklare, man har ikke en hypotese og denne analyse bliver ofte brugt til hypotesegenerering som en indledende test (Nunnaly & Bernstein, 1994)

Bekræftende faktoranalyser bruges som navnet antyder bliver denne analyse brugt til at bekræfte eller finde support for en hypotese omkring faktorstrukturen.

En lignende statistisk metode er Principal Component Analysis (PCA), som også har til formål at reducere kompleksiteten af den observerede data. PCA er ligesom faktoranalyse en statistisk teknik man kan anvende på et sæt variabler, når man vil undersøge hvordan variablerne danner klynger, som varierer sammen. PCA producerer ligesom faktoranalyse lineære kombinationer af de observerede variabler, hvor hver kombination er en komponent (Gregory, 2011; Tabaschinck & Fidell, 2014).

I PCA benytter man alt varians til at udvinde komponenter, hvor man i faktoranalyse kun bruger den varians, der er delt af items. Derfor er det en forudsætning for faktoranalyse, at den skala man tester har en latent refleksiv variabel. Hvis ikke dette er tilfældet, bør man bruge PCA (Gregory, 2011).

Faktoranalyse og sumscorer

Faktoranalyse kan blandt bruges til at retfærdiggøre brugen af sumscores fra en skala som proxy for den latente variabel. To forudsætninger for brugen af sumscores er nemlig endimensionalitet og målingsinvarians (Fried, van Borkulo, Epskamp, Schoevers, Tuerlinckx & Borsboom, 2016).

Endimensionalitet har vi beskrevet hvordan man kan undersøge med faktoranalyse. Målingsinvarians henviser til at skalaerne fungerer ens ved de forskellige målinger, ikke hvorvidt alle får samme score i skalaen, men hvorvidt skalaen måler det samme hos alle testpersoner. Faktoranalyse kan bruges til at teste målingsinvarians, hvis skalaerne måler det samme, skal faktorstrukturen være ens. Det vil sige der skal ikke være store udsving i faktorstruktur på tværs af grupper (van de Schoot, Lugtig & Hox, 2012).

Valideringen af depressionsskalaer

De fleste depressionsskalaer valideres ud fra deres korrelation med en given diagnostisk guldstandard. Som guldstandard anvendes oftest et klinisk interview, hvilket kan være struktureret, delvist struktureret eller ustruktureret (Mitchell & Coyne, 2010). Alternativt kan valideringen bestå i en undersøgelse af den konvergerende validitet (eng: convergent validity) med en anden skala, som allerede er valideret. Generelt anses SCID-5 interviewet, udviklet til DSM manualen, samt PSE og PHQ, som de primære guldstandarder (Mitchell & Coyne, 2010). Testværktøjers effektivitet angives ofte ved at angive deres sensitivitet og specificitet, holdt op mod et af disse kliniske interviews, som referencestandard (Genders et al 2012).

Dette vil altså sige, at diagnosen givet til den enkelte deltager, ved det kliniske interview, anses som den sande diagnose og det undersøges hvor præcist skalaen kan indfange dette. Samtidig betyder den varierende brug af kliniske interviews, at forskellige studier validerer testværktøjer ud fra forskellige referencestandarder.

Mitchell & Coyne (2010) angiver at det er vigtigt at skelne mellem centrale begreber som ofte bruges synonymt i litteraturen. Særligt vigtigt i denne sammenhæng, er begrebet screening, som indebærer at der ikke søges at fastsætte en diagnose, men at bestemme kandidater til en videre udredning ved et diagnostisk interview. De fleste depressionsskalaer udvikles med det formål at fungere som screeningsværktøj og ikke som diagnoseredskab (Santor et al., 2006; Mitchell & Coyne, 2010).

Centrale begreber i testning

To helt centrale begreber inden for testvalidering er testens sensitivitet (andelen af syge som tester positive) og specificiteten (andelen af ikke-syge som tester negativt). Disse er vigtige i

testsammenhæng fordi disse har afgørende betydning for hvad, der kan udledes at et testresultat. Specifikt kan både sensitiviteten og specificiteten udtrykkes som konditionelle sandsynligheder, hvilket tydeliggør hvad der er muligt at inferere ved et testresultat (Lalkhen & McCluskey, 2008).

Sensitiviteten kan udtrykkes som den konditionelle sandsynlighed for at teste positiv givet at man har sygdommen. Det er defineret som følger, hvor alle indgående faktorer er udtrykt i absolutte tal:

$$Sensitivitet = \frac{Ægte\ positive}{Prævalens} = \frac{Ægte\ positive}{Ægte\ positive + Falsk\ negative}$$
$$= 1 - Falsk\ negativ\ rate$$

På samme vis kan specificiteten udtrykkes:

$$Specificitet = \frac{Ægte\ negative}{1 - Prævalens} = \frac{Ægte\ negative}{Ægte\ negative + Falsk\ positive}$$
$$= 1 - Falsk\ positiv\ rate$$

Det følger af disse definitioner at både sensitiviteten og specificiteten kan optimeres uden at bruge en test, ved hhv. at definere alle testtagere som positive og negative. Derfor er udfordringen i praksis at finde et kompromis mellem sensitiviteten og specificiteten. Dette kompromis fastlægges for depressionsskalaer oftest med Youden's index, som kombinerer disse to og dermed udgør et mål for præcisionen af en test. Youden's index er defineret som følger, hvoraf det ses at værdien vil svinge mellem 0 og 1 (Youden, 1950):

$$Youden's\ index\ =\ sensitivitet\ +\ specificitet\ -1$$

Dette kan også skrives som:

$$\frac{Ægte\ positive}{Ægte\ positive\ +\ Falsk\ negative} + \frac{Ægte\ negative}{Ægte\ negative\ +\ Falsk\ positive} - 1$$

Youden's index kan også ses som det optimale punkt i en Receiver Operating Characteristic kurve (ROC). En ROC-kurve er et plot for alle ægte positiv og falsk positive rater ved forskellige cut-offs (Bewick, Cheek & Ball, 2004).

Depressionsskalaer er udviklet til anvendelse som screeningsværktøjer, og de er derfor oftest udformet med sensitivitet som prioritet (Mitchell & Coyne, 2010). Det primære rationale er altså at minimere antallet af falske negative. Raten af falske negative hænger direkte sammen med sensitivitet, i det de er komplementære værdier. Dette ses ud fra definitionen:

$$Falsk\ negativ\ rate\ =\ falsk\ negativ\ /(ægte\ positiv\ +\ falsk\ negativ)$$
$$= 1 - Sensitivitet$$

På samme måde er sensitiviteten og raten af falsk positive, komplementære værdier, som det ses ud fra definitionen:

$$Falsk\ positiv\ rate\ = 1 - Specificitet$$

$$= falsk\ positiv\ /\ (falsk\ positiv\ +\ ægte\ negativ)$$

Positiv prædiktiv værdi (PPV) betegner den konditionelle sandsynlighed for at være syg givet at testen er positiv.

$$Positiv\ prædiktiv\ værdi\ = Ægte\ positiv\ /\ (Ægte\ positiv\ +\ Falsk\ positiv)$$

Negativ prædiktiv værdi (NPV) betegner den konditionelle sandsynlighed for at være rask givet at testen er negativ.

$$Negativ\ prædiktiv\ værdi\ = Ægte\ negativ\ /\ (Ægte\ negativ\ +\ falsk\ negativ)$$

Sensitivitet og specificitet, kan sammen med sample size og prævalens bruges til at lave en såkaldt confusion matrix, som kan eksemplificere og illustrere de ovenstående ligninger.

For en hypotetisk depressionsskala med en sensitivitet på 0,9 og specificitet på 0,7, som anvendes på en population af 100 mennesker, hvor 10 % har depression, får vi følgende confusion matrix:

					N	100
					Prævalens	0,1
					Sens.	0,9
					spec.	0,7
		Ægte				
		Positiv	Negativ	Sum	PPV	NPV
FALSK	Positiv	9	27	36	0,25	0,984375
	Negativ	1	63	64		
	Sum	10	90			

[figur 1: confusion matrix for hypotetisk depressionsskala]

Ud fra denne confusion matrix kan de ovenstående forhold udledes, herunder eksempelvis forholdet mellem PPV, NPV, sample size (N), prævalens, sensitivitet og specificitet.

Optimering af cut-off

Der er ikke mange meta-analyser af optimale cut-off værdier for depressionsskalaer. De analyser der er udført, benytter sig af bivariate modeller, som er random effects model, der fokuserer på en fællesnormalfordeling af logit transformeret sensitivitet og specificitet (Gilbody, Richards, Brealey & Hewitt, 2007; Manea, Gilbody & McMillan, 2012; Vilagut et al, 2016; Steinhauser, Schumacher & Rücker, 2016).

Logit er logaritmen af $p/(1-p)$, hvor er p er sandsynligheden og kaldes også for log-odds (Wooldridge, 2016). Den bivariate model har to niveauer og samler sensitivitet og specificitet. Ægte positive og falske negative fra et positivt resultat er antaget at være uafhængige og at have bionomiale distributioner. Der bliver taget hensyn til variansen mellem studierne via random effekter.

I disse standard modeller er præmissen altså at hvert studie kun bidrager med et sæt sensitivitet og specificitet (Reitsman, Glas, Rutjes, Scholten, Bossuyt & Zwinderman, 2005).

Det fører til et problem da SROC (Summary Receiver Operator Curve) ikke er entydigt defineret og der er mange måder at definere en lige linje i et logit koordinatsystem. Derudover vil SROC måske bliver overestimeret da rapportering af optimale cut-off værdier for de enkelte studier, opstår en risiko for at dette vil føre til en bias i estimatet af både sensitivitet og specificitet for de enkelte cut-off niveauer. Riley et al. (2015) har undersøgt denne effekt og påviser ved simulationer, at meta-analyse estimater afledt af inkomplet data, er betydeligt oppustede (eng: inflated). De konkluderer, at ikke-biasede estimater forudsætter enten at have al data tilgængelig fra de enkelte studier eller at korrigere statistisk for manglende data.

Kliniske interviews

Som nævnt tidligere bliver skalaerne typisk valideret op imod et klinisk interview. Vi vil i dette speciale primært fokusere på to interviews, Structural Clinical Interview for DSM (SCID) og Composite International Diagnostic Interview (CIDI)

SCID er et semistruktureret interview, der hjælper testlederen med at teste forskellige diagnosticeringshypoteser (Spitzer, Williams, Gibbon & First, 1992). Interviewet fungerer ud fra en tilgang baseret på et beslutningstræ, hvor de enkelte svar fra patienten afgør i hvilken retning testlederen skal gå videre. SCID indeholder 9 moduler hvor 7 af dem repræsenterer major axis 1 diagnosticeringsklasserne.

Som nævnt tidligere, i den historiske gennemgang, ændrede DSM-III markant på konceptualiseringen af flere mentale lidelser, hvilket betød at flere af de

diagnosticeringsinterviews der allerede eksisterede blev forældede. DSM-III inkluderede diagnosekriterier for stort set alle mentale lidelser, hvilket muliggjorde udviklingen af SCID. SCID blev derfor oprindeligt udviklet til at diagnosticere for DSM-III, men siden er der udviklet en SCID-I for DSM-IV (Allen, 1998) og en SCID-5 for DSM-V (First, Williams, Karg & Spitzer, 2016). SCID er blevet brugt i flere tusinde forskningsstudier siden sin introduktion i 1986 (Mitchell & Coyne, 2010).

Den semistrukturerede struktur betyder, at testlederen har mulighed for at spørge ind til patientens svar, hvor mange af spørgsmålene i SCID ellers kan besvares ved simple bekræftende eller afkræftende svar. Dette gør testlederen i stand til at stille opfølgende spørgsmål og bede patienten komme med eksempler på hvordan et symptom manifesterer sig (First et al., 2016).

CIDI blev udviklet på anbefaling af WHO og inkluderer spørgsmål fra interviewskalaerne Diagnostic Interview Schedule (DIS) og Present State Examination (PSE). CIDI er et struktureret interview, hvilket gør at CIDI også kan bruges af lægmænd der ikke har nogen viden om psykopatologi (Robins et al., 1988). CIDI er designet til at udregne prævalenser af mentale lidelser, på tværs af kulturer, i epidemiologiske studier og bliver ofte brugt i disse sammenhænge (Wittchen, 1994).

Beslutningstagning - heuristikker og bias

Inden for feltet beslutningstagning er emnet heuristikker og bias det mest undersøgte og sættes særligt i forbindelse med Kahneman og Tversky (Leonhardt, 2016). Kahneman (2011) har senere defineret begrebet heuristik som en simpel procedure, der leder til tilstrækkelige, men uperfekte svar. Dette gør at heuristikker kan være effektive og fordelagtige, men samtidig også leder til systematiske og forudsigelige fejl, hvilke benævnes som bias (Kahneman & Tversky,

1974). En række heuristikker og bias er beskrevet i litteraturen, her præsenterer vi et lille udvalg af særlig relevans for dette projekt.

Tversky & Kahneman (1974) beskriver at repræsentativitetsheuristikken gør sig gældende ved beslutninger, der inkluderer kategoriseringer. Indenfor den kliniske psykologi, kunne dette være, hvorvidt et individ har en psykisk lidelse eller ej, eller alternativt hvorvidt individet har lidelse a eller b. Heuristikken beskriver, at agenten vil placere individet i den gruppe, hvor individet bedst repræsenterer agentens billede af denne kategori.

Kahneman & Tversky's oprindelig eksempel lyder :"Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality [...]" (s. 1124, Tversky & Kahneman, 1974).

Når en person skal vurdere Steves mest sandsynlige erhverv, så som bibliotekar, læge eller landmand, vil dette bero på i hvor høj grad Steve passer på de stereotyper.

Risikoen ved denne heuristik er at den kan føre til basis-sats neglekt (Fiedler & von Sydow, 2015). Dette fordi man i vurderingen af hvor godt et nyt tilfælde matcher tidligere tilfælde, glemmer hvor ofte denne type tilfælde forekommer statistisk. Her viser både Kahneman & Tversky og senere forskning, at yderligere viden kan forringe folks evner til at forudsige og bedømme sandsynligheder (Tversky & Kahneman, 1974; Fielder & von Sydow, 2015).

Tilgængelighedsheuristikken (eng: availability heuristic) beskriver den proces at vurdere hyppighed ud fra "den lethed, hvormed man kommer i tanke om eksempler" (s. 156, Kahneman, 2014).

Forankringseffekt (eng: anchoring) er en bias der optræder når folk overvejer en bestemt værdi for en ukendt kvantitet, før de estimerer denne kvantitet. Det endelige estimat vil være

tættere på det tal, folk først overvejede (Kahneman, 2011). I et dagligdagsperspektiv er dette også kendt i forbindelse med førstehåndsindtryk af andre personer. Kahneman (2011) angiver videre at forankringseffekten er en af de mest pålidelige og solide resultater fra eksperimentalpsykologien.

Det er understreget mange steder, at brugen af heuristikker ikke udelukkende er negativ, men at udfaldet afhænger af den konkrete kontekst (Lilienfeld, Ritschel, Lynn, Cautin & Latzman, 2014; Kahneman & Klein, 2009). Eksempelvis vil den simple heuristik, at gå ud fra hvor lidende patienten virker udadtil, ganske sandsynligt være god i mange tilfælde. Omvendt kan dette være et problem i mødet med en patient, der undertrykker sine følelser overfor klinikeren.

Metode

Søgestrategi

For at kunne belyse specialets problemformulering har vi søgt efter relevant litteratur i søgemaskinen Summon (SDU) samt databaserne PsycINFO og PubMed. Søgningsprocessen bestod af to systematiske søgninger, og flere kædesøgninger. I kædesøgningerne blev litteraturen fundet gennem udvalgte studiers referencer. Vores første hovedsøgning bestod af søgematricen ”Major Depression AND Screening AND (Test Validity OR Test Reliability)”. Denne søgning gav 197 resultater på PsycINFO og 1178 resultater på PubMed, hvilket førte til en ny søgematrice ”Major Depression AND Screening AND (Test Validity OR Test Reliability) AND Meta-Anlysis”, hvilket gav 6 resultater på PsycINFO og 22 resultater på PubMed. Vi endte dog

med ikke at bruge mange af disse studier, men via kædesøgninger på en af disse artikler, kom vi frem til Yang & Gorestein (2013) studiet, der ligger til grund for dele af vores analyse.

Den anden hovedsøgning rettede sig mod screeningsdelen af opgaven og bestod af søgematricen ”Diagnostic Accuracy AND Major Depression AND Meta-Analysis”. Denne søgning gav 23 resultater på PsycINFO og 64 på PubMED. Det var via denne søgning vi fandt Vilagut, Forero, Barbaglia & Alonso (2016) studiet, der ligger til grund for store dele af vores analyse. Studiet bliver ligesom alle andre studier, der er inkluderet i specialet udvalgt ud fra en manuel sortering af abstracts.

En betydelig del af litteraturen er, som beskrevet, kommet fra kædesøgning, hvilket kan medføre bias i resultaterne.

Regressionsmodeller

Til at forstå analysedelen er det nødvendigt at forstå matematikken bag.

Regressionsanalyse er en statistisk metode, der kan bruges til at bestemme forholdet mellem variabler i et sample, og derfor estimere forholdet mellem variabler i populationen (Wooldridge, 2016).

Den mest basale lineære regressionsmodel er bivariat og kaldes simpel lineær regression (Keller & Gaciu, 2015). Simpel lineær regression har den følgende formel:

$$y_i = a + bx_i + e_i \qquad (4.1)$$

Her er y værdien altså en funktion af x værdien, som skaleres med faktoren b. Parameteren a angiver skæringspunktet og e angiver fejl (error term), og er differencen mellem den enkelte y værdi og den estimerede y værdi som regressionsmodellen bestemmer.

Anvendelsen af simpel regressionsanalyse forudsætter at en række antagelser gør sig gældende for data. Dette vil vi ikke gå ind i her, men den vigtigste antagelse er, at der er et lineært forhold mellem variablerne (Wooldridge, 2016)

I analysen anvender vi R pakken (Rstudio, 2015) Diagmeta, som anvender multipel regression. Multipel regression betegner regression hvor der indgår mere end en uafhængig variabel

$$y_i = a + bx_i + c_i + e_i \quad (4.2)$$

Når en regressionsmodel har flere uafhængige variabler og gør brug af såkaldt random effects, betegnes dette oftest som en mixed model (Woolridge, 2016). Ved vores anvendelse af Diagmeta er der netop tale om en sådan såkaldt blandet model (eng: mixed model).

Eksempel:

Hvis man tager *m* store skoler tilfældigt fra flere tusind, og hvis *n* elever på samme alder er valgt tilfældigt fra hver skole. Vi kender deres score på en test. Lad Y_{ij} være scoren for elev *j*, på skole *i*. Så vil man kunne lave en model af forholdene for disse mængder således:

$$Y_{ij} = \mu + U_i + W_{ij}, \quad (4.3)$$

Hvor μ er den gennemsnitlige score for hele populationen, U_i er den skole-specifikke random effect, som den måler forskellen på den gennemsnitlige score for skole *i* og den

gennemsnitlige score for hele landet. W_{ij} er den individuelle-specifikke random effekt, afvigelsen fra *j* elevens score fra gennemsnittet for *i* skolen.

Modellen kan blive endnu mere forklarende ved at tilføje flere forklarende variabler. Det vil kunne hjælpe med at forklare forskellene i score mellem grupperne. Hvis man tilføjer køn og socioøkonomisk status (SØS), kan man skrive:

$$Y_{ij} = \mu + \beta_1 køn_{ij} + \beta_2 søs_{ij} + U_i + W_{ij} \qquad (4.4)$$

Hvor $køn_i$ er en dummy variabel for køn og $søs_{ij}$ er en dummy variabel for forældrenes socioøkonomiske status. Det der gør dette til en mixed model er de to dummy variabler, også kaldet fixed-effects.

En regressionsmodel, der udelukkende bruger random effeks kan kun bruges, hvis de specifikke effekter er ukorrelerede med de andre kovariater i modellen. U_i må altså ikke korrelere med de andre parametre i modellen. Hvis U_i korrelerer med de andre parametre, bliver random effects biased og det er vil være bedre at benytte sig af fixed effects.

En fordel ved random effects er, at det kan kontrollere for heterogenitet i data, der ikke er observeret eller korreleret med den uafhængige variabel (Woolridge, 2016).

Diagmeta

Som skrevet benytter vi os af R-pakken Diagmeta til vores analyse. Diagmeta pakken er baseret på en ny tilgang til meta-analyse lavet af Steinhauser, Schumacher og Rücker (2016). Denne model har tidligere været anvendt til at undersøge optimalt cut-off af biomarkører i medicin (Steinhauser, Schumacher & Rücker, 2016).

