Udvikling og validering af patientrapporterede spørgeskemaer – del 2

Cecilie Balslev Willert[1], Lisbet Rosenkrantz Hölmich[1] & Kristian Thorborg[2]

STATUSARTIKEL

1) Plastikkirurgisk Afdeling V, Herlev Hospital
2) Ortopædkirurgisk Afdeling og Fysioterapien, Amager-Hvidovre Hospital

Ugeskr Læger 2015;177:V04150290

Oplysninger fra patienter indgår i stigende grad i klinisk og forskningsmæssig praksis. Begrebet *patient-reported outcome measurement* (PROM) anvendes efterhånden også på dansk. Denne artikel er anden og sidste del af en gennemgang af dette emne. Der henvises til første del for gennemgang af god metode til udvikling og oversættelse af sådanne spørgeskemaer samt definition af begreber [1].

I denne del gennemgås feltundersøgelse og validering af patientrapporterede spørgeskemaer, hvilket bør foretages på baggrund af test af spørgeskemaet på en større gruppe patienter og har til formål at afdække spørgeskemaets målemæssige egenskaber.

På baggrund af et internationalt konsensusarbejde har en større ekspertgruppe udarbejdet *the consensus-based standards for the selection of health measurement instruments* (COSMIN) *checklist,* blandt andet for at facilitere bedømmelse og sammenligning af den metodologiske kvalitet af spørgeskemaundersøgelser. COSMIN-gruppen har udarbejdet artikler [2, 3] med beskrivelse af de psykometriske måleegenskabers mest essentielle kvaliteter og indbyrdes forhold (**Figur 1**) med en dertilhørende tjekliste [4]. I denne artikel er der taget udgangspunkt i dette omfattende arbejde.

FELTUNDERSØGELSE: TEST PÅ EN STØRRE GRUPPE PATIENTER

Når et spørgeskema er udviklet og pilottestet jf. metoder omtalt i del I [1], skal det besvares af en større gruppe patienter i en feltundersøgelse [5]. Formålet er at fravælge dårligt fungerende items og at undersøge, hvordan besvarelserne fordeler sig på item-svar-skalaerne [5]. Til det kan der bl.a. benyttes matematiske beregninger, der tager udgangspunkt i psykometriske metoder, hvis teori gennemgås nedenfor.

Classical test theory (CTT) er i dag den mest anvendte ved udvikling og validering af spørgeskemaer [6, 7]. CTT baseres på det udsagn, at en observeret score består af en sand score og fejl, hvor den sande score er den, man ville få, hvis den samme person udfyldte det samme spørgeskema uendeligt mange gange [5, 8, 9]. CTT forudsætter, at spørgeskemaet er unidimensionelt, dvs. at der kun måles på ét begreb, og at items er korrelerede, fordi de er manifestationer af det samme begreb [5]. Ved CTT-validering udføres der statistiske analyser på den samlede score, f.eks. beregning af reliabilitet. Valideringen er situations- og populationsafhængig og bør derfor gentages, hvis skemaet skal bruges i en ny sammenhæng [6-8].

Ved CTT kan faktoranalyse (FA) anvendes til itemreduktion. Med FA undersøger man, hvilke faktorer (eller dimensioner) som indgår i spørgeskemaet baseret på itemkorrelation, idet man antager, at items med høj korrelation hører til samme faktor. Items, der ikke har korrelation til nogen faktor, kan slettes, ligesom faktorer, der har lav betydning for begrebet, kan udgå [5]. Med FA undersøger man også, om en skala er unidimensionel, altså om der er én dominerende faktor, og den kan kun bruges ved en refleksiv model (se del I) [1, 5, 7].

Item response theory (IRT) er en metode, som i stigende grad benyttes i sundhedsvidenskabelig forskning. Den forudsætter foruden unidimensionalitet, at svaret på ét item er uafhængigt af svaret på et andet item [7, 8].

For at forstå IRT, kan man f.eks. ved måling af depression forestille sig et kontinuum, der afbilder graden af depression fra mild til svær. Den enkelte patient kan placeres et sted på kontinuummet, og det samme kan items. Dvs. at et item, der besvares på en bestemt måde ved svær depression, ligger i slutningen af dette kontinuum. Med IRT tager man på den måde højde for, at items har forskellige »sværhedsgrader« [5].

For at placere items det rigtige sted på dette kontinuum kræves en itemanalyse, der foretages på baggrund af spørgeskemabesvarelser fra en større gruppe

FAKTABOKS

- Feltundersøgelse og validering foretages på baggrund af afprøvning af spørgeskemaet på en større patientgruppe.
- En validering afdækker et patientrapporteret spørgeskemas målemæssige egenskaber som et udtryk for dets kvalitet.
- *Consensus-based standards for the selection of health measurement instruments checklist* er et internationalt konsensusarbejde, der faciliterer bedømmelse og sammenligning af den metodologiske kvalitet af valideringsstudier.
- *Classical test theory* og *item response theory* er begge psykometriske metoder med forskellige egenskaber, der kan bruges ved både udvikling og validering af et patientrapporteret spørgeskema.
- Vigtige psykometriske måleegenskaber er bl.a. overordnet reliabilitet, validitet og responsivitet, der hver især har flere underkategorier.

patienter. Itemanalysen kan også bruges til itemreduktion. Hvis to items ligger næsten samme sted på kontinuummet, er det ene item overflødigt og kan udgå. Omvendt kan man også se, om nogle steder på kontinuummet er tomme, og man bør tilføje items [5].

IRT har den fordel frem for CTT, at der bruges intervalskalaer i stedet for ordinale skalaer (se del I) [1], hvilket giver den samlede score større statistisk sikkerhed [6, 7, 10].

Der er internationalt ikke absolut enighed om, hvilken metode der er den mest korrekte i forbindelse med udvikling, feltundersøgelse og validering af et spørgeskema, og det er uden for denne artikels formål at gå ind i denne diskussion. COSMIN-gruppen har opstillet kriterier for god metode ved begge tilgange [2].

ET SPØRGESKEMAS MÅLEEGENSKABER

Et spørgeskema har forskellige kvaliteter, såkaldte måleegenskaber. Nogle måleegenskaber kan være vigtigere end andre, afhængigt af spørgeskematypen. F.eks. er måleegenskaben responsivitet særlig vigtig ved et evaluerende spørgeskema. Måleegenskaberne bør vurderes, inden spørgeskemaet tages i brug, så det er fastsat, hvad spørgeskemaet »kan«, inden man skal fortolke besvarelserne. Nedenfor gennemgås de forskellige måleegenskaber.

OVERORDNET RELIABILITET

Reliabilitet som paraplybegreb defineres som graden, i hvilken resultatet af et spørgeskema er fri for målefejl [3]. En mere dybdegående definition er, i hvilket omfang den samlede score for patienter, der ikke har haft ændring i deres tilstand, er den samme ved gentagne målinger [3]. Tre parametre afspejler den overordnede reliabilitet: intern konsistens, reliabilitet og målefejl.

Intern konsistens

Intern konsistens er et mål for, i hvor høj grad items korrelerer med hinanden [2]. Måling af intern konsistens forudsætter, at spørgeskemaet er opbygget efter en refleksiv model, og at skalaen er unidimensionel, hvorfor FA bør udføres forinden [2, 5].

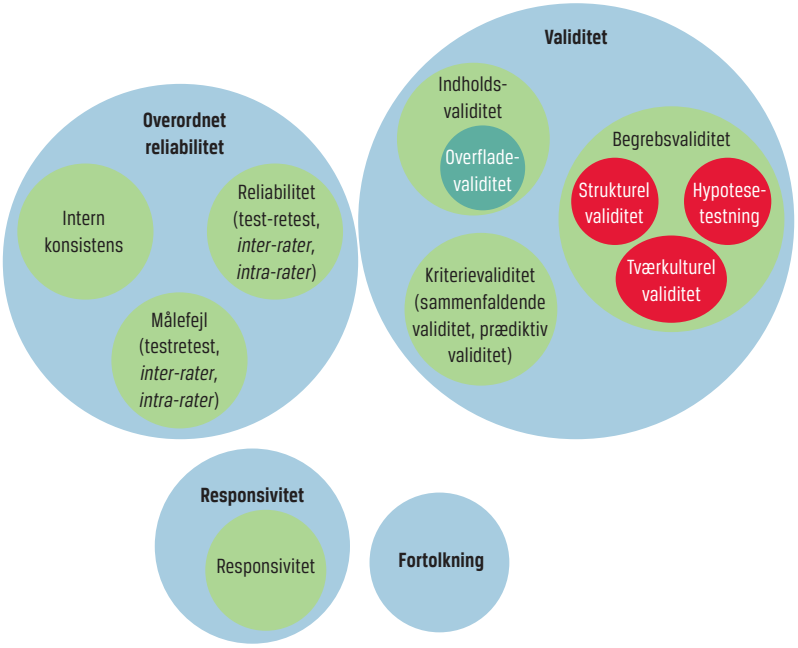
Den alment benyttede metode til måling af intern konsistens ved CTT er beregning af Cronbachs α, der har en værdi på 0-1, og bør være 0,70-0,90 [5].