Det grundlæggende rationale bag Diagmeta er at sammenlægge data på tværs af studierne, til to overordnede score-populationer. Den ene population består af de syge, hvor sensitiviteten optimeres og den anden består af de ikke-syge, hvor specificiteten optimeres. Diagmeta tilgangen inkluderer på denne vis data for samtlige cut-off værdier angivet i de enkelte studier og udnytter derfor den totale mængde tilgængelig data. Det står i modsætning til en tilgang hvor data lægges sammen, med ét datapunkt for hvert studie, og hermed undgås risikoen for at estimatet bliver for optimistisk grundet selektiv rapportering af optimale cut-offs (Riley et al. 2015).

Modellen anvender lineær regression til bestemmelse af parameterværdierne. Data for sensitivitet og specificitet transformeres derfor med en lineær funktion. Vi følger her Steinhauser et al. (2016) og betegner for begge tilfælde denne funktion med *h*. Funktionen er for henholdsvis sensitivitet og specificitet givet ved:

$$h\big(Sp(x)\big) = \frac{x-\mu_0}{\sigma_0} \qquad (5.1)$$

$$h\big(1 - Se(x)\big) = \frac{x-\mu_1}{\sigma_1} \qquad (5.2)$$

Koefficienten x, som samtidig er inputtet til den indre funktion, betegner her en cut-off værdi. Differencen mellem et givent cut-off og den gennemsnitlige score for populationen

(henholdsvis syg og ikke-syg) divideres altså med standardafvigelsen for samme population. Bemærk her at transformationen gøres på specificitetens komplementære værdi.

Diagmeta anvender mixed models regression og inkluderer otte overordnede statistiske modeller. Den primære forskel på disse er antallet af random effects. Den største model har både skæringspunktet og hældning som random effects, mens den mindste kun har et fælles skæringspunkt som random effect. Forskellen består altså i graden af differentiering grupperne imellem. Det følger af den underliggende funktion i R, LMER (Linear Mixed-Effects Models), at de større modeller kræver mere data (RStudio, 2015). I vores analyse bruger vi modellen different random intercepts and different random slopes (DIDS). Denne model er givet ved:

$$h\left(\frac{TN_{si}}{N_{0s}}\right) = \alpha_0 + a_{0s} + (\beta_0 + b_{0s})x_{si} + e_{si} \quad (5.3)$$

$$h\left(\frac{FN_{si}}{N_{1s}}\right) = \alpha_1 + a_{1s} + (\beta_1 + b_{1s})x_{si} + f_{si} \quad (5.4)$$

Alfa betegner her de fixede skæringspunkts-værdier. Beta er de fixede hældnings-værdier. Indeks 1 betegner den syge population og 0 den ikke-syge population. X_{si} betegner cut-off værdien i for studiet s. Parameteren *a* er tilfældig skæring og *b* er tilfældig hældning.

Statistiske analyser

Alle statistiske analyser og simulationer er lavet i enten Excel eller Rstudio. Diagmetamodellen og faktoranalysesimulationerne er lavet i Rstudio (Rstudio, 2015).

Analyse

Introduktion til analysen

Som beskrevet i teoriafsnittet, er der flere problemer med bivariate modeller, derfor står det klart, at der er betydelige faldgruber ved estimering på meta-analytisk niveau og at der er brug for nye metodiske tilgange.

Vi vil i analysen undersøge en række forhold i valideringsstudier for CES-D skalaen. Vi anvender det samme udvalg af studier som Vilagut et al. (2016), som har lavet den nyeste meta-analyse for skalaen. Totalt indgår 29 studier (n = 15116). I nogle delanalyser anvender vi samtlige 29 og i andre vil kun et udvalg indgå. Det største studie har 2008 deltagere, og det mindste har 34 deltagere. Det gennemsnitlige antal deltagere er 367. Bilag 1 viser den samlede liste over studier, samt nøgletal for disse.

Inklusionskriterierne for studierne i Vilagut et al. (2016) er: 1) studiet er et valideringsstudie og rapporterer nøjagtigheden (eng: accuracy) af CES-D til diagnosticering af depression, 2) samplet består af den generelle population eller en sampling fra det primære sundhedsvæsen (eng: primary care), 3) diagnosticeringen er udført med standardiserede diagnostiske interviews, baseret på enten ICD (CIDI) eller DSM (SCID), 4) studiets sprog er enten engelsk eller spansk (Vilagut et al., 2016).

Et par forhold ved datagrundlaget er dog anderledes end Vilagut et al. (2016). Vi mener Vilagut et al. (2016) har lavet en fejl i data for studiet Zich, Attkisson & Greenfield (1990) (nr. 28 på bilag 1). Studiet inkluderer to cut-off værdier, og Vilagut et al. (2016) har byttet rundt på data for disse to. Dette er dog studiet med det mindste deltagerantal og indvirkningen heraf bør derfor være minimal.

At vi anvender samme data som Vilagut et al. (2016), men en anden statistisk tilgang, er særligt af betydning for analysens anden del, hvor vi estimerer et optimalt cut-off for CES-D skalaen. Det at vi analyserer de samme studier muliggør en sammenligning med deres resultater, fordi vi kan isolere betydningen af den metodiske tilgang.

Vi vil i første del af analysen se på indvirkningen af prævalensen i de undersøgte samples, og vise betydningen af, hvordan det optimale cut-off defineres. I anden del vil vi beregne de optimale cut-off værdier for skalaen. De optimale cut-offs udregnes på tværs af alle studier, der indgår i vores beregninger. I tredje del ser vi på interaktionen mellem hvilket klinisk interview, der er anvendt i de enkelte studier og hvilket cut-off, der findes som optimal værdi, det vil sige det kliniske interview ses som en moderende variabel. I fjerde del af analysen ser vi på resultaterne og konklusionerne af valideringsstudier med fokus på hvorvidt de tager højde for de strenge antagelser bag de analyser, de bruger.

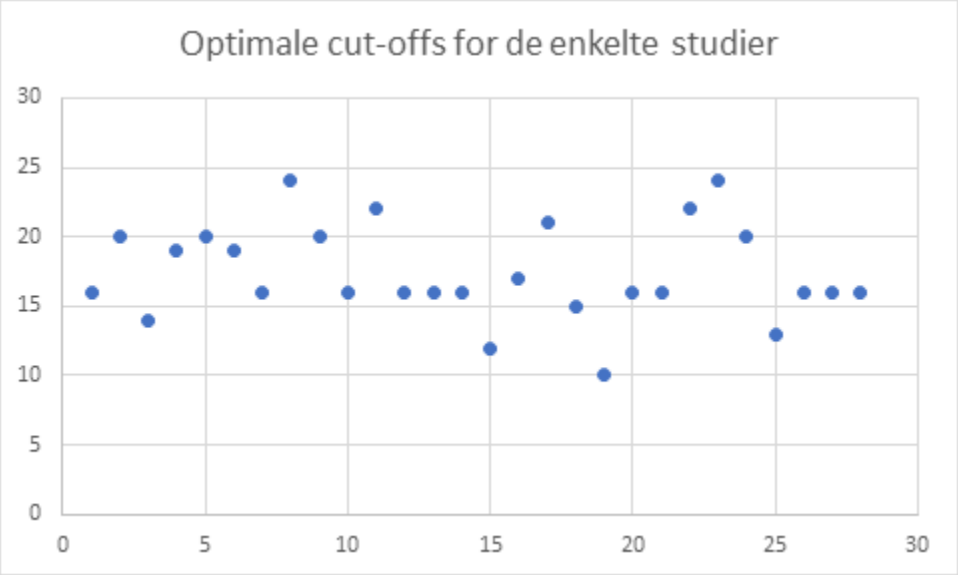
Vi anser ikke resultatet som en definitiv bestemmelse af det optimale cut-off og heller ikke som et midlertidigt estimat af samme. Begrænsningerne i resultatet kan inddeles i to grupper, som henholdsvis begrænsninger i datamaterialet og i metodiske antagelser. Datamaterialet udgøres, som beskrevet, af det litteraturudsnit som Vilagut et al. (2016) anvender. Der kan være tilkommet nyere studier siden, hvilket vi ikke har forholdt os til her. Dertil har vi helt undladt at se på kvaliteten af disse studier og kan derfor heller ikke udtale os herom.

Prævalensanalyse

Denne analyse undersøger sammenhængen mellem prævalensen af depression i de enkelte studier og det optimale cut-off for studiets sample (se teoriafsnit). Vi har her inkluderet

samtlige 28 studier. Der er en markant variation i prævalensen af depression i studiernes samples. Den laveste prævalens er 1,6% og den højeste er 47%.

Flertallet af studierne baserer deres konklusioner på enten Youden's index eller på Receiver Operating Characteristic (ROC) kurver. Dette er de facto det samme, fordi bedste punkt på ROC kurven optimerer summen af de samme to parametre, sensitiviteten og specificiteten, som Youden's index (se evt. teoriafsnittet). For at sikre ensretning, laver vi en gennemgang af de optimale cut-offs baseret på Youden's index. Her finder vi at mindste værdi for optimalt cut-off er 10 og højeste værdi er 24 (se bilag 1 for data). Studiernes fordeling på optimalt cut-off kan ses på figur 2

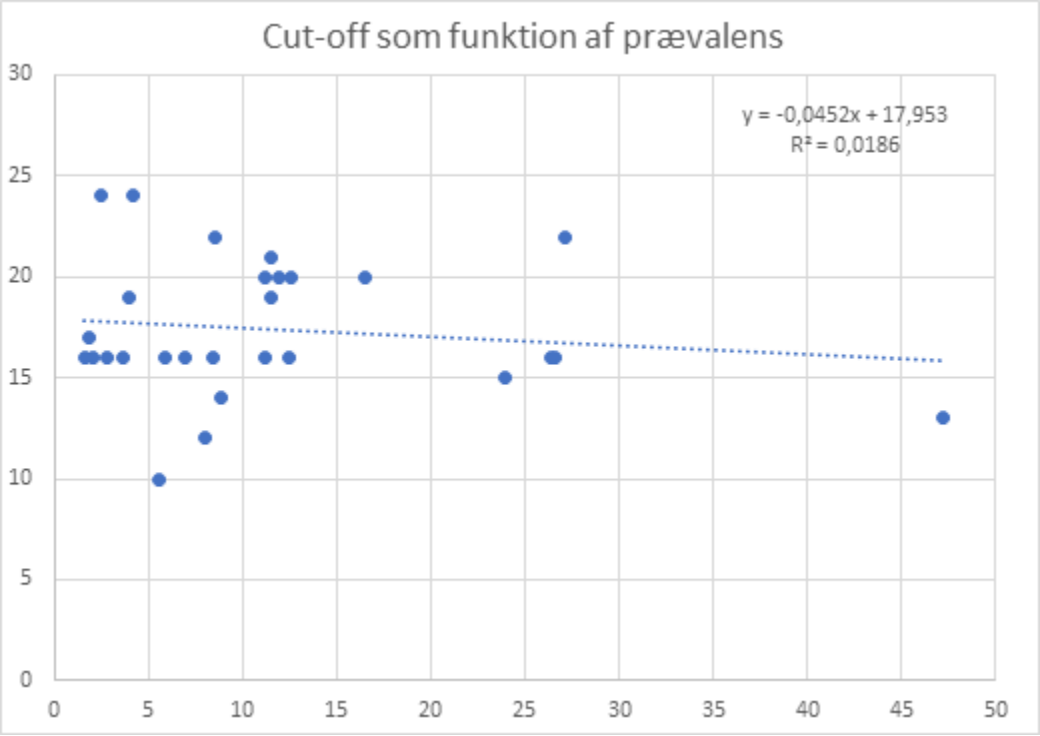


[figur 2: x = studienummer, y=optimum cut-off ud fra Youden's index]

Den store variation i prævalensen i studierne (spænd: 1,6-47%), kan forventes at forårsage en markant heterogenitet i estimeringen af optimalt cut-off studierne imellem. Dette

fordi det, alt andet lige, vil være vigtigere at have en høj sensitivitet ved en højere prævalens, ud fra et nytteetisk perspektiv, som er typisk inden for sundhedsøkonomi (Pedersen, 2013).

Vi har derfor videre undersøgt indvirkningen af prævalensen i studierne på bestemmelse af optimalt cut-off. Dette har vi gjort med en lineær regressionsanalyse, med optimalt cut-off som den afhængige variabel, og prævalensniveau som uafhængig variabel. Figur 3 viser et spredningsdiagram over sammenhængen mellem prævalens og optimalt cut-off.



[figur 3: x =prævalens, y = optimum cut-off ud fra Youden's index]

Der ses et markant fravær af korrelation mellem prævalens i de undersøgte samples og det resulterende optimale cut-off. Dette kan ses uden en regressionanalyse, men udtrykkes kvantitativt ved at værdien af den kvadrerede korrelationskoefficient er 0,0186, hvilket ofte fortolkes som at den forklarede varians er 1,86% (Wooldridge, 2016).

En mulig forklaring på dette kan findes i det anvendte optimeringskriterium, Youden's index. Flere forhold omkring betydningen af prævalensen i det undersøgte sample, kan udledes direkte fra definitionen af Youden's index. Fordi sensitiviteten og specificiteten vægtes ligeligt og denne ligevægtsfaktor er invariant, vil vægtningen af de enkelte tilfælde i hver kategori ændre sig med ændringer i forekomsten af depression i samplet. Definition er, som også angivet i teoriafsnittet:

$$Youden's\ index\ =\ sensitivitet\ +\ specificitet\ -1 \quad (6.1)$$

Det kan vises at definitionen medfører at Youden's index også kan skrives som:

$$Youden's\ index = (1-p)\ *Ægte\ positive_c\ +\ p*Ægte\ positive_c \quad (6.2)$$

Her er p igen prævalensen og $Ægte\ positive_c$betegner det absolutte antal ægte positive ved cut-off værdien *c*. Det mest interessante her, er betydningen af dette og ikke logikken i omskrivningen. Fordi vægtningen af ægte positive versus ægte negative varierer med prævalensen, vil vægtningen kun være ens for ét niveau af prævalens, nemlig 50%.

De to studier med de mest ekstreme niveauer af prævalens vil kunne demonstrere denne effekt, ved at vise den maksimale forskel i vægtning. Studiet med den mindste prævalens, på 1,6% vægter ægte positive med 0,984 (98,4%) og ægte negative med 0,016 (1,6%). Det giver en ratio på 61,5:1 for vægtningen af ægte positive til ægte negative. Ved en ændring i cut-off skal antallet af falsk positive altså mindskes med 61,5 tilfælde for at modsvare en marginaleffekt på én ekstra falsk negativ. For studiet med den højeste prævalens (47,27%), vil vægtningen være

0,5273 (52,73%) for ægte positive og 0,473 (47,27%) for ægte negative. Det udgør dermed et eksempel på at ægte positive og ægte negative vægtes næsten ligeligt og er det eneste tilfælde hvor dette er gældende.

I modsætning til hvad vi har beskrevet ovenfor, hvor vægtning på individniveau varierer med prævalensen, kan man i stedet undgå denne tilfældighed, ved at holde betydningen af de enkelte diagnosticeringer konstante. Dette følger logikken at vigtigheden af at diagnosticere den enkelte rigtigt, er uafhængig af prævalensen i et aktuelt sample (Lalkhen & McCluskey, 2008).

Nytten kan både opgøres i en objektiv enhed, som monetær værdi, og en mere subjektiv enhed, som eksempelvis livskvalitet (Pedersen, 2013). Af pragmatiske årsager, vil vi her anvende den første af de to, i form af estimerede omkostninger. Med dette valg på plads, kan vi nu beregne værdien.

Formlen vi anvender her er:

$$EU_c = TP_c * U_{TP} + FN_c * U_{FN} + FP_c * U_{FP} + TN_c * U_{TN} \qquad (6.3)$$

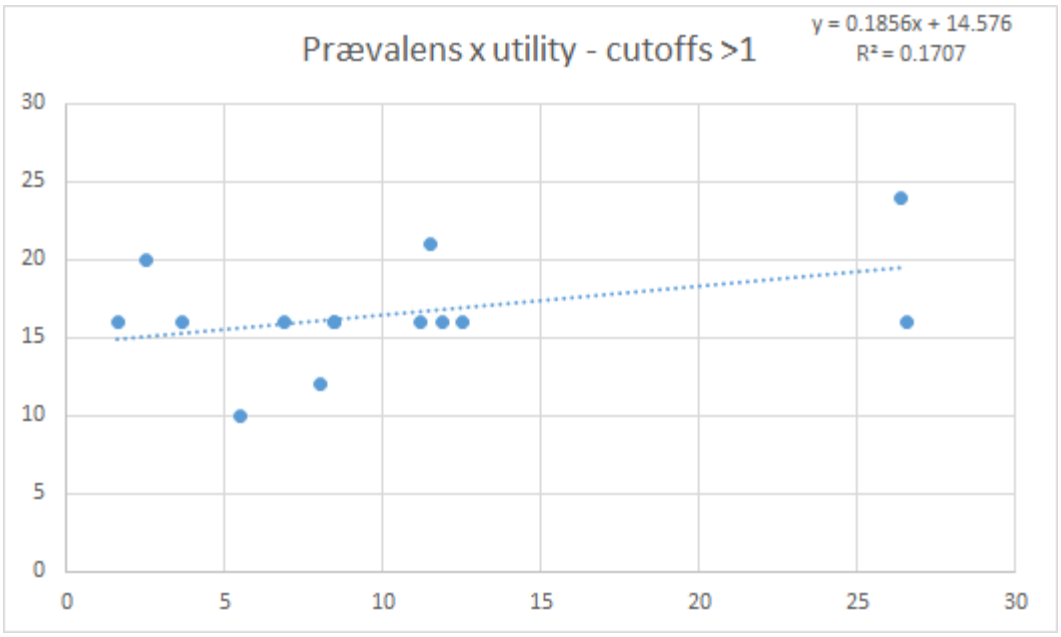
I stedet for at optimere Youden's index, definerer vi her optimum, som det cut-off for hvilket værdien af denne ligning maksimeres. Da vi angiver nytten i omkostninger, ændres fortegnet imidlertid, og nytten maksimeres ved at minimere værdien af ligningen. Symbolerne er her delvist de samme som tidligere, i det *FP, TN, TP, FN* angiver de fire udfald fra confusion matrix'en, henholdsvis ægte positiv, falsk negativ, falsk positiv, ægte negativ. Sænket skrift c henviser til talværdien for et specifikt cut-off og U angiver nytten (engelsk: utility). Eksempelvis er U_{TP} nytten for et individ der, i denne kontekst, er depressiv og hvor testresultatet udgør en ægte positiv.

Hakkaart-Van Roijen et al. (2006), har tidligere estimeret omkostningerne for testning og behandling af depression. Nytten fastsætter vi her ud fra deres angivelser. Vi angiver her blot værdierne uden at korrigere disse til en dansk kontekst. Værdierne, angivet som omkostninger for samfundet i Euro, i de 4 kategorier, er som følger: Ægte negativ = 0, Falsk positiv = 124, Ægte positiv = 9635, Falsk negativ = 11309.

Det er vigtigt at bemærke at målet med disse beregninger, udført på baggrund af hollandsk data, ikke er at bestemme et optimalt cut-off, men at illustrere betydningen af metodiske forhold, nærmere bestemt at cut-off kan optimeres på flere måder. Af denne grund begrænser vi også redegørelsen for de anvendte omkostningsestimater, og henviser her til Hakkaart-Van Roijen et al. (2006). En række overvejelser om begrænsninger ved disse antagelser og muligheden for at bruge denne metodik i praksis, gennemgås senere.

Vi beregner med ovenstående ligning 6.3, det optimale cut-off for samtlige studier hvor vi har data tilgængeligt for mere end et cut-off niveau. Dette fordi det ikke lader sig gøre at optimere cut-off værdien for studier hvor der ikke er rapporteret data for mere end ét cut-off niveau.

Når vi anvender denne fremgangsmåde, finder vi sammenhængen, som vist på figur 4:



[figur 4: x = prævalensniveau, y = optimalt cut-off ud fra nytteværdi optimum]

Med denne definition af optimalt cut-off finder vi en positiv sammenhæng. R-kvadreret går fra 0,01 til 0,17. For dette udsnit af studier kan 17% af variationen i optimale cut-offs altså forklares ud fra prævalensen i de enkelte studier (Wooldridge, 2016).

Vi ser ingen grund til at det undersøgte forhold skulle afvige systematisk fra den brede population i studierne inkluderet her. Ligeledes ser vi ingen grund til at antage, at der skulle være en (anderledes) sammenhæng for andre skalaer end CES-D. Hvis de inkluderede studier er repræsentative, peger dette derfor på at prævalensniveau har betydning for bestemmelsen af optimalt cut-off, ikke kun i teorien, men også i praksis.

Diskussion af prævalensanalysen

I denne del af analysen demonstrerede vi, at der ikke findes nogen sammenhæng mellem prævalensniveau og optimalt cut-off, når dette defineres ud fra Youden's index. Videre

gennemgik vi teoretisk hvorfor dette gør sig gældende. Dette er interessant fordi det viser en helt central begrænsning ved Youden's index som optimeringskriterium.

Det grundlæggende problem er, at den centrale begrænsning i Youden's index i praksis fører til en arbitrær vægtning af sensitivitet og specificitet (Smits et al, 2007). Det har som følge, at resultaterne, typisk rapporteret med henvisning til en ROC-kurve, ikke kan læses uafhængigt af et studies sample. Videre kan dette element af tilfældighed forplante sig i systematiske reviews og meta-analyse estimater hvis disse udføres ved en ukorrigeret sammenlægning af de enkelte studiers fund.

Denne demonstration leder til overvejelser om berettigelsen af Youden's index som optimerings-kriterium. Youden's anvendes bredt i litteraturen (Vouloumanou, Plessa, Karageorgopoulos, Mantadakis & Falagas., 2011) og vi har sjældent set omtale af begrænsninger herved. Vi kan konstatere at prævalensen er stærkt varierende i vores udvalg af studier og ser ingen grund til at dette ikke skulle være repræsentativt for den bredere population af valideringsstudier. Denne variation i prævalensen gør at resultaterne i forskellige studier baserer sig på forskellige antagelser om værdien af sensitivitet og specificitet. Dette ser ikke ud til at være et bevidst valg, idet der ikke argumenteres for at lade vægtning afhænge af prævalensen. Dette betyder at sammenligningsgrundlaget eroderes af en ubegrundet, varierende og arbitrær vægtning af sensitivitet og specificitet.

Videre undersøgte vi samme forhold ud fra nytteværdi, defineret ved omkostninger, og fandt her en sammenhæng med korrelation r = 0,42 (R kvadreret = 0,17). Ved en simpel regressionsmodel med prævalens som forklarende faktor, er cirka 82% af variationen i optimalt cut-off altså stadig uforklaret. Det gør det interessant at overveje hvilke yderligere faktorer, der kan tænkes at indvirke på resultaterne, både som systematisk afvigelse og støj. En kilde til systematisk

afvigelse kan være anvendelse af forskellige guldstandarder, hvilket vil mindske korrelationens styrke, fordi støjen vil øges i det samlede datamateriale. Risikoen for dette undersøger vi senere i del 3 af analysen. Yderligere kan inklusionen af de mindre studier være en betydelig faktor, qua tilfældighed i resultaterne som følge af større sampling fejl for disse. Disse faktorer fører til spørgsmålet om hvorfor Youden's index anvendes så ofte. Det er tidligere blevet påpeget at Youden's index er det mest simple og netop er udbredt grundet simpliciteten (Vouloumanou et al., 2011).

Foruden begrænsningerne i Youden's index, er selve variationen i studiers estimater af optimalt cut-off en udfordring og et potentielt problem. Denne variation studierne imellem, og det at den kun i begrænset omfang kan forklares ud fra forskelle i prævalens, er interessant fordi valideringsstudier ofte drager konklusioner omkring deres bestemmelse af et optimalt cut-off for en bestemt population, ofte en nationalitet. Med den betydelige heterogenitet i studiernes fund, opstår en risiko for at selve den statistiske grundidé går tabt, i det resultatet bliver et estimat af et optimalt cut-off for samplet og ikke for populationen. Som nævnt ovenfor kan der her være et argument for at eksempelvis sprogligt-kulturelle forskelle kan lede til forskelle imellem populationer, særligt mellem forskellige nationaliteter. En åbenlys indvending kunne derfor være, at forskellene studier imellem kan afspejle virkelige forskelle mellem populationer og ikke varians. Dette ser vi som en faktor af begrænset betydning, da vi mener at variationen i de enkelte populationer er markant større end variationen mellem studier. Når det samtidig gør sig gældende, at mange valideringsstudier har et begrænset deltagerantal, vil denne variation ikke konvergere. Videre vil det være svært at afgøre hvilken del af forskellen imellem to populationer der skyldes målefejl, da populationen sjældent vil kunne isoleres som faktor i valideringsstudier. Enhver forskel vil derfor kunne konfundere resultatet, hvilket vil være særligt farligt ved mindre

deltagerantal. Foruden nødvendigheden af at studier anvender samme guldstandard, vil forskelle i de klinikere der udfører de kliniske interviews kunne lede til konfundering. Dette vil vi komme ind på i diskussionen.

Beregninger ud fra nytteværdi, forudsætter en enhed for denne og her valgte vi at anvende et estimat af samfundets omkostninger. Definitionen af det optimale cut-off blev derfor, det for hvilket omkostningerne for samfundet var mindst. Smits et al. (2007) har beskrevet at nytteværdi i forbindelse med diagnoser kan optimeres fra fire perspektiver. Dette er henholdsvis patientens, sundhedsforsikringens (eng: health care provider), samfundets og forskeres perspektiver. Det kan diskuteres om vi har valgt det mest relevante af disse. Særligt vil patientens nytte som kriterium antageligt have stærk intuitiv appel for flertallet. Vores valg af perspektiv skyldes dels det pragmatiske hensyn, at det er simpelt at anvende og dels at omkostningerne ved behandling ikke kan ignoreres. Et grundlæggende princip indenfor sundhedsøkonomi er alternativomkostninger (Pedersen, 2013).

Ideelt set skulle alle have en optimal og uindskrænket behandling, altså skulle patientens nytte optimeres, men fordi midlerne er begrænsede, vil én patients behandling som udgangspunkt begrænse en andens mulighed for behandling. Vi mener derfor ikke det anlagte perspektiv er uden relevans, men påpeger samtidig, at dette perspektiv ikke kan stå alene. Endelig bygger den nytteværdi-baserede tilgang også (indirekte) på antagelser om effekten af depressionsbehandling, særligt når omkostningsniveau tages som perspektiv - men dette vender vi tilbage til. Disse overvejelser er dog sekundære, da det vigtigste her er at vise indvirkningen af metodiske valg og at der er alternative tilgange til Youden's index.

Det logiske skridt herefter er at undersøge, hvorvidt vi vil få et resultat, der differentierer sig fra Vilagut et al. (2016), hvis vi analyserer alt data.

Screeningsværktøjers nøjagtighed

I denne analyse vil vi estimere et optimalt cut-off niveau for CES-D skalaen. Vores grundlæggende rationale for analysen, er at udnytte mest mulig information, ved dels at inkludere samtlige studier og dels at inkludere al tilgængelig data fra disse studier. Begge disse dele står i modsætning til hvad der tidligere er gjort.

Vilagut et al. (2016) angiver at deres undersøgelse er den første meta-analyse af CES-D skalaen.

De inkluderede studier spænder over en række nationaliteter og både kliniske og ikke-kliniske sub-samples. Denne bredde i studierne står i kontrast til den generelle tendens i valideringsstudier, at disse typisk er fokuseret mod at validere en given skala indenfor en afgrænset population, typisk en nationalitet. Det typiske argument, for denne afgrænsning er sprogligt-kulturelle forskelle, lande imellem (Manea et al., 2012; Gilbody et al., 2007). Dette ser vi som et reelt argument, men vi mener det er uheldigt at argumentet står alene, fordi man i forsøget på at undgå denne faldgrube kan overse andre faldgruber, som vi ser som vigtigere. Problematisk er her særligt samplingen, fordi deltagerantallet ofte er begrænset. Ved en tilgang med en afgrænset demografi, men også en begrænset sample size, vil man nedprioritere intra-populations varians til fordel for inter-populations varians. Det mener vi er en fejl fordi variationen på intra-populations niveau må antages at være større end variationen på inter-populations niveau. Her bemærker vi fraværet af et teoretisk grundlag for at depressionsbegrebet skulle variere med nationalitet (Manea et al., 2012; Vilagut et al., 2016).

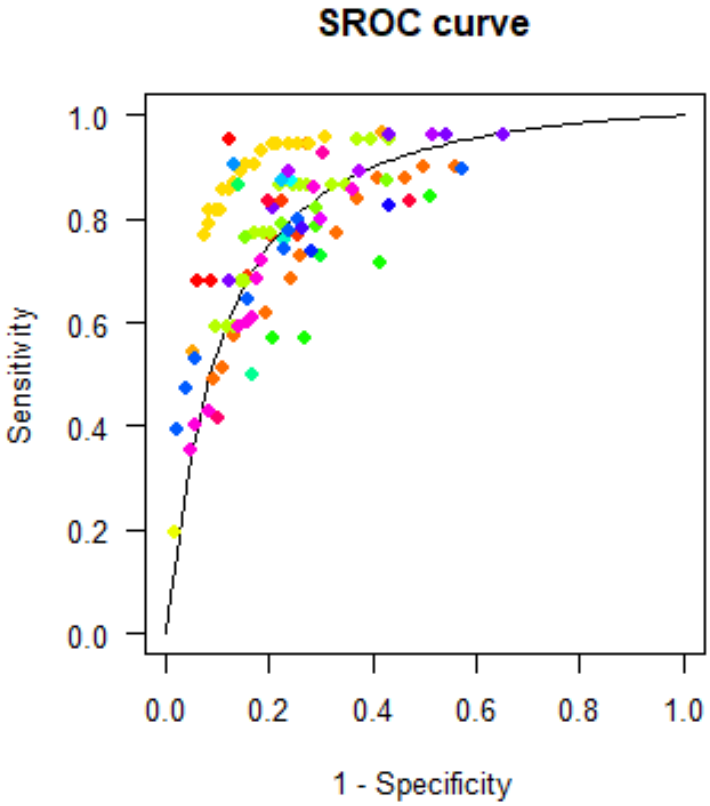
Vores rationale er derfor: en større og mere heterogen sampling har større sandsynlighed for at klarlægge de reelle underliggende tendenser end en mindre og demografisk indsnævret sampling.

De fleste meta-analyser af depressionsskalaer diagnostiske nøjagtighed en bivariat model (se evt. teoriafsnittet). Det har som beskrevet den konsekvens, at hvert studie kun kan bidrage med én cut-off værdi. Meta-analytikeren skal derfor udvælge et repræsentativt cut-off niveau for hvert enkelt studie og inkludere de tilhørende værdier for sensitiviteten og specificiteten.

Den første analyse inkluderer al den tilgængelige data fra de inkluderede studier. Totalt er der 110 cut-off værdier og 220 datapunkter da hver cut-off værdi har både en sensitivitet og specificitet tilknyttet. Antallet af rapporterede cut-off værdier varierede mellem 1 (n=16) og 18 (n=1).

Fra Diagmeta pakken anvender vi funktionen af samme navn, med parameteren "model" sat til DIDS og den totale mængde data som input. Alle indstillinger i modellen er her sat til deres standard niveau (Steinhauser et al., 2016). Når de indgående studier ses som bedste estimat af populationen af depressive patienter er cut-off 17,16. Ved dette cut-off opnås en sensitivitet på 0,8 (95% konfidensinterval 0,708 – 0,869) og en specificitet på 0,762 (95% konfidensinterval: 0,713 – 0,806).

Sammenhængen mellem sensitiviteten og specificiteten ved hvert cut-off niveau for de enkelte studier er vist på figur 6.



[Figur 6: Summary receiver operation characteristics (SROC) kurve for studierne inkluderet i analysen. Hvert punkt angiver sensitivitet og specificitet for et cut-off niveau.]

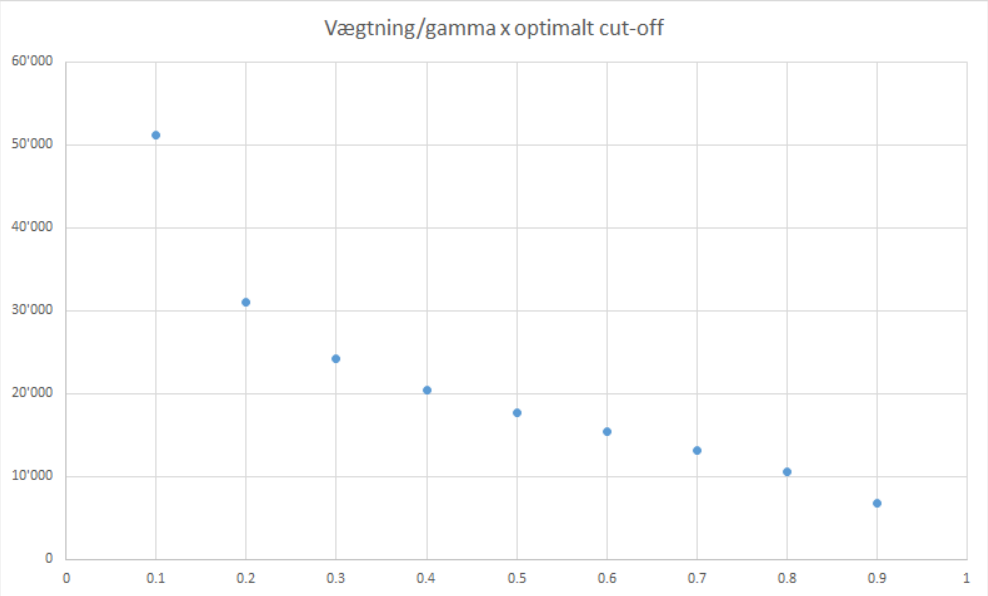
Ved eksperimentel tilgang fandt vi, at dette resultat er robust over for outliers. Eksempelvis er betydningen af at inkludere studiet Roberts (1991), som finder en særligt høj optimal cut-off værdi på 24 og er det andet største studie (hvilket øger dets vægtning), en ændring til 16,26 som estimat af bedste cut-off (forskel 0,9)

Resultatet af vores analyse afviger fra Vilagut et al. (2016), som finder den optimale cut-off værdi 20. I praksis må vi afrunde vores estimat til nærmeste heltal, altså 17. Den eksakte forskel mellem estimaterne er derfor 3 sumscore point på CES-D skalaen.

Studiet Cho (1993) er det eneste studie, for hvilket vi har data for begge disse værdier og er samtidigt det største af de undersøgte studier, med 2008 deltagere. Antallet af ægte positive

testresultater ville i dette sample være 74 ved cut-off 17 og 70 ved cut-off 20, en forskel på fire tilfælde (5,4%). Antallet af falske positive vil ved cut-off 17 være 353 og vil ved cut-off 20 falde til 278, en forskel på 75 tilfælde (21,2%).

Alt dette bygger som nævnt på en lige vægtning af sensitivitet og specificitet. Ud fra logikken udledt i foregående analysedel, viser figur 7 det optimale cut-off vi finder ved forskellige vægtninger af sensitivitet og specificitet.



[Figur 7. X = vægtning af sensitivitet, y = resulterende optimalt cut-off]

Ved at se på prævalensen i samplet, kan vi udregne hvilket forhold af vægtning der svarer til de absolutte værdier for nytte, vi anvendte i prævalensanalysen. Differencen mellem en falsk negativ og en ægte positiv er 11309-9635= 1674. Differencen mellem en falsk positiv og en ægte negativ er 124-0 = 124. Ratioen mellem disse to tal er 1674/124 = 13,5, hvilket givet en vægtning på 13,5/(13,5+1) = 0,931 for sensitivitet.

Vi kan nu ud fra samme model som ovenstående igen beregne det optimale cut-off, her med parameteren lambda sat til 0,931. Med denne vægtning finder vi det optimale cut-off 6,25.

Sensitiviteten er her 0,993 med konfidensintervallet 0,977 til 0,998 og specificiteten er 0,225 med konfidensintervallet 0,121 til 0,367

Vi har her ikke sammenholdt dette med implikationen for diagnosticering ved det største inkluderede studie. Dette af den simple grund at ingen enkeltstudier inkluderer så lavt et cut-off niveau.

Diskussion af denne analyse

Til estimeringen af det optimale cut-off knytter sig en række metodiske overvejelser, særligt omkring selve det at definere et optimalt cut-off. Videre kan denne analyse ikke anskues i et vakuum, i det resultatet hænger tæt sammen med den efterfølgende analyse af forskellen imellem interviews, fordi denne undersøger graden af heterogenitet i det overordnede datamateriale. Endelig er resultaterne begrænset af den statistiske usikkerhed, som har været svær at vurdere.

Vores estimat, som beregnet ved ligevægtning af sensitivitet og specificitet, ligger tæt op ad den oprindelige anbefaling for CES-D skalaen (Radloff, 1977). Dette er interessant i sig selv, men her er det centralt at denne lighed i estimaterne opstår netop når vi ligevægter sensitiviteten og specificiteten, hvilket vi uddybende har gennemgået problemerne ved i det foregående analyseafsnit. Imidlertid er det interessant her at se deres begrundelse for det optimale cut-off.

I originalstudiet anvendte man ikke Youden's index eller tilsvarende og videre forholdt man sig ikke til vægtningen af sensitivitet og specificitet. Baggrunden for at fastlægge 16 som standard cut-off angives i originalstudiet i det følgende: "Seventy percent of the patients but only

21% of the general population scored at and above an arbitrary cut-off score of 16" (s. 393, Radloff, 1977). Dette er problematisk fordi denne cut-off værdi, i originalstudiets egen beskrivelse, er fastsat på arbitrær basis, og fordi dette er fastsat ud fra et enkeltstående studie foretaget i 1977. Dette gør det bemærkelsesværdigt at cut-off scoren 16 herefter blev fastlagt som standard for CES-D skalaen og fortsat i dag, mere end 40 år senere, er uændret.

Videre er det interessant at vi finder et markant andet resultat end Vilagut et al. (2016) selv når vi estimerer det optimale cut-off ved ligevægtning af sensitivitet og specificitet, og estimaterne bygger på de samme 28 studier. Der er to potentielle årsager til denne difference: forskelle i den anvendte data og forskelle i den statistiske tilgang.

Vilagut et al. (2016) angiver i deres metodeafsnit, at de har udvalgt én cut-off værdi for hvert studie, med henvisning til at deres statistiske tilgang gør dette nødvendigt. Videre har de prioriteret cut-off værdien 16, fordi det er den anbefalede cut-off værdi for skalaen. For de studier, som ikke har rapporteret data for cut-off 16, har de udvalgt det cut-off hvor sensitiviteten og specificiteten er bedst.

Vi har kontaktet Vilagut et al. (2016) for at sikre en præcis oversigt over de data de har inkluderet i deres estimering af det optimale cut-off. Dette ud fra rationalet, at vi ved at gentage analysen med den eksakt samme data, kan opstille differencen i vores og deres fund og beskrive hvor stor en del der må tilskrives henholdsvis det metodiske og det datamæssige. Vi har imidlertid ikke modtaget svar fra gruppen og er derfor ikke sikre på hvilken data de har anvendt. Derfor har vi ikke prøvet at replikere deres eksakte data set-up. Det gør det svært at vurdere hvilken del af differencen i estimaterne som skyldes hhv. forskelle i data og i den statistiske tilgang.

Når vi ser på den svage baggrund for anbefalingen af cut-off niveau 16 for CES-D skalaen, er det interessant at Vilagut et al. (2016) har prioriteret denne til fordel for cut-off niveauer som har bedre tradeoffs mellem sensitivitet og specificitet. Vilagut et al. (2016) angiver selv at kun 11% af studierne ikke har angivet data for det anbefalede cut-off niveau. Da datasættet består af 28 studier i alt må der altså være tale om tre studier, for hvilke Vilagut et al. (2016) har anvendt andre cut-off værdier end 16. Hvis dette er tilfældet vil modellen de opstiller, aflede et resultat som meget eksakt afspejler tendensen i de inkluderede studier for cut-off niveauet 16. Omvendt vil inklusionen at blot tre andre værdier, medføre en markant tilfældighed i estimatet af, hvordan sensitiviteten og specificiteten ændrer sig med cut-off niveauet. Vilagut et al. (2016) angiver at det optimale cut-off afhænger af konteksten for anvendelsen af skalaen. Dette er vi, som forklaret i prævalensanalysen, fuldt ud enige i. Imidlertid angiver Vilagut et al. (2016) som hovedpointe at cut-off værdien 20 må anses som det bedste estimat af et optimalt cut-off. Denne konklusion er baseret på det begrænsede datagrundlag finder vi overraskende, fordi vi ser denne som markant svækket af usikkerheden i data.

Endelig fandt vi med vægtningen baseret på beregningerne af nytteværdi fra foregående afsnit, at disse ville resultere i et optimalt cut-off på 6. Det er vigtigt at understrege de betydelige begrænsninger i antagelser, som afspejler at målet med denne del er at demonstrere hvor afgørende den metodiske tilgang er og ikke at optimere det bedste nuværende estimat af en ny standard cut-off værdi. Antagelserne om nytteværdien af de forskellige udfald er betydeligt arbitrær her, men eksemplificerer at det optimale cut-off kan være et helt andet end standardværdien, når spørgsmålet tilgås nytteetisk.