Reliabilitet

Reliabilitet defineres som den andel af variationen i patienternes resultater i en spørgeskemaundersøgelse, der skyldes den »sande« forskel [3]. Reliabilitet er således et udtryk for, hvor godt man vha. et spørgeskema skelner mellem patienterne på trods af målevariation [11]. Det undersøges ved en test-retest, hvor de samme patienter udfylder spørgeskemaet to gange med et passende tidsinterval (2-14 dage uden ændringer i patienternes tilstand) [7, 12, 13]. Reliabiliteten beregnes med *intra-class*-koefficienten (ICC), der har en værdi på 0-1 og bør være > 0,70 [5, 12].

FIGUR 1

Consensus-based standards for the selection of health measurement instruments **taksonomiske oversigt over diverse psykometriske måleegenskaber og deres indbyrdes forhold [4].**



Målefejl

Enhver måling vil være behæftet med fejl i en eller anden grad, og målefejl defineres som den systematiske og tilfældige variation, der er indeholdt i en patients samlede score og ikke skyldes sande ændringer i begrebet [3]. Målefejl undersøges ligesom reliabilitet ved test-retest og er et udtryk for målevariationen i den enkelte patients score udtrykt i absolutte tal i den enhed, som man anvender i målingen [5, 11]. Målefejl udtrykkes f.eks. med *standard error of measurement* og *minimal detectable change* og kan bl.a. beregnes ud fra ICC [5, 11].

VALIDITET

Validitet er et udtryk for, i hvor høj grad man med et spørgeskema måler det begreb, som det har til formål at måle, altså gyldigheden [3]. Der er flere undertyper af validitet.

Indholdsvaliditet inklusive overfladevaliditet

Ved indholdsvaliditet vurderer man, om spørgeskemaet i tilstrækkelig grad berører alle relevante aspekter af begrebet. Man bør stille følgende spørgsmål under udviklingen: »Berører alle items relevante aspekter af begrebet?« og »er alle items relevante for målgruppen

mht. f.eks. alder, køn, sprog og sygdomskarakteristika?« [2, 5].

Indholdsvaliditet er en kvalitativ egenskab, der ikke kan beregnes. I stedet må man forsøge at opnå såkaldt datamætning ved f.eks. under udviklingen at blive ved med at lave fokusgruppe- og/eller enkeltpersoninterview, indtil der ikke dukker nye emner op [14].

Overfladevaliditet (*face validity*) defineres som graden, i hvilken (items i) et spørgeskema ser ud til tilstrækkeligt at reflektere det begreb, der måles på [3]. Det er altså et slags førstehåndsindtryk af spørgeskemaet, og det kan f.eks. være vigtigt, hvis man skal vurdere, om et spørgeskema egner sig til ens eget studie.

Kriterievaliditet

Kriterievaliditet defineres som graden, i hvilken resultaterne fra et spørgeskema er en tilstrækkelig refleksion af en »guldstandard« [3]. Kriterievaliditet er primært aktuelt i de tilfælde, hvor man har et spørgeskema (»guldstandard«), som man ønsker at forkorte [2]. Man beregner, hvor stor overensstemmelse der er mellem resultaterne fra standardskemaet og resultaterne fra det nye spørgeskema ved f.eks. at beregne en korrelation [5].

Begrebsvaliditet

Begrebsvaliditet bruges, når man ikke har en »guldstandard« i form af et andet spørgeskema at sammenligne med. Man belyser i stedet validiteten ved at opstille specifikke teoretiske modeller (hypoteser) med relation til begrebet for herefter at teste dem [2, 5, 7, 12, 13]. COSMIN-gruppen definerer begrebsvaliditet som graden, i hvilken resultaterne fra et patientrapporteret spørgeskema er i overensstemmelse med disse hypoteser [3]. Ved hypotesetestning formulerer man a priori-hypoteser om, hvordan spørgeskemaets resultater vil korrelere med resultater fra andre spørgeskemaer, hvor man måler enten et lignende begreb (konvergerende validitet) eller et helt andet begreb (divergerende validitet), for herefter at afprøve dem [5]. Hypoteserne bør være teoretisk begrundende og specifikke med hensyn til både i hvilken retning, man forventer korrelationen (positiv eller negativ), og størrelsen af den [2, 5].

Øvrige undertyper af begrebsvaliditet

To andre former for validitet nævnes kort her. Tværkulturel validitet er aktuel, når et spørgeskema oversættes til et nyt sprog eller skal bruges i en ny kultur. Det defineres som graden, i hvilken items i den nye version tilstrækkeligt reflekterer besvarelserne af items i originalversionen (se del I) [1, 3].

Strukturel validitet omhandler strukturen i spørgeskemaet. Det defineres som graden, i hvilken resultaterne fra et spørgeskema reflekterer begrebets relevante dimensioner tilstrækkeligt og undersøges f.eks. ved at bruge FA [3, 5].