Teoretiske afvejninger og forskellen i disse, vil altså kunne være at stor betydning i praksis.

For de metodiske antagelser har vi prioriteret at beregne det optimale cut-off ud fra en lige vægtning af sensitivitet og specificitet. Dette er, som også tidligere angivet, den typiske tilgang i litteraturen og er om ikke andet aktuelt ift. at kunne sammenligne med Vilagut et al. (2016).

Usikkerheden i estimatet har vi desuden undersøgt ved at gentage undersøgelsen med forskellige studier trukket ud af databasen. Her fandt vi, at når vi fjernede det største studie, med 2008 deltagere, ændrede bedste estimat sig mindre end et sumscore point på CES-D skalaen. Denne ændring skal ses i lyset af at dette studie indvejer særligt tungt og dermed udgør den største påvirkning af det samlede estimat. Dette indikerer at resultatet er forholdsvis robust overfor den type tilfældighed. Udvalget af studier skulle altså have været markant anderledes for at finde et væsentlig forskelligt estimat af bedste cut-off.

Vigtigheden af at der anvendes et optimalt cut-off, og dermed relevansen af denne undersøgelse, hænger logisk direkte sammen med omfanget af brug af testen. Interessant er det her at langt det meste af diagnosticering og behandling foregår ved egen læge. Nærmere bestemt foregår 90% ved egen læge, hvor depressionsskalaer typisk anvendes ved mistanke om depression (Ulrich, 2016). Dette gør det interessant i hvor høj grad alment praktiserende læger vurderer sig i stand til at underkende patienters testresultater. Vi formoder at flertallet af alment praktiserende læger i høj grad vil lade testresultatet fra en depressionsskala afgøre den videre behandlingsplan. Hvis dette er tilfældet spiller det anbefalede cut-off en potentielt ekstremt central rolle i den danske udredning for og behandling af depression.

Analyse af sammenhængen mellem klinisk interview og optimal cut-off værdi

Rationalet i denne del af analysen er at undersøge om der er en systematisk forskel imellem de strukturerede interviews. Hvis studier, der anvender ét klinisk interview, som guldstandard, finder en højere optimal cut-off værdi end studier, der anvender et andet interview, vil dette alt andet lige indikere en sådan forskel. Omvendt kan et fravær af en sådan forskel ikke i sig selv garantere en høj inter-scale reliabilitet, fordi dette blot er en undersøgelse af systematiske forskelle i hvor stor en anden der vil score over grænseværdien. Denne analyse tester udelukkende om der er en sådan forskel.

Den samlede gruppe af studier anvendte minimum seks forskellige kliniske interviews, herunder angav enkelte studier ikke hvilket interview, der blev anvendt. De to mest brugte interviews var CIDI, anvendt i 11 studier (n total=8775) og SCID, anvendt i 5 studier (n total=2399). Vi har begrænset denne analyse til disse to interviews, fordi datamaterialet er utilstrækkeligt for de resterende.

Vi ønskede at undersøge dette ud fra samme fremgangsmåde som i foregående analyse, altså i én samlet analyse. Dette ville være en mixed model med de indgående studier grupperet efter anvendt interview som fixed effect, og inkludere en test af signifikansniveau for forskellen mellem de to grupper. Diagmeta pakken understøtter imidlertid ikke denne udvidede type af test på nuværende tidspunkt (Steinhauser et al., 2016). Vi udfører i stedet to separate tests, hvor studierne der anvender henholdsvis CIDI og SCID indgår i hver deres analyse. Her anvender vi modellen DIDS, ligesom i foregående analyse, da der er tilstrækkelig data til at modellen kan konvergere.

For studierne, der anvender CIDI, finder vi det optimale cut-off niveau: 14,83. For studierne der anvender SCID, finder vi det optimale cut-off niveau: 21,56. Differencen opgjort i CES-D sum score point er altså 6,73, hvilket vi i overførslen til praksis må afrunde til 7 point. Hvis vi afrunder disse estimater af bedste cut-offs, får vi 15 for CIDI og 22 for SCID. Studiet Cho et al. (1993) er det eneste studie, for hvilket vi har data for begge disse værdier og er samtidigt som tidligere beskrevet det største af de undersøgte studier. Antallet af positive testresultater ville i dette sample være 4 ved cut-off 15 og 11 ved cut-off 22, en forskel på 7 tilfælde (175%). Antallet af falske positive vil ved cut-off 15 være 414 og vil ved cut-off 22 falde til 235, en forskel på 179 tilfælde (43,2%).

Vi finder altså ved denne tilgang at det bedste estimat er en forskel på 7 sumscore point i optimalt cut-off for CES-D skalaen, afhængigt af om denne optimeres mod CIDI eller SCID.

Dette er en forskel af betydelig praktisk relevans, hvilket ses af implikationen for hvilken andel af populationen i studiets største indgående studie der diagnosticeres. Forskellen har som implikation at: a) de to kliniske interviews har ikke samme gennemsnitsværdi målt ved CES-D og vil derfor divergere systematisk fra hinanden, idet CIDI interviewet vil have højere sensitivitet og lavere specificitet, og b) som logisk følge af dette vil inter-skala reliabiliteten imellem de to interviews være begrænset. Resultatet af analysen stemmer overens med tidligere indikationer af at 12 måneders prævalensen af depression, varierer med brugen af forskellige kliniske interviews (Mitchell & Coyne, 2010). Dette resultat har flere vigtige implikationer og er samtidig begrænset af flere forhold. Disse implikationer venter vi med at diskutere og fokuserer i dette afsnit på begrænsningerne specifikt for denne analyse.

Analysen her er begrænset af at vi kun har undersøgt forholdet for CIDI og SCID, og kun ud fra CES-D sumscorer. Vi kan derfor ikke angive noget direkte for de andre interviews, men blot inferere sandsynlige forhold ud fra lighederne skalaerne imellem. Videre kan vi ikke kategorisk udelukke at en del af forskellen dækker over en interaktionseffekt, således at forskellen er særligt stor når den opgøres på CES-D skalaen. En sådan kunne teoretisk være tilstede hvis CES-D skalaen har større lighed med det ene interview. Der er dog intet specielt ved CES-D skalaens opbygning der giver grund til at antage at dette skulle være tilfældet, og det virker derfor som en mulighed uden videre praktisk relevans.

Dette fører til spørgsmålet om hvorvidt denne sammenligning er repræsentativ for en generel heterogenitet imellem kliniske interviews. Vi er ikke bekendte med eksistensen af studier der undersøger flere diagnostiske interviews i en sammenligning som her. Teoretisk kunne dette belyses ved at sammenligne en række interviews i en netværks meta-analyse og sammenholde differencen mellem SCID og CIDI, med den gennemsnitlige forskel mellem to tilfældigt udvalgte interviews. I fraværet af et sådant sammenligningsgrundlag må vi derfor, som ovennævnt, se på lighederne skalaerne imellem. Når vi ser på rationalet bag de to interviews og indholdet af dem, er der ikke noget der indikerer at netop disse to interviews kunne forventes at være særligt heterogene relativt til det bredere felt af kliniske interviews. Modsat ville det teoretisk forventes at de to interviews var relativt homogene, fordi de er udviklet ud fra DSM og ICD, og rationalet for ICD diagnosen har været at tilpasse den til definitionen i DSM.

På praktisk niveau er det imidlertid vigtigt da CIDI og SCID er blandt de absolut hyppigst anvendte interviews. Hvis det skulle vise sig at CIDI og SCID er de to indbyrdes mest heterogene interviews, er den betydelige forskel imellem dem derfor stadig af stor relevans. Dette fører os videre til overvejelser om datamæssige begrænsninger for resultatet.

Vores analyse baserer sig på et højt deltagerantal og samtidig på relativt få studier, særligt for SCID, hvor der indgår fem studier. Da Diagmeta sammenregner de enkelte samples til én population og vejer disse efter deres størrelse, undgås risikoen for at støjen i resultaterne fra mindre studier vil kunne påvirke resultatet uforholdsmæssigt. Det er derfor ikke ubetinget et problem at den samlede population fordeler sig over et mindre antal studier, med stor heterogenitet i størrelse og prævalens. Trods fordelene ved at anvende Diagmeta, er det imidlertid fortsat af betydning om der er systematiske forskelle studierne imellem. Diagmeta modellen forholder sig til heterogenitet i studierne ved at antage at disse alle tilhører den samme overordnede population (Steinhauser et al., 2016). Det er afgørende for denne metode at variationen mellem studiernes udførelse er begrænset, herunder særligt faktorer udover hvilke klinisk interview der anvendes. Når dette er tilfældet kan betydningen af hvilket klinisk interview der anvendes isoleres som variabel og studierne kan meningsfuldt ses som én population. Desto mindre variationen mellem studierne er desto mindre er betydningen af det lave antal studier.

En generel metodemæssig svaghed i valideringsstudier af depressionsskalaer er, at den enkeltes diagnose ofte kun fastsættes af én kliniker. Der anvendes altså ikke eksempelvis en tilgang hvor to klinikere skal være enige ved uafhængige diagnoser. Dette betyder at diagnosen er påvirket af både det kliniske interview og af den aktuelle fagperson, hvilket vi vender tilbage til i afsnittet om beslutningstagning og diagnoser. Denne variation vil kunne lede til systematiske forskelle imellem studier, særligt ved mindre studier hvor et begrænset antal klinikere medvirker og forskellen imellem klinikere derfor ikke udlignes.

Det netop gennemgåede udgør et eksempel på en årsag til en bredere metodisk udfordring i at vurdere den præcis usikkerhed i estimater ud fra Diagmeta.

Analyse af faktorstruktur og intern konsistens

Der ligger som nævnt en antagelse om en endimensionel struktur i det underliggende begreb, bag stort set alle de oftest anvendte analyser i valideringsstudier. Vi har derfor lavet en gennemgang af en række valideringsstudier for at se hvordan resultaterne og fortolkningen af disse, stemmer overens med antagelserne.

Vi vil her i en kortere analyse forsøge at belyse et par potentielle problemer ved dette, ud fra empirien i de gennemgåede studier og et par simulationer af data.

Datamateriale og fokus

Vi tager udgangspunkt i et review af 118 valideringsstudier for BDI-ll med fokus på de psykometriske egenskaber lavet af Wang & Gorenstein (2013). De søger at give et overblik over brugbarheden af BDI-ll på tværs af populationer og har derfor blot gennemgået studierne og ikke lavet en meta-analyse. Vi har udvalgt de 30 første af disse studier og lader disse repræsentere populationen af studier for skalaen. Vores analyse er ikke tænkt som en systematisk gennemgang af resultater og fortolkninger, men har blot til formålet at vise mulige problemstillinger.

Vi har her valgt at lægge fokus på BDI-ll på grund af den ekstreme udbredelse og anvendelse af skalaen. At vi følger Wang og Gorenstein (2013) skyldes at dette er den nyeste og bredest dækkende oversigt over studier rettet mod de psykometriske egenskabe af BDI-ll, samt at dette allerede er citeret mere end 350 gange. Wang og Gorenstein (2013) rapporterer resultater fra faktoranalyse, Cronbach's Alpha, Test-retest målt med Pearson korrelation, korrelationer med

andre depressionsskalaer, angstskalaer og andre psykologiske skalaer. Her har vi fokuseret på hvad studierne finder når de undersøger faktorstrukturen og den interne konsistens i skalaen.

Faktorstruktur

Wang og Gorenstein (2013) inkluderer 89 faktoranalyser baseret på de 74 studier i deres review. Af disse er 32 EFA, 45 CFA, 11 PCA og 1 studie benytter sig af multidimensional scaling. De 89 faktoranalyser finder mellem 2 og 4 faktorer med et faktortypetal på 3.

Vi finder at ingen af de 30 studier, vi læser igennem, skriver overvejelser omkring problemer med kausaliteten i skalaen. I flere af studierne bruges tre argumenter for at underbygge skalaens validitet.

Det første af disse er, at studierne nævner at de finder en faktor, som alle items loader på, og anvender dette som argument for at sige noget om skalaens dimensionalitet og kausalitet. Det andet er at der findes en faktor, der alene kan forklare en stor del af variansen. Ved dette argument angives oftest at denne faktor forklarer x %. Vi mener at man skal være meget varsomme med sådanne fortolkninger, da disse resultater kan være forårsaget af støj.

Derudover lægger en del studier vægt på at de finder samme mønster i faktorladninger som tidligere studier og bruger dette som et argument, for at de har replikeret disse resultater.

Disse argumenter kan være svære at forholde sig til og er ikke videre falsificerbare fordi der ikke er retningslinjer for hvornår en skala kan siges at være valideret.

Vi vil her undersøge tyngden af disse argumenter ved at etablere en baseline at sammenligne resultaterne med, så disse kan ses i en kontekst fremfor et vakuum. Denne baseline etablerer vi ved at simulere data.

Forbenius-Perrons læresætning siger, at en gruppe variabler der alle er ikke-negativt korreleret, altså en positiv korrelationsmatrice, vil have en faktor med en egenværdi højere end alle andre faktorer (egenværdi > 1) og alle variabler vil være positivt loaded på denne faktor (Krijnen, 2004; Pillai, Suel & Cha, 2005; Livshits, MacDonald & Radjavi, 2017).

Det er derfor muligt at udvinde en faktor, der kan se ud som om den ”forklarer” en betydelig del af variansen af en positiv korrelationsmatrice, der er kunstigt skabt og altså ikke er skabt af en latent variabel og som alle variabler vil loade positivt på. Det vil vi forsøge at vise her:

Vi starter med at skrive en funktion, der simulerer en tilfældig korrelationsmatrice med 7 variabler udelukkende bestående af variabler med ikke-negative korrelationer.

Vi sætter maksimum for korrelationen til 1 og minimum for korrelationen til 0 og angiver at der ikke behøver at herske såkaldt diagonal dominans (eng: diagonal domination), hvilket betyder at summen af de ikke-diagonale værdier ikke behøver være lavere end summen af de diagonale værdier (Deaconu, 2014).

Vi skriver funktionen således at den bliver ved med at generere matricer indtil den laver en, der passer på disse specifikationer.

Funktionen giver os den følgende korrelationsmatrice:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	1.00000000	0.02270653	0.02960963	0.6093289	0.2957229	0.2292860	0.1365431
[2,]	0.02270653	1.00000000	0.35435758	0.4117427	0.5651395	0.1562648	0.3083198
[3,]	0.02960963	0.35435758	1.00000000	0.1802606	0.6179441	0.8173178	0.2789120
[4,]	0.60932891	0.41174267	0.18026061	1.0000000	0.6286682	0.4342119	0.1522889
[5,]	0.29572291	0.56513954	0.61794409	0.6286682	1.0000000	0.6147022	0.6588177
[6,]	0.22928602	0.15626484	0.81731784	0.4342119	0.6147022	1.0000000	0.1590632
[7,]	0.13654313	0.30831976	0.27891200	0.1522889	0.6588177	0.1590632	1.0000000

[figur 6 : tilfældigt genereret ikke-negativ korrelationsmatrice]

Nu skriver vi en funktion som laver en korrelationsmatrice som ovenfor, men denne gang simulerer vi et datasæt fra 500 testpersoner med ovenstående korrelationer og på denne matrice laver vi en maximum-likelihood faktoranalyse med én faktor. Vi får så en faktor, der er associeret med 42 procent af variansen svarende til en egenværdi på 3 (42/(100/7)) i den tilfældigt genererede korrelationsmatrice.

Ved denne fremgangsmåde vil der være et element af tilfældighed. Vi gentager derfor simuleringen 100 gange og lader resultatet konvergere mod den forventede værdi, som det er almindeligt i monte carlo simulationer. Herved finder vi at at den gennemsnitlige proportion af variansen, der kan forklares af en faktor, er 38 % svarende til en egenværdi på 2,7 (38/(100/7)).

Dette viser at en faktoranalyse, af data uden et sandt kausalitetsforhold, men blot med uden negativ korrelation, vil finde en faktor som kan forklare en størredel af variansen.

Cronbach's Alpha

Vi undersøger i hvilket omfang de studier, der inkluderede både Cronbach's Alpha og faktoranalyse, beskriver problemer med CA når de finder mere end en faktor.

Bruddet på endimensionalitet, betyder også brud på tau-ækvivalens, hvilket betyder at $s_{xixj} \neq s^2_t/k^2$ og at tælleren i ligning 2.1 derfor ikke er lig med den sande true score varians.

$$a = k/(k-1)((S_{i'}S_j s_{ij})/s^2_x) = k^2 Mean(s_{ij})/s^2_x \qquad (2.1)$$

Derudover kan endimensionalitet føre til korrelation mellem itemstøj. Det kan betyde at inter-item kovariansen ændrer sig fra s^2_t/k^2 til s^2_t/k^2+s_{eiej} (se evt. teori). Alpha vil stige, men det vil reliabiliteten ikke, derfor vil $a > p_{xx'}$.

Samtlige af de 30 studier, beskriver at deres CA-analyser viser at studierne har god reliabilitet og intern konsistens. Brugen af CA i skalaer, der ikke opfylder de antagelser vi lige har opremset er problematisk. Det er ikke til at sige hvorvidt CA bliver overestimeret eller underestimeret, det er blot muligt at konkludere at CA er fejlestimeret (Yang & Green, 2011).

Diskussion af disse fund:

Ingen af de studier vi har læst nævner noget om de teoretiske forudsætninger for faktoranalyse, hvilket er problematisk, men måske forståeligt med tanke på studierne.

Det vil ikke give nogen mening at lave faktoranalyse, hvis ikke man har god grund til at vurdere at ens data passer ind i en refleksiv model. Som nævnt i teoriafsnittet vil en faktoranalyse af data fra en formativ model være problematisk, da man ikke inkluderer al varians i modellen.

Vi finder det stærkt bemærkelsesværdigt at ingen af studierne nævner overvejelser omkring dette, når ingen af dem finder en endimensionel struktur.

Stort set alle de studier vi har læst, henviser til tidligere og lignende valideringsstudier af BDI-II og udførligt redegøre for hvad disse studier har fundet. Dette indikerer at forfatterne er opmærksomme på at andre studier har fundet mere end en faktor, inden de laver deres egne undersøgelser. Dette får imidlertid ikke forfatterne til at angive overvejelser af mere fundamental karakter. Af sådanne årsager mener vi, at det ville være særligt relevant at overveje berettigelsen

af den teoretiske tilgang, herunder om det at operationalisere skalaen som en refleksiv latent variabel er understøttet. Af denne grund vil vi herefter i et separat afsnit adressere netop dette spørgsmål, mens vi her blot konstaterer at dette ikke gøres i de valideringsstudier vi har gennemgået.

Det at ingen af de 30 studier adresserer disse forhold, ser vi tre mulige årsager til: 1) De antager at faktorstrukturen er anderledes i det subsample de har, det vil sige de antager BDI-depression er anderledes i f.eks. Indien frem for USA. Et problem der er meget lig det vi beskriver i diskussionen af rationalet bag det at optimere cut-off på specifikke populationer og med små sample size. 2) De antager at det er så sikkert at BDI kan forklares af en refleksiv latent variabel, at de finder det unødvendigt at adressere det. 3) De er ikke bekendte med antagelser bag de modeller de anvender.

Det er ikke til at sige med sikkerhed, hvilken årsag der er den rigtige og det behøver selvfølgelig ikke være det samme for alle 30 studier.

Ydermere er det problematisk at flere af artiklerne antyder kausalitetsforhold i data, hvilket må siges slet ikke at kunne retfærdiggøres. Det kan diskuteres hvorvidt man under visse antagelser, måske forbeholdent kan sige noget om endimensionalitet ud fra en faktoranalyse, men det kan absolut ikke lade sig gøre når man finder resultater, der direkte modsiger dette.

I vores simulationer viser vi, at flere andre argumenter fra faktorstudier, som det at alle items loader på en faktor eller det at en faktor kan forklare en vis procentdel af variansen, det kan være problematiske tolkning, når man ved tilfældighed kan simulere sig frem til det samme.

Wang & Gorenstein (2013) skriver intet om de mulige udfordringer med studierne, hverken bruddet på antagelserne bag Cronbach's Alpha (de beskriver, det sågar som et mål for

skalaernes reliabilitet) eller de mulige brud på målingsinvarians (faktorstrukturen varierer mellem studierne).

Diskussion

Indledning

Ved analysens tredje afsnit, om forskelle imellem de to kliniske interview SCID og CIDI, ventede vi med at diskutere implikationerne af vores resultater. Dette skyldes at disse danner en overgang fra analysens første to dele til de senere dele af analysen. Det kliniske interview anses som guldstandarden for diagnosticering af depression og danner fundamentet for validering af depressionsskalaer. Enhver kritik af guldstandarden er derfor en udfordring for valideringen af depressionsskalaer. Som angivet i indledningen, afspejler dette skifte at vi på den ene side finder betydelige udfordringer for hele den nuværende tilgang til depression, men omvendt anser det for usandsynligt at tilgangen vil ændre sig markant foreløbigt.

Først og fremmest er det interessant at to forskellige kliniske interviews ser ud til at afvige i hvor sensitive de er for depression, eller udtrykt anderledes: hvordan de implicit definerer depression. Dette introducerer et element af tilfældighed i diagnoseprocessen, fordi brugen af interviews varierer, ligesom brugen af depressionsskalaer varierer. Hvis disse resultater kan replikeres betyder det i praksis at der vil være forskel på at have en depression, som defineret ud fra SCID, fremfor ud fra CIDI. Dermed vil diagnosen ikke kunne stå alene uden en specificering af hvilken metode der er anvendt ved diagnosticeringen af den enkelte patient.

Reliabiliteten for diagnosticeringen af depression begrænses dermed ikke blot af usikkerhed pga. forskelle imellem klinikere (inter rater reliabilitet), men også af systematisk bias i hvilket interview en given del af sundhedssystemet anvender. I en klinisk kontekst betyder det

at patienten kan gå til én kliniker og få én diagnose og til en anden og få en anden eller ingen diagnose. Dette fører til en mere fundamental diskussion om berettigelsen af at have flere operationaliseringer af det samme begreb, som det er tilfældet når der eksisterer en række interviews som bruges aktivt. Hvorvidt dette kan forenes med at tale om ét fælles og invariant begreb, gennemgår vi senere.

Vi vil her gennemgå en række aspekter af potentielle svagheder ved det kliniske interview som guldstandard for diagnosticering.

Kliniske interviews

Vi finder det relevant her at skelne mellem to måder at tilgå validiteten af kliniske interviews som guldstandard. Disse opstiller vi som følgende:

1) Et individ defineres som værende depressivt, når en trænet kliniker ved et klinisk interview vurderer en person som depressiv. Guldstandarden defineres hermed ved et aksiom og der er ikke grundlag for en undersøgelse af dets berettigelse.
2) Et individ defineres som værende deprimeret eller ikke ud fra en sand underliggende tilstand (ontologisk), som kan variere i sværhedsgrad. Guldstandarden vil derfor være den bedste approksimation af tilstedeværelsen (og graden) af denne tilstand. Dette lægger op til diskussion, fordi det er givet at en klinikers vurdering ikke vil stemme perfekt overens med tilstedeværelsen af denne tilstand og derfor vil være suboptimal.

Vi har ikke set præcis denne opstilling i litteraturen, men noterer os at den tilgang de fleste valideringsstudier bruger, de facto er den samme som den første tilgang vil foreskrive, da

der ikke ses en kritik af guldstandarden, men blot forsøg på at optimere overensstemmelsen med denne.

Succeskriteriet for en guldstandard er at denne er valid og at operationaliseringen af denne faciliterer reliabilitet. Det er her vigtigt at validiteten vil svækkes af en lav reliabilitet, mens reliabilitet omvendt ikke i sig selv kan garantere validitet – at en diagnose har et validt grundlag, er ikke en forudsætning for at klinikere vil kunne være enige om bedømmelsen af den. Fra denne vinkel er det interessant at se på hvad klinikerne ser ud til at kunne. Dette kan vi belyse fra to vinkler. Den ene er den teoretiske vinkel, hvor vi kan gå til forskningen i beslutningstagning og ekspertiseudvikling og se på hvad man vil kunne forvente på dette domæne. Den anden vinkel er den empiriske og her vil vi starte.

Empiri for kliniske interviews

Det at psykologens diagnose anvendes som guldstandard, bringer et fundamentalt spørgsmål op, som vi introducerede i indledningen: hvordan kan selve guldstandarden valideres? Ethvert alternativ til en guldstandard kan testes for om det korrelerer med den etablerede guldstandard og dermed udviser konvergerende validitet. En mulig tilgang er at starte med at se på reliabiliteten for depressionsdiagnosen.

Test-retest reliabiliteten af DSM-5 diagnoser er testet i forbindelse med udviklingen af DSM-5. Reliabiliteten er her udtrykt ved kappa-koefficienten, som beskriver reliabiliten kontrolleret for den forventede værdi ved tilfældig diagnosticering. Denne værdi vil være en funktion af prævalensen. Her fandt man at kappa værdien for depression er 0,28 (Regier et al., 2013). Til sammenligning fandt man for Borderline en kappa værdi på 0,54, en diagnose som ellers har været hårdt kritiseret og vides at være svær at diagnosticere (Biskin & Paris, 2012).

Anerkendelsen af dette har flere steder ført til konstateringer af at DSM-III og DSM-IV var bedre end DSM-5, så vidt angår reliabilitet (Mitchell & Coyne, 2010). Dette er interessant og kan lyde faktuelt, men denne kritik er ikke nødvendigvis berettiget. Chmielewski et al. (2015) har undersøgt betydningen af den metodiske tilgang til reliabilitetsestimaterne. DSM-5 undersøgelserne er baseret på test-gentest reliabilitet, hvor den enkelte patient testes i to omgange af forskellige klinikere. Undersøgelserne for DSM-4 (og tidligere) var baseret på lydoptagelser, hvor den enkelte patient kun interviewes en gang og begge klinikere diagnosticerer ud fra dette. De finder en interessant forskel på tværs af diagnoser. For lydoptagelsesmetoden finder de den gennemsnitlige kappa værdi 0,8 og for test-gentest metoden finder de kappa værdien 0,47. Lieblich et al. peger på denne baggrund på at de lave kappa værdier for DSM-V sandsynligvis ikke er resultatet af en svækket definition af diagnosen, men snarere af en forbedret metode som fremviser en svaghed der hele tiden har været gældende for diagnosen.

Det kan være svært at forholde sig til betydningen af kappa værdier, særligt fordi disse hænger sammen med prævalensen i det aktuelle sample. Effekten kan tydeliggøres ved et eksempel. Ved en prævalens på 20% vil en kappa værdien på 0,28 opnås ved at psykiaterne stiller samme diagnose i 42,4% af tilfældene. Fordi den forventede værdi af en tilfældig diagnosticering er en enighed lig prævalensen, er differencen altså 22,4 procentpoint.

Disse kappa værdier er bemærkelsesværdigt lave. Dette peger ligesom den tidligere diskuterede forskel imellem interviews på problemet i at den samme person kan gå til to forskellige klinikere uden at resultatet af diagnosticeringen vil være det samme. En grund til dette kan være at klinikere ofte ikke benytter sig af diagnosekriterierne (Blashfield & Herkov, 1996; Garb, 2005). Det er her interessant at forstå hvorfor disse kappa værdier er så lave. Her har

det tidligere været diskuteret fra et beslutningstagningsperspektiv at klinikere blandt andet vil operere ud fra repræsentativitets-heuristikken (Garb, 1996). Denne vil, som præciseret i teoriafsnittet, kunne indvirke både positivt og negativt, afhængigt at den specifikke kontekst. Klinikeren vil være mere tilbøjelig til at stille diagnosen depression desto mere patienten fremstår i lighed med klinikerens billede af en depressiv person. Her er det interessant at det tidligere er fundet i et sample af over 3000 deprimerede individer, at disse indbyrdes havde mere end 1000 forskellige unikke kombinationer af depressionssymptomer (Fried & Nesse, 2014).

Denne ekstreme heterogenitet peger på at depressive patienter vil kunne fremstå meget forskelligt, fordi deres beskrivelser af symptomer vil variere. En kliniker kan derfor eksempelvis møde en patient der har udtalt hypersomni og manglende appetit, mens en anden klient kan have hyposomni og en øget appetit. Her vil repræsentatitivets-heuristikken kunne forventes at føre til suboptimale vurderinger og dette udgør dermed én mulig forklaring på de lave kappa værdier.

At det er svært at finde undersøgelser fokuseret direkte på klinikerens evner skyldes i høj grad at dette netop anses for at udgøre guldstandarden. Dette gør sig eksempelvis gældende for en række studier som har undersøgt hvorfor alment praktiserende læger overser mange depressionstilfælde, hvorved det er her implicit at klinikeren har ret (Mitchell & Coyne, 2010). Et bud er at lægerne, som følge af diverse heuristikker og biases, måske danner sig et forhastet indtryk.

Dette er måske en plausibel forklaring, men imidlertid også en faktor som ser ud til være gældende for klinikere. Tidligere studier har fundet at psykiatere forevist videooptagelser af interviews formede diagnostiske indtryk indenfor et minut (Gauron & Dickinson, 1966) og at psykiatere generelt ofte er klar til at bestemme en diagnose indenfor få minutter (Kendell, 1973).

Ekspertiseudvikling i diagnosticering

Når man kigger til forskningen indenfor læring og ekspertiseudvikling, finder man at den kliniske psykologi og psykiatri er et svært område at blive ekspert indenfor. Særligt har Shanteau (1992) peget på, at analysen af eksperter ofte kan komme til at dreje sig om en række personlige faktorer, mens betydningen af den konkrete opgave og dens miljø overses.

Når vi ser på veludførte større enkeltstudier af terapieffekt viser der sig et ganske klart billede. Fundene varierer, men den klare tendens er at der findes enten ingen eller en negativ erfaringseffekt blandt kliniske professionelle – den mere erfarne psykolog hjælper ikke klienten mere end den nyuddannede (Vollmer et al., 2013, Van Oppen et al. 2010, Groove et al. 2000, Ægisdottir et al., 2006). Enkelte studier har fundet en positiv korrelation med behandlingseffekt ved svære lidelser, men negativt korreleret ved lette og moderate lidelser (Mason et al., 2016).

Vi er i mindre grad bekendte med undersøgelser af erfaringseffekten specifikt rettet mod diagnosticering. I lighed med vilkårene for psykologen der udøver psykoterapi er her dog også dårlige vilkår for udvikling, navnlig et minimum af feedback tilgængeligt for klinikeren. Her kan man spørge om der set i lyset af dette kan forventes at finde en erfaringseffekt i forbindelse med diagnosticering.

Det er tidligere undersøgt om der er særlige grupperinger som ofte fejldiagnosticeres (Perez-Stable, Miranda, Ying & Munoz, 1990). Her er tidligere fundet at dette særligt er folk der er ældre, lavere uddannede, har flere besøg ved lægen og anvender mere medicin.

Klinisk interviews og depressionsskalaer

Ved sammenligning af depressionsskalaer og diagnostiske interviews er det vigtigt at gøre det klart om formålet er screening eller diagnosticering. Fordi de fleste depressionsskalaer er udviklet til screening er deres standard cut-off forsøgt optimeret til dette.

Det er vigtigt at forstå at skalaers præcision afhænger af hvilket cut-off der anvendes. Her har der været en tendens i litteraturen til at anvende det anbefalede cut-off i sammenligninger mellem skalaer og diagnostiske interview (Zimmerman, Walsh, Friedman, Boerescu & Attiuhllah, 2018). En sådan sammenligning vil helt naturligt divergere, fordi dette cut-off vil være bevidst lavt sat og derfor resultere i et højt antal falsk positive. Her bliver udfordringen at mange studier i stedet søger at optimere cut-off niveauet, men anvender for små sample sizes og generaliserer på et spinkelt grundlag, som beskrevet i vores analyseafsnit.

Flere studier har sammenlignet selvvurderingsskalaer og interview-baserede skalaer. Zimmerman et al. (2018) sammenlignede henholdsvis tre og to af disse, og fandt at effektstørrelserne var lige store. Modsat har Cuijpers, Li, Hofmann & Andersson (2010) fundet at interview-baserede skalaer har større effektstørrelse.

Det er her interessant at Zimmerman et al. (2018) lavede deres undersøgelse for at se på hvorvidt skalaerne kan differentiere mellem tidspunktet inden behandling og tidspunkter efter behandling. Manglen på kontrolgruppe er her et problem, fordi en skala både bør kunne differentiere på præ til post behandling, men også bør være uændret for en kontrolgruppe. I fraværet af en kontrolgruppe vil det derfor se bedst ud for den skala som viser størst ændring, uanset hvad denne ændring skyldes. Med denne præmis for undersøgelsen kan en skala derfor optimeres ved at maksimere anti-temporal invarians – den absolutte antitese til de fundamentale antagelser i klassisk testteori.

Et spørgsmål er her om selvvurderingsskalaer kan anvendes som en del af et klinisk interview eller helt erstatte det kliniske interview. Howard Garb har her beskrevet at omend klinisk vurdering er begrænset af en række heuristikker og biases, kan dette om ikke andet som minimum bruges til at hjælpe klienten med at udfylde en skala (Mitchell & Coyne, 2010). Vi finder det interessant at Garb her blot nævner et konkret argument for at foretage det kliniske interview. Dette reducerer den professionelles rolle markant og angiver at det er svært at finde belæg for klinikerens særlige ekspertise udi diagnosticering.

Sammenfatning

Der ses altså en lav inter-rater reliabilitet ved diagnosticering af depression. Når vi ser på klinikerens muligheder for at lære af sine erfaringer og bredden af de klienter klinikeren vil møde, som alle indgår under diagnosekriterierne for depression, er det tydeligt at udfordringen er massiv. Det står uklart i hvor høj grad det at kombinere selvvurderingsskalaer med kliniske interviews vil kunne forbedre reliabiliten af diagnosticeringen.

Denne uklarhed i de nuværende sammenligninger af kliniske interviews og depressionsskalaer fører os mod at diskutere om depressionsbegrebet bør operationaliseres som en refleksiv latent variabel.

Er det en refleksiv variabel?

De forskellige afsnit i analysedelen omkring Wang & Gorenstein (2013) artiklen berørte fra hver deres vinkel det forhold at grundlaget for de enkelte konklusioner er, at depression ses som en refleksiv latent variabel. Netop det at flere af de forudgående afsnit har henvist til dette

spørgsmål, understreger dets centralitet. Vi vil i dette afsnit se på berettigelsen af at se på depression som en refleksiv latent variabel.

Alle de statistiske modeller bygger på en central antagelse, nemlig den at skalaerne er refleksive latente variabel modeller. Spørgsmålet for dette afsnit er derfor helt centralt for både den metodemæssige tilgang og for fortolkningen af de statistiske analyser. Dette danner fundamentet for depressionsforskningen og er derfor afgørende for alle de forhold hvor depressionsbegrebet anvendes.

Endimensionalitet og intern konsistens

Refleksive og formative begreber kan adskilles ved hvorvidt et begreb er endimensionelt eller flerdimensionelt (Edwards, 2011). Dette skyldes at man ved refleksive begreber antager, at alle items/proxier er udvalgt således at de direkte afspejler den samme latente variabel. Dette fører til at forholdet må være endimensionelt (Edwards, 2011; Bollen & Lennox, 1991; Diamantopoulos & Siguaw, 2006).

Som reviewet i analysen finder, viser faktorstudierne af BDI-II studierne et meget tydeligt resultat, idet ingen af de 74 studier fandt et endimensionelt forhold understøttet. Dette ser ud til at være repræsentativt for depressionskalaer generelt, idet gennemgange af de andre store skalaer som HAM-D, CES-D og MADRS har fundet samme tendens (Shafer, 2006; Quilty, Robinson, Rolland, Fruyt, Rouilion & Bagby, 2013). Dette viser at de mest brugte depressionsskalaer, entydigt ikke finder et flerdimensionelt forhold.

At der på konsistent basis findes flere faktorer i faktoranalyser af de mest brugte depressionsskalaer, er et markant problem, når begrebet ses som refleksivt. Som angivet ovenfor, er dette imidlertid ikke noget, der har forhindretet et eneste af de valideringsstudier vi har

gennemlæst, i at bekræfte validiteten af deres skala og anse denne som en refleksiv model. Dette problem omtales meget begrænset, og når det gør er det til tider med henvisning til at den interne konsistent, målt ved Cronbach's alpha, retfærdiggør at se skalaen som en refleksiv model. Hvilket som beskrevet tidligere er yderst problematisk.

Vi gennemgik BDI-ll skalaen i vores teoriafsnit og her kan man indvende at det er en skala oprindelig udviklet i 1961, som kun i mindre grad er ændret i 1996, og at den derfor ikke nødvendigvis er den bedste skala. Dette gør sig også gældende for de andre skalaer vi har nævnt ovenfor, idet fire af de fem mest brugte depressionsskalaer alle er lavet før 1980 (Santor et al., 2006).

Et oplagt spørgsmål er her om de skalaer vi har beskrevet, ikke til fulde er repræsentative, fordi nyere udviklede skalaer måske i højere grad er endimensionelle. Der er nyere depressionsskalaer f.eks. Major Depression Inventory (MDI), en skala udviklet af den danske psykiater Per Bech, og som er udviklet som kriterierne for depression ICD-10 i spørgeskemaform. MDI er dog ikke endimensionel (Bech, Timmerby, Martiny, Lunde & Soendergaard, 2015; Christensen, Oernboel, Nielsen & Bech, 2019. Selv skalaer der er endimensionellen så som HAMD-6 der en viderebygning på HAM-D, hvor man har cuttet antal items fra 17 til 6 (Bech, Gram, Dein, Jacobsen, Vitger & Bolwig, 1975) er baseret på en ide om en refleksiv latent variabel og kan bruges til at bekræfte kausalitetsforholdet mellem variabler og indikatorer.

Netværksanalyse og sammenhængen mellem variabler

Fælles for disse principper for adskillelse af refleksive og formative modeller er at de, fra forskellige vinkler, berører hvorvidt man med den refleksive model kun forventer at måle en ting og derfor forventer homogenitet mellem items.

Tæt op ad bruddet på kriterierne om endimensionalitet og intern reliabilitet, ligger nogle nyere undersøgelser af hvordan symptomer forholder sig til hinanden og hvordan livsbegivenheder manifesterer sig i symptomer.

Nyere forskning har anvendt såkaldte netværksanalyser, hvilket i modsætning til andre testteorier, inkluderer en tidsfaktor, til at belyse spørgsmål om kausalitet (Borsboom, 2017). Disse netværksanalyser kan særligt sige noget om to aspekter: symptomernes forhold indbyrdes, hvilket for refleksive variabler forventes at være nul, og livsbegivenheders påvirkning af de enkelte items. Disse netværksanalyser har særligt fokuseret på BDI-II og HAM-D, hvilket igen skyldes den store udbredelse af disse skalaer i interventionsforskning.

I undersøgelsen af livsbegivenheder, er det blandt andet fundet, at kærestesorger fører til markante ændringer i symptomerne; tab af appetit og anhedonia. Modsat vil dødsfald i den nærmest omgangskreds særligt have en betydelig indvirkning på træthed (eng: fatigue). Generelt viser disse studier at en række livsbegivenheder har en betydeligt heterogen påvirkning af depressionssymptomer (Keller, Neale & Kendler, 2007). Dette bryder med antagelsen fra den refleksive variabel model om at man ikke kan påvirke symptomer udenom den latente variabel (Edwards & Bagozzi, 2000).

Ved netværksmodeller, illustreres det indbyrdes forhold mellem de enkelte symptomer typisk ved at samtlige symptomer angives i en model, hvor tykkelsen på stregen mellem to symptomer/items angiver korrelationen mellem disse to (Borsboom, 2017; Borsboom & Cramer, 2013). Heraf kan grafisk udledes hvilke symptomer, der korrelerer med hvilke og hvor der er

stærke korrelationer. Når en refleksiv variabel optegnes i en sådan model, vil man teoretisk forvente at to ting gør sig gældende: a) at samtlige items i netværket er korreleret med samtlige andre items, og b) alle disse korrelationer er af samme styrke. Begge disse punkter er en direkte følge af antagelsen om envejskausalitet i modellen og antagelsen om lokal uafhængighed.

Når netværksmodeller tegnes for depression, finder man markante forskelle i centraliteten af items. De enkelte items korrelerer i forskellige klynger, og udgør små netværk. Mens nogle items er meget centrale, og korrelerer med en stor del af de andre items, er andre meget decentrale og korrelerer kun med enkelte andre (Borsboom, 2017; Fried, Epskamp, Nesse, Tuerlinckx & Borsboom, 2016). Dette betyder at items ikke er indbyrdes udskiftelige, hvilket er en af antagelserne bag klassisk test teori og den latente variabel model (Edwards & Bagozzi, 2000). Bruddet med denne antagelse betyder at alle indikatorer ikke er lige gode for begrebet og der gælder derfor ikke såkaldt tau-ækvivalens. Det er problematisk, blandt andet fordi det bryder med antagelsen om lokal uafhængighed, som også er en forudsætning for at kunne se et begreb som en refleksiv latent variabel (Edwards, 2011).