RESPONSIVITET

Responsivitet defineres som spørgeskemaets evne til at detektere ændring over tid i det begreb, der måles på [3]. Responsivitet er kun relevant ved evaluerende spørgeskemaer, der har til formål at måle, hvor meget patienterne ændrer sig f.eks. efter behandling [5]. Responsivitet undersøges enten som kriterieresponsivitet eller hypoteseresponsivitet, der er analoge til kriterievaliditet og begrebsvaliditet mht. både teori og metode; den eneste forskel er, at man her sammenligner ændringerne i score [5].

FORTOLKNING

Fortolkning dækker over graden, i hvilken man kan tillægge de kvantitative samlede scorer eller ændringer i samlet score fra et spørgeskema en klinisk betydning [3]. Den »mindste betydningsfulde ændring« er den mindste ændring i en patients samlede score, som patienten selv opfatter som betydningsfuld [2]. Den kan eksempelvis bestemmes ved brug af et eksternt anker [12, 15], f.eks. et ekstra spørgsmål, hvor patienterne angiver ændringer i tilstanden (f.eks. tristhed ved depression). Herefter beregnes forskellen fra *baseline*-scoren hos de patienter, der angiver en lille ændring. Der findes i dag flere forskellige metoder til bestemmelse af den mindste betydningsfulde ændring, som alle giver forskellige resultater, og der er endnu ingen enighed om, hvad der er den foretrukne [16].

KONKLUSION

I to artikler har vi beskrevet metoder til udvikling og validering af patientrapporterede spørgeskemaer. Eftersom spørgeskemaer bruges i stigende grad i både klinisk øjemed og forskningsøjemed, er det vigtigt med en basal forståelse for, hvordan man vurderer kvaliteten af dem. Vi har forsøgt at forklare avancerede psykometriske begreber på en måde, der forhåbentligt er forståelig for de fleste læsere. Vi har med disse artikler ønsket at skabe fokus på god metode ved både udvikling, oversættelse, feltundersøgelse og validering af patientrapporterede spørgeskemaer, når hovedformålet inden for sundhedsområdet er at få gyldige vurderinger af patientrelevante, selvoplevede problemer.

SUMMARY

Cecilie Balslev Willert, Lisbet Rosenkrantz Hölmich & Kristian Thorborg:

Developing and validating of patient-reported questionnaires – part 2

Ugeskr Læger 2015;177:V04150290

Patient-reported outcome measurements (PROMs) are often questionnaires which provide and rate the patient's point of view in the measurement of subjective clinical phenomena such as pain or quality of life. In the second of two articles we describe field-testing, the psychometric theories Classical Test Theory and Item Response Theory, psychometric validation and the measurement properties of a PROM. The latter is based on the COSMIN (consensus-based standards for the selection of health measurement instruments) guidelines, which are developed to assist in development and evaluation of PROMs.

KORRESPONDANCE: *Cecilie Balslev Willert*. E-mail: ceciliebwi@gmail.com

ANTAGET: 13. juli 2015

PUBLICERET PÅ UGESKRIFTET.DK: 12. oktober 2015

INTERESSEKONFLIKTER: Forfatternes ICMJE-formularer er tilgængelige sammen med artiklen på Ugeskriftet.dk

LITTERATUR

1. Willert CB, Rosenkrantz HL, Thorborg K. Udvikling af validering af patientrapporterede spørgeskemaer – del 1. Ugeskr Læger 2015; 177:V08140450.
2. Mokkink LB, Terwee CB, Knol DL et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010; 10:22.
3. Mokkink LB, Terwee CB, Patrick DL et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.
4. Mokkink LB, Terwee CB, Patrick DL et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-49.
5. Vet HCW. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press, 2011:338.
6. Cano SJ, Hobart JC. The problem with health measurement. Patient Prefer Adherence 2011;5:279-90.
7. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford: Oxford University Press, 2008:431.
8. de Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 2010;44:109-17.
9. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15:72-101.
10. Bowling A. Measuring health: a review of quality of life measurement scales. 3. ed. London: Open University Press, 2005:211.
11. de Vet HC, Terwee CB, Knol DL et al. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.
12. Terwee CB, Bot SD, de Boer MR et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
13. Grønvold M. Metoder i livskvalitetsforskning. Ugeskr Læger 2008; 170:825-9.
14. Patrick DL, Burke LB, Gwaltney CJ et al. Content validity-establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1-eliciting concepts for a new PRO instrument. Value Health 2011;14:967-77.
15. Revicki DA, Cella D, Hays RD et al. Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes 2006;4:70.
16. Terwee CB, Roorda LD, Dekker J et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol 2010;63:524-34.