Sammenfattende må det siges at både de konsistente fund af flere faktorer for de mest brugte skalaer, den begrænsede information om intern konsistens og endelig de netop gennemgåede fund fra netværksanalyser, udgør markante brud på antagelserne bag den refleksive latent variabel model. På denne baggrund må vi derfor sige at depression, som anskuet ved diverse skalaer og interviews, ikke kan ses som en refleksiv latent variabel.

Hvad er der af muligheder for diagnosen?

Den ovenstående gennemgang af en række udfordringer for depressionsbegrebet har to mulige årsager. Enten dækker depressionsdiagnosen over en lidelse, som meningsfuldt kan betragtes som en refleksiv latent variabel, og dette er blot ikke lykkedes i de mest anvendte

skalaer. Alternativt betegner begrebet måske et sæt af lidelser, som har flere facetter og som ikke kan ses som én refleksiv latent variabel.

Der er to mulige veje at gå med begrebet: tilgangen til begrebet kan gentænkes, idet begrebet i stedet kan ses som en formativ latent variabel eller begrebets definition kan helt gentænkes.

Den første mulighed fører til et spørgsmål, som vi kort kom ind på ovenfor. At de mest velundersøgte og mest anvendte skalaer har store udfordringer betyder ikke, at der ikke findes bedre skalaer.

Dermed kan der være bedre skalaer derude. Der er over 280 depressionsskalaer og mange af disse varierer meget i deres teoretiske og empiriske grundlag (Santor et al., 2006).

Hvis sidstnævnte mulighed er tilfældet kan svaret ligge i at nyere skalaer blot skal operationalisere begrebet anderledes.

Edwards (2011) har beskrevet en tendens til, at forskere til tider konstaterer at deres begreb ikke er internt konsistent og derfor konkluderer, at en skala ikke afspejler en refleksiv struktur. men derimod en formativ struktur. Dette er misforstået og en logisk brist 'affirming the consequent', som opstår når man konstaterer at et resultat gør sig gældende, men overser at flere forhold kan lede til dette resultat. En skala kan udvise lav intern reliabilitet som resultat af at man fejlagtigt forsøger at arbejde med et begreb som dækker over et formativt forhold. Imidlertid kan en skala også udvise lav intern reliabilitet, når der er tale om et refleksivt begreb, fordi den lave interne reliabilitet også kan være forårsaget af en svag operationalisering af begrebet (Edwards & Bagozzi, 2000; Diamantopoulos & Siguaw, 2006). Eksempelvis vil en skala for mæslinger, som fejlagtigt indeholder irrelevante items, udvise en begrænset intern reliabilitet, men dette vil være

vidne om en dårlig begrebsoperationalisering og ikke en afkræftelse af at mæslinger er en, ontologisk set, virkelig lidelse.

Det kan ikke udelukkes at en bedre og refleksiv skala måske har været lavet, men er blevet kasseret, af disse grunde. Vi forholder os dog til de skalaer, der er bliver benyttet og de ser ikke ud til at være refleksive.

Hvis depression, som brugt i skalaerne, i stedet betragtes som en formativ variabel, vil relationen mellem items ikke på samme måde udgøre et problem. I den formative variabel model kan items frit korrelere med hinanden, udenom den latente variabel, modsat af den refleksive model. Anskuet som en formativ model vil nogle af de mest centrale udfordringer derfor ikke længere være i konflikt med modellen og begrebet ville på dette punkt stå stærkere (Edwards, 2011).

For formative variabler gælder det imidlertid, at de items der måles på, antages at være uden støj og direkte afspejle det man ønsker at måle (Diamantopoulos & Siguaw 2006; Edwards, 2011). Variablen defineres derfor som summen af det der måles. Hvis depression ses som en formativ variabel har det den konsekvens, at der for forskellige depressionsmål er tale om forskellige begreber, der måles. Diagnosticeringen vil så på et teoretisk plan, være opdelt i skalaspecifik depression for eksempel Hamilton-depression og Beck-depression.

Disse to veje at gå fører til at se på, om depression kan betegnes som en naturlig art, fordi dette er centralt for at se på hvorvidt begrebet bør ses som refleksivt eller formativt.

Forudsætningen for dette er, at begrebet depression dækker over en egentlig lidelse, og dermed kan betegnes som en såkaldt naturlig art. Dette er interessant fordi diagnosen, som beskrevet i teoriafsnittet, ikke er defineret ud fra ætiologi eller patofysiologi. Vi mener en række

forhold peger mod at depression ikke er en naturlig art og derfor ikke bør ses som en refleksiv latent variabel.

Depression som en naturlig art

Vi introducerede i teoriafsnittet, under beskrivelsen af den refleksive latent variabel model, begrebet naturlig art som en art der ontologisk set, findes. Det fundamentale spørgsmål her er derfor ganske simpelt: dækker depression over et ontologisk set virkeligt fænomen? Vi vil her belyse dette fra forskellige vinkler, ved at se på tilblivelsen af diagnosen, hvad forskningen i genetiske anlæg for depression har ført til, samt hvorvidt arveligheden af depression er et argument for depression som naturlig art.

Skalaernes tilblivelse og indbyrdes heterogenitet

Fire af de fem største skalaer er som tidligere nævnt lavet før 1980, hvor DSM-III udkom. Ingen af skalaerne inkluderer alle symptomerne fra DSM-III, og flere, såsom HAM-D, adskiller sig markant fra DSM-III. Det er derfor svært at forestille sig eksempelvis både HAM-D og DSM-III vil indfange den samme refleksive latente variabel. Som nævnt i teoriafsnittet varierer de mest anvendte depressionsskalaer både i deres teoretiske grundlag, i den måde de er blevet konstrueret på og vigtigst, i items. Derfor er det måske slet ikke muligt, selvom der skulle være en afgrænset depressionslidelse, som udgør en naturlig art, at alle skalaerne måler den. Dette fører til spørgsmålet om skalaerne generelt er for indbyrdes forskellige til at det er meningsfuldt at spørge om de udgør en naturlig art.

Flere af skalaerne har indbyrdes minimale overlap. Eksempelvis har BDI og CES-D et overlap på 0,35 i det såkaldte Jaccard Index, hvor et overlap på 0,40 indikerer et moderat overlap. CES-D overlapper med HAM-D 0,26 og BDI har et overlap med HAM-D på 0,42 (Fried, 2016). Hvis de alle skulle måle den samme refleksive latente variabel, må præmissen være at den eneste latente variabel de måler, er at finde i de items de alle måler og at resten af items blot er unikt støj, der ikke korrelerer.

Den nuværende konceptualisering af depression fra DSM og ICD er i store træk baseret på et studie fra 1957, hvor forfatterne selv konkludere at deres version af depression muligvis er forkert (Cassidy et al., 1957). Som beskrevet i teoriafsnittet om diagnosens historie, er det tydeligt at valget af netop denne definition af depressionsbegrebet var mere eller mindre tilfældig. Lederen arbejdsgruppen, der skulle udvikle diagnosen, beskrev selv at diagnosen var mere politisk end videnskabelig (Leite et al., 2017).

Når vi ser på dette fundament for begrebet, virker det usandsynligt at man gennem denne proces har formået at operationalisere depressionsbegrebet perfekt eller blot korrekt i overensstemmelse med en naturlig art.

Genetik

Som nævnt tidligere peger særligt netværksanalyser på at items i depressionsskalaer ikke har samme effekt på hinanden (Borsboom, 2017; Borsboom & Cramer, 2013). Det harmonerer med fund, der viser at de forskellige depressionssymptomer har forskellige biologiske markører. Hassler et al. (2004) finder store biologiske forskelle imellem de forskellige symptomer, hvilket stemmer overens med andre studier (Kendler, Aggen & Neale, 2013). Eksempelvis finder Myung et al. (2012), at symptomerne følelse af skyld og søvnløshed har forskellige biomarkører.

På samme måde varierer sociale risikofaktorer også for de forskellige symptomer (Keller et al., 2007; Keller & Nesse, 2006).

Når vi ser på begrebet depression og ikke på symptomerne enkeltvis, er der særligt siden 1990'erne beskrevet mange fund af såkaldte kandidatgener i litteraturen. Her har Border et al. (2019) lavet den hidtil største meta-analyse, rettet mod de 18 mest undersøgte kandidatgener. Deres meta-analyse fandt, at ingen af de undersøgte gener hænger betydeligt sammen med risikoen for at udvikle depression (Border et al., 2019). Dette er særligt markant fordi studiet har ekstremt høj statistisk styrke. Videre angiver de, at deres resultater står i stærk kontrast til store dele af litteraturen, hvor der ofte rapporteres fund af langt større effekter end hvad deres studie finder belæg for. Her peger Border et al. (2019) på at litteraturen generelt er betydeligt svækket af netop lav styrke i studierne.

I den litteratur vi har gennemgået, finder vi to tendenser: a) når de enkelte depressionssymptomer undersøges findes der ikke et ensformigt mønster af gener der peger i samme retning, og b) når depression undersøges som samlet begreb er de mest opsigtsvækkende fund omkring kandidatgener måske primært et udslag af statistisk varians. Begge disse tendenser er svært forenelige med opfattelsen af depression som en naturlig art.

Arvelighed

Arveligheden af depression er interessant for denne diskussion, fordi den ofte omtales og fordi den har stærk intuitiv appel. Det kan synes oplagt, at hvis det man måler, når man måler og tester for depression, er arveligt, må depressionsbegrebet også være et ontologisk virkeligt fænomen.

Der er også publiceret flere studier, der undersøger depressions arvelighed både i forskellige populationer og som meta-analyser (McGue & Christensen, 2003; Reid, Arcese, Sardell & Keller, 2011; McMahon, 2018; Fernandez-Pujals et al., 2015). Arveligheden af depression er fastlagt og er tidligere estimeret til at være på ca. 37 % (Fernandez-Pujals et al., 2015). Imidlertid er arveligheden af begrebet ikke nødvendigvis helt så tungtvejende et argument for depressionsbegrebet, som det kan synes intuitivt. Vi vil her forsøge at vise hvorfor argumentet ikke ubetinget er berettiget.

Den klassiske måde at måle arvelighed på, er at måle kovariansen mellem individer, der deler nogle, men ikke alle deres gener, og nogen der deler noget af, men ikke hele deres miljø. Forskellen i kovariansen, kan så bruge til at måle indvirkninger af de forskellige komponenter af variansen (Griffiths, 2000)

Standardmodellen for arvelighed ser således ud:

$$X = A + D + C + S$$

X er her arveligheden af det træk man vil måle, A er den additive genetiske komponent, D er den ikke additive genetiske komponent, C er miljøvariabler der er fælles for alle afkom i en familie og S er miljøvariabler, der bidrager unikt til et individ og alt støj i målingen af trækket. Antagelsen bag denne model er at alle komponenter er fuldstændig uden indbyrdes korrelation. Arvelighed bliver typisk delt op i bred arvelighed (H^2) og snæver arvelighed (h^2). H^2 er al genetisk påvirkning af det træk man vil måle. Hvor h^2 kun er den additive genetiske komponents

påvirkning af det træk man vil måle (Fernandez-Pujals et al., 2015; Griffiths, 2000). H^2 kan udregnes som:

$$\frac{\sigma_A + \sigma_D}{\sigma_A + \sigma_D + \sigma_C + \sigma_S}$$

h^2kan udregnes som

$$\frac{\sigma_A}{\sigma_A+\sigma_D+\sigma_C+\sigma_S}$$

Fællesnævnerne er nemme at opnå ved at sample hele populationen. Det besværlige er at estimere variansen af A og D. Af denne grund laver man tvillingestudier (Griffiths, 2000). Hvis man tager enæggede tvillinger, der er opvokset sammen, og antager at de på andre punkter er identiske med resten af populationen, så opnår man kovariansen:

kov = var(A)+var(D)+var(C)

Hvis man derimod tager enæggede tvillinger, der er opvokset hver for sig, og antager at de ellers er helt ligesom resten af populationen og vigtigst, at deres miljø er komplet tilfældigt efter fødslen, får man kovariansen:

kov = var(A)+var(D)

Målet for H^2 er kovariansen af enæggede tvillinger vokset op hver for sig, divideret med den totale varians.

Arvelighed her (H^2) fortæller hvor meget depression vil variere mellem personer, der er genetisk identiske, men ellers er tilfældigt distribueret ud i en population, til sammenligning med hvor meget depression varierer på tværs af en hel population.

Arvelighedsscorer siger altså ikke hvor stor en procentdel af depression der er genetisk eller hvor stor en chance, der er for at personer har samme depressionsscore/status. En arvelighed af depression på 0,40 siger derfor ikke at 40 % af variansen er genetisk.

Hvis disse begrænsninger i estimeringen af arvelighed lægges til side og hvis vi antager at depressionsbegrebet er arveligt, vil dette fortsat ikke være et ubetinget stærkt argument for begrebet som en naturlig art. En formativ latent variabel model, som altså dækker over et socialt konstrueret begreb, kan også have en arvelighed. F.eks. har personlighedstrækket "Extraversion" en arvelighed på 0,54 (Bouchard, 2004) og højde har en arvelighed på 0,80. Det betyder at summen af "extraversion" og højde er arvelig, selvom de to træk ikke har noget med hinanden at gøre.

Det er også blevet fundet, at depressionssymptomer varierer meget i deres arvelighed, fra 0-35 H^2(Jang, Livesley, Taylor, Stein & Moon, 2004). Hvilket betyder at det kan være en lignende situationen i depressionsbegrebet.

Arveligheden af depression udgør derfor ikke et entydigt argument for depressionsbegrebet som naturlig art. Dette udgør et aspekt hvor den mulige inferens er ensidig, i det ovenstående ikke er direkte er bevis for at depression ikke skulle være en naturlig art, men blot taler imod anvendelsen af arveligheden af depression som argument.

Sammenfatning

Baggrunden for depressionsbegrebets nuværende form er præget af en betydelig tilfældighed, hvori det er særligt bemærkelsesværdigt at denne ikke alene er videnskabeligt udledt. Dette gør at vi ser det som usandsynligt at netop denne definition af begrebet skulle være i eksakt overensstemmelse med en naturlig art. Dette er i højere grad en usandsynliggørelse af ideen om depression som naturlig art end det er et direkte bevis imod ideen. Når vi ser på den netop gennemgåede diskussion af arveligheden af begrebet har vi på samme vis ikke et bevis imod ideen, men derimod et argument imod et hyppigt anvendt argument.

Forskningen i genetikken og biologien bag depressionsbegrebet vurderer vi som det stærkeste argument imod ideen. Det at symptomer varierer i biomarkører, arvelighed og risikofaktorer, taler imod at depressionsbegrebet dækker over en samlet størrelse. Samt det at kandidatgenerne for depression ser ikke ud til at kunne forklare en betydelig del af depressionsbegrebet. Alt dette peger mod, at depression ikke skal ses som en naturlig art.

Efter denne diskussion vil vi se på hvordan mål for depression stemmer overens med andre psykologiske mål, for derefter at kunne diskutere betydningen af disse to afsnit samlet.

Depression og ekstern validitet - Sammenhængen med invalideringsgrad

Indenfor forskningen bruges begrebet invalideringsgrad ofte, som et fænomen af lignende karakter med begrebet livskvalitet (Fried & Nesse, 2014). Depressionsbegrebet kan forventes at korrelere med mål såsom livskvalitet og invalideringsgrad. Dette følger naturligt af at depression er en stærkt invaliderende lidelse, som rammer en række facetter af livsførelsen (American Psychiatric Association, 2013).

Ligesom der er stor variation i genetikken bag og i risikofaktorer for depressionssymptomer, er der også stor forskel i invalideringsgraden af de forskellige symptomer. Overordnet set findes der moderate korrelationer mellem depression og invalideringsgrad (McKnight & Kashdan, 2009). Her er der tidligere argumenteret for, at denne korrelation er forventelig, men også at den er stærkt varierende og at depressionskalaer ikke nødvendigvis bør stå alene som mål for den depressives mentale tilstand (McKnight & Kashdan, 2009). Mere interessant er det imidlertid at se på sammenhængen mellem depression og invalideringsgrad på item niveau.

Der er en del forskning, som peger på at depressionssymptomer varierer i deres invalideringsgrad. Fried & Nesse (2014) fandt blandt andet en variation i invalideringsgraden af depressionssymptomer på psykosociale mål. De fandt at symptomer som lavt humør, koncentrationsbesvær, udmattelse og tab af interesse alle kunne forklare over 10 % procent af variansen i invalideringsgraden. Modsat kunne hypersomnia, insomnia, og vægtforøgelse alle kun forklare under 1,5 % af variansen. Symptomer varierer også på tværs af subdomæner. Tab af interesse har f.eks. stor indflydelse på ens sociale aktiviteter. Andre studier har fundet lignende resultater med DSM-III og psykosociale mål (Tweed, 1993). Denne heterogenitet mellem items har den afgørende betydning, at de enkelte værdier af sumscoren ikke er entydigt forbundet til én invalideringsgrad. Samlet taler disse fund derfor imod brugen af sumscores, da en HAM-D score på 20 ikke nødvendigvis er værre for en person end en HAM-D score på 15. Invalideringsgraden er afhængig af hvilke symptomer sumscoren dækker over. Sumscoren er derfor en dårlig proxy for invalideringsgraden.

Dette forhold er også interessant, fordi det vil være ideelt at kunne forbinde hvert enkelt niveau af depressiv lidelse, entydigt til ét niveau af invalideringsgrad. Tidligere anvendte vi i analysen af prævalensniveauer, nytteværdien af forskellige testresultater ved depression. Ideelt set skulle denne kunne opgøres i livskvalitet. Dette ville imidlertid forudsætte et entydigt forhold mellem depressiv lidelse og invalideringsgrad, fordi nytteværdien kræver en enhed at opgøres i. De gentagne fund af heterogenitet i invalideringsgraden af forskellige depressionssymptomer, indikerer at der ikke er grundlag for på nuværende tidspunkt at opgøre nytteværdien og relaterede begreber på grundlag af depressionsskalaer. Dette forhold begrænser på den ene side muligheden for at anvende nytteetisk-orienterede estimater, fordi disse ikke kan baseres på de ideelle enheder, såsom livskvalitet eller invalideringsgrad. Dette førte også til, at vi i analysen anvendte den simple distinktion mellem hvorvidt det enkelte individ var deprimeret eller ikke deprimeret. Dette betegnes som en 'threshold loss function' (Smits et al., 2007). Omvendt peger dette på relevansen af at faktorere andre parametre ind i sådanne analyser, såsom den økonomiske vinkel vi anlagde i analysen af prævalens.

Den differentierede invalideringsgrad betyder ligesom det foregående afsnit om arvelighed og genetik gør, at man ikke bør se depression som en naturlig art. Forskellen i invalideringsgrad betyder også at man mister en masse information ved at se på sumscores fremfor de individuelle symptomer.

Konsekvenserne af at det ikke er en refleksiv latent variabel

De vigtigste konklusioner fra de forudgående to afsnit er: a) depression må anskues som en formativ og ikke en refleksiv variabel, hvilket betyder at forskellige skalaer og interviews for depression retter sig mod forskellige depressionsbegreber, og b) forskellige depressionsbegreber

kan ikke forventes at have samme valens, fordi de enkelte symptomer viser sig at være heterogene når de holdes op mod eksterne begreber.

Fundene er interessante, fordi etableringen af substantiv validitet inden man har fastlagt begrebsvaliditeten, betyder man risikerer at have opsamlet en vidensbase, der pludselig er ubrugelig (MacKenzie et al., 2005). Vi vil her se på betydningen af disse to fund, i forskellige kontekster.

Forskning

Set fra et forskningsorienteret perspektiv bringer disse fund helt åbenlyse problemer. Et område som i nyere tid er vokset eksponentielt, er forskningen i ændrede hjernestrukturer og neurofysiologiske forhold i forbindelse med depression (Wonch et al., 2016; Huang et al., 2017; Gramer et al., 2014). Eksempelvis undersøges det om depression bevirker ændringer i hjernen der er synlige ved FMRI-scanninger, hvor ændringer i størrelsen på hippocampus er et ofte gengivet eksempel (Schmaal et al., 2016). Det er angivet både i fagbøger og populærvidenskabelige kilder, at der er tydeligt synlige forskelle i hjernerne hos depressive individer (Hildebrandt, 2016). Dette er bl.a. brugt som argument for vigtigheden af hurtig behandling af depression. Imidlertid har nyere forskning sået tvivl om klarheden af denne effekt, ved at nå skiftende konklusioner om de strukturelle ændringer. Således har der været fund både af at depressive har større hypothalamus, og af at depressive har mindre hypothalamus (Fried, 2015). Foruden en åbenlys risiko for at sådanne variationer må tilskrives brugen af for små sample sizes, er sådanne fund absolut aktuelle for vores emne her.

Denne type projekter hvor man søger at se sammenhænge mellem strukturer på neurologisk niveau og depressive versus ikke depressive, vil logisk være hæmmet hvis der ikke

er klarhed om hvorvidt det enkelte individ er deprimeret eller ej. En indvending kan her være, at det enkelte studie typisk gennemgående vil klassificere deltagerne ved samme operationalisering af depression og studier derfor ikke vil være eksponeret for sådan variation imellem deltagerne. Selv i dette tilfælde vil resultatet dog være at de sammenhænge man måtte finde, er optimeret til netop én operationalisering af depressionsbegrebet, med en tvivlsom generaliserbarhed til følge. Som et andet eksempel kan serotoninhypotesen nævnes. I over 50 år har man antaget at depression skyldes et fald i neurotransmitteren serotonin hos deprimerede En hypotesen der stadig ikke er blevet bekræftet (Cowen & Browning, 2015).

Brugen af sumscores – i klinisk og i forskningsmæssig sammenhæng

Sumscores bliver brugt som en proxy for sværhedsgraden af depression og forskelle i sumscores over tid (interventionsforskning) bliver forstået som ændringer i det underliggende depressionsbegreb. Denne anvendelse af sumscoren for depression bruges på en række områder.

Sumscores bliver ofte brugt til at korrelere med andre mål eller vurdere effekten af en intervention. Det giver kun mening, hvis depressionscores rent faktisk er en proxy for sværhedsgraden af depression. En ændring i sumscores fra f.eks. 10 til 5, giver kun mening, hvis de to scores måler det samme. Forudsætningerne for dette er, at en skala kun måler en ting og at det begreb man måler er ens over tid.

Hvis skalaerne er formative, giver korrelationerne et større problem, for hvis ikke der er støj i item-scorer er forskellen mellem testene, forskelle i begreberne. Korrelationen mellem BDI-II og HAM-D på 0,66-0,75 (Yang & Gorenstein, 2013) betyder at skalaerne kun overlapper

43,6-56,3 %, hvilket giver store problemer for overførbarheden af resultater på de to skalaer og gør det nødvendigt at gøre resultater unikke for skalaerne.

Interventionsforskning

Et område hvor brugen af sumscores for depressionsskalaerne har vist sig særligt problematisk, er forskning i virkningen af SSRI-præparater.

Forskning i effekten af SSRI-præparater udgør et interessant eksempel på vigtigheden af, hvorvidt det er berettiget at se et begreb, som en refleksiv latent variabel og derfor at måle den med en sumscore. Forskningsstudier rettet mod SSRI-præparater benytter næsten udelukkende HAM-D skalaen (Bagsby, Ryder, Schuler & Marshall, 2004). Meta-analyser af SSRI-præparater har nået varierende konklusioner, hvilket har ført til fortsatte diskussioner, ikke bare om graden af effekt, men også om tilstedeværelsen af en effekt af SSRI-præparater (Kirsch, Deacon, Huedo-Medina, Scoboria, Moore & Johnson & 2008; Cipriani et al., 2018)

Hieronymus, Emilsson, Nilsson & Eriksson (2016) reanalyserede data fra SSRI-studier i en metanalyse, hvor de udelukkende anvendte det første item fra HAM-D skalaen, nedtrykt humør. Studiet inkluderede 6669 deltagere og fandt som hovedfund at 29 ud af 32 studier fandt en signifikant behandlingseffekt med dette modererede begreb, mod 18 ud af 32 ved den fulde HAM-D skala. Af yderligere interesse, fandt de at SSRI-præparater har en markant forskelligartet indvirkning på forskellige items i HAM-D. Det er netop denne heterogenitet, antitesen til tau ækvivalens, som er årsag til at Hieronymous et al. (2016) finder et markant anderledes resultat. SSRI har en statistisk signifikant effekt over placebo på item ”sad mood”, men har en negativ effekt på items som ”loss of weight” ”genitale symptoms”. Dette påviser et åbenlys paradoks; flere af bivirkningerne af SSRI’er optræder som items i HAM-D skalaen.

Tilstedeværelsen af bivirkninger kan derfor influere sumscoren, fordi de indgår i modellen som proxier for depressionsbegrebet. Fried (2015) viser at flere HAM-D symptomer overlapper med de typisk bivirkninger af forskellige SSRI-præparater. Det har stor betydning for fortolkningen af effekten af SSRI'er og kan føre til uhensigtsmæssige medicinske beslutninger.

Effekten af diagnosen – lever den op til dens formål?

Hvis vi skal belyse diagnosens berettigelse, må vi spørge om diagnosen lever op til dens formål. Vi mener at det altoverskyggende formål med diagnosen må være at fastsætte en behandlingsplan for patienten og medvirker til at optimere behandlingen.

Fundamentet for dette må være at diagnosen indeholder brugbar information til den videre behandling. Det bringer to spørgsmål op: er behandlingen for depression særligt adskilt fra andre psykiske lidelser? Og i hvor høj grad kan depression behandles?

Den almindelige medicinske behandling for depression er SSRI præparater. Disse bruges imidlertid også i vid udstrækning til angst behandling og behandling af andre psykiatriske problemstillinger (Cassano, Baldini Rossi & Pini, 2002; Mao & Zhang, 2015).

Det har henover en årrække været omdiskuteret ikke kun i hvilken grad disse virker, men også hvorvidt de virker (Kirsch et al., 2008). Store meta-analyser har fundet forskellige resultater og har varieret i deres inklusionskriterier. Det er nævneværdigt at SSRI som oftest er førstevalget, repræsenterer en tilgang som er uændret henover de sidste 30 år (Fried, 2015)

Hvordan ser effekten ud for psykoterapi?

Det er vores indtryk at mange kliniske psykologer i deres dagligdag ikke arbejder ud fra diagnoser (Mitchell & Coyne, 2010). Hele tilgangen med at bruge diagnoser er primært udbredt i

psykiatrien, frem for i den kliniske psykologi. Disse forhold mindsker i sig selv relevansen for de patienter som behandles uden for psykiatrien, hvilket som tidligere nævnt er estimeret til at være 90 % (Davidsen, 2009). Forskning har vist at, hvis man giver depressive patienter, psykoterapi specifikt udviklet til angstpatienter, finder man samme effekt som, hvis man giver dem psykoterapi udviklet til depressive patienter (Weitz, Kleiboer, van Straten & Cuijpers, 2018). Det kan dog skyldes de begrænsninger i brugen af sumscores vi tidligere har nævnt.

Det fører til spørgsmålet: er det behandlingen, der ikke virker eller er det målingen der ikke kan indfange den?

Afslutning på diskussionen

I vores diskussion af hvorvidt at skalaerne kan siges at indfange en refleksiv latent variabel, fandt vi at det ikke var tilfældet. Når det fund isoleres, er der to mulige årsager: den ene er en grundlæggende begrænsning i depressionsbegrebet, den anden mulige årsag er at de mest anvendte skalaer ikke lykkes med at operationalisere begrebet. Imidlertid pegede vores videre diskussion af depressionsbegrebet på at dette ikke skal ses som en naturlig art. Dette taget i betragtning er det ubetinget en fejl at se begrebet som en refleksiv latent variabel.

Begrebet må derfor enten ses som en formativ latent variabel eller alternativt helt gentænkes. Dette forhold og den manglende afspejling i litteraturen, ser vi som det primære fund i dette speciale og en betydelig udfordring for de psykologiske egenskaber af depressionsskalaer og for synet på depressionsbegrebet i det hele taget.

Vi mener det bør være den primære agenda i fremtidig forskning at besvare de helt fundamentale spørgsmål, da det er grundlaget for alle udledninger fra skalaerne.

Konklusion

I analysens første del fokuserede vi på indvirkningen af prævalensen i de enkelte studier for CES-D skalaen. Her fandt vi at når vi anvendte nytteværdi som kriterium for at optimere cut-off værdien, var der en betydelig sammenhæng mellem prævalens og optimalt cut-off.

I analysens anden del estimerede vi det optimale cut-off for CES-D skalaen ud fra en ny metodisk tilgang til meta-analyse og fandt et resultat der divergerede markant fra den originale meta-analyse af de samme studier.

I analysens tredje del sammenlignede vi de to kliniske interviews CIDI og SCID. Her fandt vi en betydelig forskel i det optimale cut-off afhængigt af hvilket interview studierne anvendte som guldstandard.

Særligt de første to analyser indikerer at der sandsynligvis er en del information der går tabt ved anvendelsen af bivariate modeller til meta-analyse bestemmelser af optimale cut-offs.

Analysens tredje pegede på at det optimale cut-off ikke kan optimeres uden at tage operationaliseringen af guldstandarden i betragtning. Dette markerede overgangen til analysens fjerde del og den efterfølgende diskussion.

Fjerde del af analysen fokuserede på brugen af faktoranalyser, mål for den interne konsistens og i hvilken grad mulige brud på disse blev omtalt i valideringsstudier for BDI-ll, belyst ved 30 studier udvalgte som repræsentative. Denne analyse viste, at samtlige af disse studier undlod at omtale betydningen af mulige brud på antagelserne bag deres statistiske analyser.

I diskussionen fulgte vi indledningsvist disse analyser op med en diskussion af hvorvidt depressionsskalaer måler en refleksiv latent variabel. Ud fra diskussionen nåede vi frem til konklusionen, at de gennemgåede fund taler stærkt imod at dette er tilfældet.

Dernæst diskuterede vi hvorvidt depressionsbegrebet udgør en naturlig art. Ud fra baggrunden for diagnosen, gyldigheden af arveligheden af begrebet som argument samt genetikken for depression, fandt vi at der dels ikke er stærke argumenter for at se begrebet som en naturlig art, og at der modsat, særligt ved det genetiske aspekt, er klare indikationer af at begrebet ikke er en afgrænset ontologisk størrelse.

Litteraturliste

Allen, J. G. (1998). User's guide for the structured clinical interview for DSM-IV axis II personality disorders: SCID-II. Bulletin of the Menninger Clinic, 62(4), 547.

American Psychiatric Association, & American Psychiatric Association DSM-5 Task Force. (2013). Diagnostic and statistical manual of mental disorders: DSM-5 (5.th ed.). Arlington, Va: American Psychiatric Association.

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The hamilton depression rating scale: Has the gold standard become a lead weight? American Journal of Psychiatry, 161(12), 2163-2177. doi:10.1176/appi.ajp.161.12.2163

Ban TA. 2014. From melancholia to depression. A history of diagnosis and treatment. Retrieved
from http://inhn.org/home.html (2019, Marts 10).
Bech P Licht R W Stage K B Abildgaard W Bech-Andersen G Søndergaard S Martiny K

Bech, P. & Coppen, A. (1990). The hamilton scales. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-75373-2

Bech, P., Allerup, P., Maier, W., Albus, M., Lavori, P., & Ayuso, J. (1992). The hamilton scales and the hopkins symptom checklist (SCL-90). A cross- national validity study in patients with panic disorders. The British Journal of Psychiatry, 160(2), 206-211. doi:10.1192/bjp.160.2.206

Bech, P., Coppen, A., & SpringerLink (Online service). (1990). The hamilton scales. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-75373-2

Bech, P., Gram, L. F., Dein, E., Jacobsen, O., Vitger, J., & Bolwig, T. G. (1975). QUANTITATIVE
RATING OF DEPRESSIVE STATES: Correlation between clinical assessment, beck's self-rating scale and hamilton's objective rating scale. Acta Psychiatrica Scandinavica, 51(3), 161-170. doi:10.1111/j.1600-0447.1975.tb00002.x

Bech, P., Licht R. W., Stage K. B., Abildgaard, W., Bech-Andersen, G., Søndergaard, S., Martiny, K. (2005). Rating Scales for affektive lidelser. Retrieved from https://www.psykiatri-regionh.dk/CCMH/Rating-scales-og spoergeskemaer/Documents/DenblaabogRatingscalesforaffektivelidelser.pdf (2019, April 4)

Bech, P., Timmerby, N., Martiny, K., Lunde, M., & Soendergaard, S. (2015). Psychometric evaluation of the major depression inventory (MDI) as depression severity scale using the LEAD (longitudinal expert assessment of all data) as index of validity. BMC Psychiatry, 15(1), 190. doi:10.1186/s12888-015-0529-3

Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961). "An inventory for measuring depression". Arch. Gen. Psychiatry. 4 (6): 561–71

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. Journal of Consulting and Clinical Psychology, 56(6), 893-897. doi:10.1037/0022-006X.56.6.893

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. Journal of Consulting and Clinical Psychology, 56(6), 893-897. doi:10.1037/0022-006X.56.6.893

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of beck depression inventories-IA and-II in psychiatric outpatients. Journal of Personality Assessment, 67(3), 588-597. doi:10.1207/s15327752jpa6703_13

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: Receiver operating characteristics curves. Critical Care, 8(6), 508-512. doi:10.1186/cc3000

Biskin, R. S., & Paris, J. (2012). Diagnosing borderline personality disorder. CMAJ : Canadian Medical Association Journal = Journal De l'Association Medicale Canadienne, 184(16), 1789-1794. doi:10.1503/cmaj.090618

Blashfield, R. K., & Herkov, M. J. (1996). Investigating clinician adherence to diagnosis by criteria: A replication of morey and ochoa (1989). Journal of Personality Disorders, 10(3), 219-228. doi:10.1521/pedi.1996.10.3.219

Blashfield, R. K., Keeley, J. W., Flanagan, E. H., & Miles, S. R. (2014). The cycle of classification: DSM-I through DSM-5. Annual Review of Clinical Psychology, 10(1), 25-51. doi:10.1146/annurev-clinpsy-032813-153639

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110(2), 305-314. doi:10.1037//0033-2909.110.2.305

Borsboom, D. (2017). A network theory of mental disorders. World Psychiatry, 16(1), 5-13. doi:10.1002/wps.20375

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. Annual Review of Clinical Psychology, 9(1), 91-121. doi:10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity.Psychological Review, 111(4), 1061-1071. doi:10.1037/0033-295X.111.4.1061

Bouchard, T. J. (2004). Genetic influence on human psychological traits: A survey. Current Directions in Psychological Science, 13(4), 148-151. doi:10.1111/j.0963-7214.2004.00295.x

Cappelleri, J. C., PhD, Jason Lundy, J., PhD, & Hays, R. D., PhD. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clinical Therapeutics, 36(5), 648-662. doi:10.1016/j.clinthera.2014.04.006

Carroll, B., Feinberg, M., Smouse, P., Rawson, S., & Greden, J. (1981). The Carroll Rating Scale for Depression I. Development, Reliability and Validation. British Journal of Psychiatry, 138(3), 194-200. doi:10.1192/bjp.138.3.194

Cassidy, W. L., Flanagan, N. B., Spellman, M., & Cohen, M. E. (1957). clinical observations in manic-depressive disease: A quantitative study of one hundred manic-depressive patients and fifty medically sick controls. Journal of the American Medical Association, 164(14), 1535-1546. doi:10.1001/jama.1957.02980140011003

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood.Organizational Research Methods, 18(2), 207-230. doi:10.1177/1094428114555994

Cho, M. J., Mościcki, E. K., Narrow, W. E., Rae, D. S., Locke, B. Z., & Regier, D. A. (1993). Concordance between two measures of depression in the hispanic health and nutrition examination survey. Social Psychiatry and Psychiatric Epidemiology, 28(4), 156-163. doi:10.1007/BF00797317

Christensen, K. S., Oernboel, E., Nielsen, M. G., & Bech, P. (2019). Diagnosing depression in primary care: A rasch analysis of the major depression inventory. Scandinavian Journal of Primary Health Care, 37(1), 105-112. doi:10.1080/02813432.2019.1568703

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. Journal of Business Research, 61(12), 1250-1262. doi:10.1016/j.jbusres.2008.01.013

Cowen, P. J., & Browning, M. (2015). What has serotonin to do with depression? World Psychiatry, 14(2), 158-160. doi:10.1002/wps.20229

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334. doi:10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests.Psychological Bulletin, 52(4), 281-302. doi:10.1037/h0040957

Cuijpers, P., Reijnders, M., Karyotaki, E., de Wit, L., & Ebert, D. D. (2018). Negative effects of psychotherapies for adult depression: A meta-analysis of deterioration rates. Journal of Affective Disorders, 239, 138-145. doi:10.1016/j.jad.2018.05.050

DeVellis, R. F. (2006). Scale development: Theory and applications. Newbury Park, Calif: Sage.

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. British Journal of Management, 17(4), 263-282. doi:10.1111/j.1467-8551.2006.00500.x

Edwards, B. C., Lambert, M. J., Moran, P. W., Mccully, T., Smith, K. C., & Ellingson, A. G. (1984). A meta-analytic comparison of the Beck Depression Inventory and the Hamilton Rating Scale for Depression as measures of treatment outcome. British Journal of Clinical Psychology,23(2), 93-99. doi:10.1111/j.2044-8260.1984.tb00632.x

Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. Organizational Research Methods, 4(2), 144-192. doi:10.1177/109442810142004

Edwards, J. R. (2011). The fallacy of formative measurement. Organizational Research Methods, 14(2), 370-388. doi:10.1177/1094428110378369

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. Psychological Methods, 5(2), 155-174. doi:10.1037/1082-989X.5.2.155

Evans, D. L., Herbert, J. D., Nelson-Gray, R. O., & Gaudiano, B. A. (2002). Determinants of diagnostic prototypicality judgments of the personality disorders. Journal of Personality Disorders, 16(1), 95-106. doi:10.1521/pedi.16.1.95.22554

Faravelli, C., Albanesi, G., & Poli, E. (1986). Assessment of depression: A comparison of rating scales. Journal of Affective Disorders, 11(3), 245-253. doi:10.1016/0165-0327(86)90076-5

Faravelli, C., Servi, P., Arends, J. A., & Strik, W. K. (1996). Number of symptoms, quantification, and qualification of depression. Comprehensive Psychiatry, 37(5), 307-315. doi:10.1016/S0010-440X(96)90011-5

Feighner JP , Robins E , Guze SB , Woodruff RA , Winokur G , Munoz R : Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry 1972; 26:57–63

Fernandez-Pujals, A. M., Adams, M. J., Thomson, P., McKechanie, A. G., Douglas H R Blackwood, Smith, B. H., . . . McIntosh, A. M. (2015). Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in generation scotland: Scottish family health study (GS:SFHS). PLoS One, 10(11), e0142197. doi:10.1371/journal.pone.0142197

Fielder, K. & von Sydow, M. (2015). Heuristics and biases: Beyond Tversky and Kahneman's (1974) judgment under uncertainty. In: M. W. Eysenck & D. Groome. Cognitive Psychology: Revisiting the Classical Studies. Los Angeles, London: Sage., 146–161.

Fleiner, T., Dauth, H., Gersie, M., Zijlstra, W., & Haussermann, P. (2017). Structured physical exercise improves neuropsychiatric symptoms in acute dementia care: A hospital-based RCT. Alzheimers Research & Therapy, 9(1), 68-9. doi:10.1186/s13195-017-0289-z

First MB, Williams JBW, Karg RS, Spitzer RL: Structured Clinical Interview for DSM-5 Disorders, Clinician Version (SCID-5-CV). Arlington, VA, American Psychiatric Association, 2016

Fried, E. I. (2016). Are more responsive depression scales really superior depression scales?Journal of Clinical Epidemiology, 77, 4-6. doi:10.1016/j.jclinepi.2016.05.004

Fried, E. I. (2016). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. Journal of Affective Disorders, 208, 191-197. doi:10.1016/j.jad.2016.10.019

Fried, E. I. (2017). What are psychological constructs? on the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. Health Psychology Review, 11(2), 130-134. doi:10.1080/17437199.2017.130671

Fried, E. I., & Nesse, R. M. (2014). The impact of individual depressive symptoms on impairment of psychosocial functioning. Plos One, 9(2), e90311. doi:10.1371/journal.pone.0090311

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. Journal of Affective Disorders, 189, 314-320. doi:10.1016/j.jad.2015.09.005

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . or not?: Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression.Psychological Assessment, 28(11), 1354-1367. doi:10.1037/pas0000275

Garb, H. N. (1996). The representativeness and past-behavior heuristics in clinical judgment.Professional Psychology: Research and Practice, 27(3), 272-277. doi:10.1037/0735-7028.27.3.272

Garb, H. N. (2005). Clinical judgment and decision making. Annual Review of Clinical Psychology, 1(1), 67-89. doi:10.1146/annurev.clinpsy.1.102803.143810

Gauron EF, Dickinson JK. Diagnostic decision making in psychiatry. Arch Gen Psychiatry. 1966;14:225–232

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the patient health questionnaire (PHQ): A diagnostic meta-analysis.Journal of General Internal Medicine, 22(11), 1596-1602. doi:10.1007/s11606-007-0333-y

Gramer, M., Feuerstein, D., Steimers, A., Takagaki, M., Kumagai, T., Sué, M., . . . Graf, R. (2014). Device for simultaneous positron emission tomography, laser speckle imaging and RGB reflectometry: Validation and application to cortical spreading depression and brain ischemia in rats. Neuroimage, 94, 250-262. doi:10.1016/j.neuroimage.2014.03.027

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. Psychometrika, 74(1), 155-167. doi:10.1007/s11336-008-9099-3

Greenberg, G. (2010). Manufacturing depression: The secret history of a modern disease. London: Bloomsbury.

Gregory, R. J. (2011). Psychological testing: History, principles, and applications (6., International ed.). Boston, Mass: Allyn & Bacon.

Griffiths, A. J. F. (2000). An introduction to genetic analysis (7. 3.printing ed.). New York: W. H. Freeman.

HACKNEY, A. C., LANE, A. R., REGISTER-MIHALIK, J., & O'LEARY, C. B. (2017). Endurance exercise training and male sexual libido. Medicine & Science in Sports & Exercise, 49(7), 1383-1388. doi:10.1249/MSS.0000000000001235

HAMILTON M. (1960). A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry, 23(1), 56–62. doi:10.1136/jnnp.23.1.56

Hamilton, M. (1967). Development of a Rating Scale for Primary Depressive Illness. British Journal of Social and Clinical Psychology, 6(4), 278-296. doi:10.1111/j.2044-8260.1967.tb00530.x

Hasler, G., Drevets, W. C., Manji, H. K., & Charney, D. S. (2004). Discovering endophenotypes for major depression. Neuropsychopharmacology, 29(10), 1765-1781. doi:10.1038/sj.npp.1300506

Hieronymus, F., Emilsson, J. F., Nilsson, S., Eriksson, E., Department of Mathematical Sciences, Mathematical Statistics, Naturvetenskapliga fakulteten, . . . Institute of Neuroscience and Physiology, Department of Pharmacology. (2016). Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression. Molecular Psychiatry, 21(4), 523-530. doi:10.1038/mp.2015.53

Hildebrandt, S. (2016, May 25). Depression kan skade hjernen. Retrieved May 20, 2019, from https://videnskab.dk/krop-sundhed/depression-kan-skade-hjernen

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. Educational and Psychological Measurement, 60(4), 523-531. doi:10.1177/00131640021970691

Horwitz, Allan & C Wakefield, Jerome & Lorenzo-Luaces, Lorenzo. (2016). History of Depression. 10.1093/oxfordhb/9780199973965.013.2.

Huang, J., Zeng, C., Xiao, J., Zhao, D., Tang, H., Wu, H., & Chen, J. (2017). Association between depression and brain tumor: A systematic review and meta-analysis. Oncotarget, 8(55), 94932. doi:10.18632/oncotarget.19843

Irwing, F. P., Booth, T. W., & Hughes, D. J. (2018). The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test development (First ed.). Hoboken: Wiley.

Jang, K. L., Livesley, W. J., Taylor, S., Stein, M. B., & Moon, E. C. (2004). Heritability of individual depressive symptoms. Journal of Affective Disorders, 80(2), 125-133. doi:10.1016/S0165-0327(03)00108-3

Kahneman, D. (2011). Thinking, fast and slow. London: Allen Lane.

Kahneman, D. (2014). At tænke - hurtigt og langsomt (2. udgave ed.). Kbh.: Lindhardt og Ringhof.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree.American Psychologist, 64(6), 515-526. doi:10.1037/a0016755

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20(1), 141-151. doi:10.1177/001316446002000116

Keller, G., & Gaciu, N. (2015). Managerial statistics (Europe, Middle East and Africa ed.). Andover: Cengage Learning.

Keller, M. C., & Nesse, R. M. (2006). The evolutionary significance of depressive symptoms: Different adverse situations lead to different depressive symptom patterns. Journal of Personality and Social Psychology, 91(2), 316-330. doi:10.1037/0022-3514.91.2.316

Keller, M. C., Neale, M. C., & Kendler, K. S. (2007). Association of different adverse life events with distinct patterns of depressive symptoms. American Journal of Psychiatry, 164(10), 1521-1529. doi:10.1176/appi.ajp.2007.06091564

Kendell R. E., (1973). Psychiatric diagnoses: A study of how they are made. Br J Psychiatry. 122:437–445.

Kendell, R. E. (1976). The classification of depressions: A review of contemporary confusion.The British Journal of Psychiatry : The Journal of Mental Science, 129(1), 15-28. doi:10.1192/bjp.129.1.15

Kendler, K. S., Aggen, S. H., & Neale, M. C. (2013). Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. JAMA Psychiatry, 70(6), 599-607. doi:10.1001/jamapsychiatry.2013.751

Kendler, K. S., Muñoz, R. A., & Murphy, G. (2010). The development of the feighner criteria: A historical perspective. American Journal of Psychiatry, 167(2), 134-142. doi:10.1176/appi.ajp.2009.09081155

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication.Archives of General Psychiatry, 62(6), 617-627. doi:10.1001/archpsyc.62.6.617

Kim, N. S., & Ahn, W. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. Journal of Experimental Psychology: General, 131(4), 451-476. doi:10.1037/0096-3445.131.4.451

Kimbrel, N. A., Meyer, E. C., DeBeer, B. B., Gulliver, S. B., & Morissette, S. B. (2016). A 12-month prospective study of the effects of PTSD-depression comorbidity on suicidal behavior in Iraq/Afghanistan-era veterans. Psychiatry Research, 243, 97-99. doi:10.1016/j.psychres.2016.06.011

Krijnen, W. P. (2004). Positive loadings and factor correlations from positive covariance matrices.Psychometrika, 69(4), 655-660. doi:10.1007/BF02289861

Krogh, J., Hjorthøj, C., Speyer, H., Gluud, C., & Nordentoft, M. (2017). Exercise for patients with major depression: A systematic review with meta-analysis and trial sequential analysis.BMJ Open, 7(9), e014820. doi:10.1136/bmjopen-2016-014820

Kumar, A., & Dillon, W. R. (1987). Some further remarks on measurement-structure interaction and the unidimensionality of constructs. Journal of Marketing Research, 24(4), 438-444. doi:10.1177/002224378702400413

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain, 8(6), 221-223. doi:10.1093/bjaceaccp/mkn041

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., nstitutionen för neurovetenskap och fysiologi. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature Genetics, 45(9), 984-994. doi:10.1038/ng.2711

Leite, R., Macedo, P., Borges, J., & Santos, T. (2017). Depression across DSM and ICD editions: Psychiatric nosology's 'Black dog'. European Psychiatry, 41, S461-S461. doi:10.1016/j.eurpsy.2017.01.509

Leonhardt, D. (2016). THE UNDOING PROJECT: A friendship that changed our minds. New York: New York Times Company.

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. Perspectives on Psychological Science, 9(4), 355-387. doi:10.1177/1745691614535216

Lorenz, T. A., & Meston, C. M. (2012). Acute exercise improves physical sexual arousal in women taking antidepressants. Annals of Behavioral Medicine, 43(3), 352-361. doi:10.1007/s12160-011-9338-1

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. Journal of Applied Psychology, 90(4), 710-730. doi:10.1037/0021-9010.90.4.710

Maier, W., Philipp, M., Heuser, I., Schlegel, S., Buller, R., & Wetzel, H. (1988). Improving depression severity assessment—I. reliability, internal validity and sensitivity to change of three observer depression scales. Journal of Psychiatric Research, 22(1), 3-12. doi:10.1016/0022-3956(88)90022-2

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the patient health questionnaire (PHQ-9): A meta-analysis. Canadian Medical Association Journal, 184(3), E191-E196. doi:10.1503/cmaj.110829

Mayes, R., & Horwitz, A. V. (2005). DSM-III and the revolution in the classification of mental illness. Journal of the History of the Behavioral Sciences, 41(3), 249-267. doi:10.1002/jhbs.20103

McDowell, I. (2006;2009;). Measuring health: A guide to rating scales and questionnaires(3.;Third; ed.). New York: Oxford University Press. doi:10.1093/acprof:oso/9780195165678.001.0001

McGue, M., & Christensen, K. (2003). The heritability of depression symptoms in elderly danish twins: Occasion-specific versus general effects. Behavior Genetics, 33(2), 83-93. doi:10.1023/A:1022545600034

McKnight, P. E., & Kashdan, T. B. (2009). The importance of functional impairment to mental

health outcomes: A case for reassessing our goals in depression treatment research. Clinical Psychology Review, 29(3), 243-259. doi:10.1016/j.cpr.2009.01.005

McMahon, F. J. (2018). Population-based estimates of heritability shed new light on clinical features of major depression. American Journal of Psychiatry, 175(11), 1058-1060. doi:10.1176/appi.ajp.2018.18070789

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. Psychological Methods, 23(3), 412-433. doi:10.1037/met0000144

Miller, P. R., Dasher, R., Collins, R., Griffiths, P., & Brown, F. (2001). Inpatient diagnostic assessments: 1. accuracy of structured vs. unstructured interviews. Psychiatry Research, 105(3), 255-264. doi:10.1016/S0165-1781(01)00317-1

Mitchell, A. J., & Coyne, J. C. (2010). Screening for depression in clinical practice: An evidence-based guide. Oxford: Oxford University Press.

Møller Pedersen, K. (2013). sundhedsøkonomi. København: Munksgaard.

Montgomery S. A. & Asberg M. (1979). "A new depression scale designed to be sensitive to change". British Journal of Psychiatry. 134 (4): 382–89.

Morey, L. C., & Benson, K. T. (2016). An investigation of adherence to diagnostic criteria, revisited: Clinical diagnosis of the DSM-IV/DSM-5 section II personality disorders. Journal of Personality Disorders, 30(1), 130.

Morey, L. C., & Benson, K. T. (2016). An investigation of adherence to diagnostic criteria, revisited: Clinical diagnosis of the DSM-IV/DSM-5 section II personality disorders. Journal of Personality Disorders, 30(1), 130.

Myung, W., Song, J., Lim, S., Won, H., Kim, S., Lee, Y., . . . Kim, D. K. (2012). Genetic association study of individual symptoms in depression. Psychiatry Research, 198(3), 400-406. doi:10.1016/j.psychres.2011.12.037

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3.th ed.). New York: McGraw-Hill.

Okun, A., Stein, R. E. K., Bauman, L. J., & Silver, E. J. (1996). Content validity of the psychiatric symptom index, CES-depression scale, and state-trait anxiety inventory from the perspective of DSM-IV. Psychological Reports, 79(3), 1059-1069. doi:10.2466/pr0.1996.79.3.1059

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). Making sense of factor analysis: The use of factor analysis for instrument development in health care research. Thousand Oaks, Calif;London;: SAGE.

Pérez-Stable, E. J., Miranda, J., Muñoz, R. F., & Ying, Y. W. (1990). Depression in medical outpatients. underrecognition and misdiagnosis. Archives of Internal Medicine, 150(5), 1083-1088. doi:10.1001/archinte.150.5.1083

Pincus, H., Davis, W., & McQueen, L. (1999). 'subthreshold' mental disorders. A review and synthesis of studies on minor depression and other 'brand names'. The British Journal of Psychiatry, 174(4), 288-296. doi:10.1192/bjp.174.4.288

Prusoff, B. A., Klerman, G. L., & Paykel, E. S. (1972). Concordance between clinical assessments and patients' self-report in depression. Archives of General Psychiatry, 26(6), 546-552. doi:10.1001/archpsyc.1972.01750240058009

Quilty, L. C., Robinson, J. J., Rolland, J., Fruyt, F. D., Rouillon, F., & Bagby, R. M. (2013). The structure of the montgomery-åsberg depression rating scale over the course of treatment for depression: Structure of the MADRS over treatment. International Journal of Methods in Psychiatric Research, , n/a. doi:10.1002/mpr.1388

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1(3), 385-401. doi:10.1177/014662167700100306

Rector, N. A., Szacun-Shimizu, K., & Leybman, M. (2007). Anxiety sensitivity within the anxiety disorders: Disorder-specific sensitivities and depression comorbidity. Behaviour Research and Therapy, 45(8), 1967-1975. doi:10.1016/j.brat.2006.09.017

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the united states and canada, part II: Test-retest reliability of selected categorical diagnoses. American Journal of Psychiatry, 170(1), 59-70. doi:10.1176/appi.ajp.2012.12070999

Reid, J. M., Arcese, P., Sardell, R. J., Keller, L. F. (2011). Additive genetic variance, heritability, and inbreeding depression in male extra-pair reproductive success. The American Naturalist, 177(2), 177-187. doi:10.1086/657977

Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology, 58(10), 982-990. doi:10.1016/j.jclinepi.2005.02.022

Rens van de Schoot Schoot, Hox, Moerbeek, M. M., J, J., Mirjam, & de, R. v. (2017). Multilevel analysis: Techniques and applications, third edition (Third ed.). Milton: Routledge Ltd

Riley, R. D., Price, M. J., Jackson, D., Wardle, M., Gueyffier, F., Wang, J., . . . White, I. R. (2015). Multivariate meta-analysis using individual participant data: Multivariate meta-analysis using IPD. Research Synthesis Methods, 6(2), 157-174. doi:10.1002/jrsm.1129

Robins, L. N., Wing, J., Wittchen, H. U., Helzer, J. E., Babor, T. F., Burke, J., . . . Towle, L. H. (1988). The composite international diagnostic interview: An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures.Archives of General Psychiatry, 45(12), 1069-1077. doi:10.1001/archpsyc.1988.01800360017003

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Rush AJ1, Carmody TJ, Ibrahim HM, Trivedi MH, Biggs MM, Shores-Wilson K, Crismon ML, Toprac MG, Kashner TM.

Rush, A. J., Carmody, T. J., Ibrahim, H. M., Trivedi, M. H., Biggs, M. M., Shores-Wilson, K., . . .

Kashner, T. M. (2006). Comparison of Self-Report and Clinician Ratings on Two Inventories of Depressive Symptomatology. Psychiatric Services,57(6), 829-837. doi:10.1176/appi.ps.57.6.829

Santor, D. A., Gregus, M., & Welch, A. (2006). FOCUS ARTICLE: Eight decades of measurement in depression. Measurement: Interdisciplinary Research and Perspectives, 4(3), 135-155. doi:10.1207/s15366359mea0403_1

Schmaal, L., Veltman, D., Van Erp, T. G. M., Smann, P. G., Frodl, T., Jahanshad, N., . . . for the ENIGMA-Major Depressive Disorder Working Group. (2016). Subcortical brain alterations in

major depressive disorder: Findings from the ENIGMA major depressive disorder working group. Molecular Psychiatry, 21(6), 806-812. doi:10.1038/mp.2015.69

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, hamilton, and zung. Journal of Clinical Psychology, 62(1), 123-146. doi:10.1002/jclp.20213

Shanteau, J. (1992). Competence in experts: The role of task characteristics. Organizational Behavior and Human Decision Processes, 53(2), 252-266. doi:10.1016/0749-5978(92)90064-E

Shear, M. K., Greeno, C., Kang, J., Ludewig, D., Frank, E., Swartz, H. A., & Hanekamp, M. (2000). Diagnosis of nonpsychotic patients in community clinics. American Journal of Psychiatry, 157(4), 581-587. doi:10.1176/appi.ajp.157.4.581

Shorter, E. (2013). How everyone became depressed: The rise and fall of the nervous breakdown. New York, NY: Oxford University Press.

Shorter, E. (2015). The history of nosology and the rise of the diagnostic and statistical manual of mental disorders. Dialogues in Clinical Neuroscience, 17(1), 59.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. Psychometrika, 74(1), 107-120. doi:10.1007/s11336-008-9101-0

Slocum, S. L. (2005). Assessing unidimensionality of psychological scales: Using individual and integrative criteria from factor analysis

Snaith, R. P., Harrop, F. M., Newby, D. A., & Teale, C. (1986). Grade scores of the montgomery-asberg depression and the clinical anxiety scales. The British Journal of Psychiatry : The Journal of Mental Science, 148(5), 599-601. doi:10.1192/bjp.148.5.599

Smits, N., Smit, F., Cuijpers, P., & De Graaf, R. (2007). Using decision theory to derive optimal cut-off scores of screening instruments: An illustration explicating costs and benefits of mental health screening. International Journal of Methods in Psychiatric Research, 16(4), 219-229. doi:10.1002/mpr.230

Smits, N. (2010). A note on youden's J and its cost ratio. BMC Medical Research Methodology, 10(1), 89-89. doi:10.1186/1471-2288-10-89

Solomon, A., Haaga, D. A. F., & Arnow, B. A. (2001). Is clinical depression distinct from subthreshold depressive symptoms? A review of the continuity issue in depression research.The Journal of Nervous and Mental Disease, 189(8), 498-506. doi:10.1097/00005053-200108000-00002

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1997). Further evidence for the construct validity of the beck depression inventory-II with psychiatric outpatients. Psychological Reports, 80(2), 443-446. doi:10.2466/pr0.1997.80.2.443

Steinhauser, S., Schumacher, M., & Rücker, G. (2016). Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC Medical Research Methodology, 16(1), 97-15. doi:10.1186/s12874-016-0196-1

Tabachnick, B. G., & Fidell, L. S. (2014). Using multivariate statistics (6th ed.). Harlow: Pearson Education Limited.

Temme, D., & Diamantopoulos, A. (2016). Higher-order models with reflective indicators: A rejoinder to a recent call for their abandonment. Journal of Modelling in Management, 11(1), 180-188. doi:10.1108/JM2-05-2014-0037

Trivedy, M. H., Rush, A, J., Ibrahim, H. M., Carmody, T. J., Biggs, M. M., Suppes, T.,… Kashner, T. M. (2004). The inventory of depressive symptomatology, clinician rating (IDS-C) and self-report (IDS-SR), and the quick inventory of depressive symptomatology, clinician rating (QIDS-C) and self-report (QIDS-SR) in public sector patients with mood disorders: A psychometric evaluation. Psychological Medicine, 34(1), 73-82. doi:10.1017/S0033291703001107

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. biases in judgments reveal some heuristics of thinking under uncertainty. Science, 185(4157), 1124-1131. doi:10.1126/science.185.4157.1124

Tweed, D. L. (1993). Depression-related impairment: Estimating concurrent and lingering effects. Psychological Medicine, 23(2), 373-386. doi:10.1017/S0033291700028476

Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., . . . Aitchison, K. . (2008). Measuring depression: Comparison and integration of three scales in the GENDEP study.Psychological Medicine, 38(2), 289-300. doi:10.1017/S0033291707001730

Ulrich, A. (2016). Depression: Giver diagnosens tilgang mening? Tidsskrift for Forskning i Sygdom Og Samfund,

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. Applied Psychological Measurement, 35(5), 380-392. doi:10.1177/0146621610392911

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. European Journal of Developmental Psychology, 9(4), 486-492. doi:10.1080/17405629.2012.686740

Vilagut, G., Forero, C. G., Barbaglia, G., & Alonso, J. (2016). Screening for depression in the general population with the center for epidemiologic studies depression (ces-d): A systematic review with meta-analysis. Plos One, 11(5), e0155431. doi:10.1371/journal.pone.0155431

Vouloumanou, E. K., Plessa, E., Karageorgopoulos, D. E., Mantadakis, E., & Falagas, M. E. (2011). Serum procalcitonin as a diagnostic marker for neonatal sepsis: A systematic review and meta-analysis. Intensive Care Medicine, 37(5), 747-762. doi:10.1007/s00134-011-2174-8

Wittchen, H. (1994). Reliability and validity studies of the WHO-composite international diagnostic interview (CIDI): A critical review. Journal of Psychiatric Research, 28(1), 57-84. doi:10.1016/0022-3956(94)90036-1

Wonch, K. E., de Medeiros, C. B., Barrett, J. A., Dudin, A., Cunningham, W. A., Hall, G. B., . . .

Fleming, A. S. (2016). Postpartum depression and brain response to infants: Differential amygdala response and connectivity. Social Neuroscience, 11(6), 600-617. doi:10.1080/17470919.2015.1131193

Wooldridge, J. M. (2016). Introductory econometrics: A modern approach (6.th ed.). Boston, MA: Cengage Learning.

Worboys, M. (2013). The hamilton rating scale for depression: The making of a "gold standard" and the unmaking of a chronic illness, 1960-1980. Chronic Illness, 9(3), 202.

World Health Organisation (2018). Depression. Retrieved May 10, 2019, from https://www.who.int/news-room/fact-sheets/detail/depression

World Health Organization. (1992). International statistical classification of diseases and related health problems (10. revision ed.). Geneva: WHO.

Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3(1), 32-35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. European Journal of Psychological Assessment, 31(4), 231-237. doi:10.1027/1015-5759/a000309

Zimmerman, M., Walsh, E., Friedman, M., Boerescu, D. A., & Attiullah, N. (2018). Are self-report scales as effective as clinician rating scales in measuring treatment response in routine clinical practice? Journal of Affective Disorders, 225, 449-452. doi:10.1016/j.jad.2017.08.024