



2013

Patientmedverkan sjukdomsspecifika PROM

Att skapa ett sjukdomsspecifikt
frågeformulär - en kort manual

Vad skall frågorna till patienten mäta?

Vilka respondenter är det ansett för?

Vilka krav på validitet ska ställas?

Går det att reducera antalet frågor?

Professor Mats Lundström

i samarbete med

Registercentrum Syd

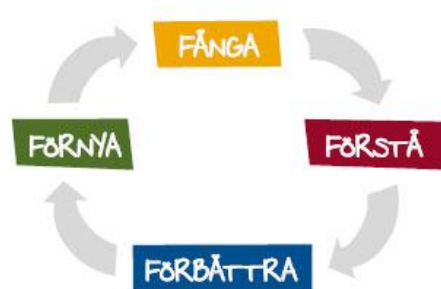
EyeNet Sweden, Karlskrona

2013-09-04



Innehållsförteckning - Att skapa ett sjukdomspecifikt frågeformulär

1. Vilket syfte skall frågeformuläret uppfylla?	4
2. Vad skall frågeformuläret innehålla?	4
2.1 Vad skall frågorna mäta?	4
2.2 Vilka respondenter är det avsett för?	5
2.3 Konstruktion av frågor	5
2.4 Reduktion av antalet frågor	6
3. Vilken design skall svarsalternativen (skalorna) ha?	7
4. Hur skall ett ifyllt frågeformulär poängbedömas?	7
5. Vilka krav på validitet skall ställas?	8
6. Vilka krav på reliabilitet skall ställas?	9
7. Vilka krav på känslighet (för förändringar) skall ställas?	10
8. Hur skall utvärderingen göras begriplig?	11
9. Hur skall man länka samman kliniska utfallsdata och patientrapporterade utfallsdata?	11
10. Referenslista	12
11. Appendix I	14
12. Appendix II	18



1. Vilket syfte skall frågeformuläret uppfylla?

Syftet med ett sjukdomsspecifikt frågeformulär är i regel att man vill undersöka hur ett visst tillstånd eller en viss behandling påverkar den hälsorelaterade livskvaliteten.

Det kan gälla att

- Utvärdera indikationer för åtgärd
- Beskriva utfall i rutinverksamhet och för olika undergrupper
- Utvärdera olika terapier/implantat/åtgärder ur ett patientperspektiv
- Belysa hälsoekonomiska aspekter

Det kan finnas andra syften som är mer eller mindre specifika för den patientgrupp man vill undersöka. Det kan inte nog understrykas hur viktigt det är att göra klart för sig syftet med det frågeformulär man vill skapa. Är inte detta klarlagt riskerar man att mäta fel saker och dra fel slutsatser.

2. Vad skall frågeformuläret innehålla?

2.1 Vad skall frågorna mäta?

Det är viktigt att klart definiera vad frågeformuläret skall mäta och vilken population det är avsett för. När man konstruerar ett sjukdomsspecifikt frågeformulär handlar det i regel om att mäta på vilket sätt den aktuella sjukdomen eller tillståndet påverkar individen. Denna påverkan kan gälla förmågan att klara dagliga aktiviteter, förmågan till självständigt leverne, förmågan till ett önskvärt socialt liv etc. Några exempel på områden ges nedan.

- Ett tillstånd eller en behandling kan påverka förmågan att leva det liv man föredrar och denna påverkan faller inom det som WHO klassificerar som ”Activity limitation”. Med detta menas hur man klarar av att utföra önskade aktiviteter med olika typer av funktionsnedsättning såsom rörelseinskränkning (på grund av ledsjukdom, kärlsjukdom etc.), synnedsättning, hörselnedsättning och liknande.
- Ett tillstånd eller behandling kan påverka det sociala livet, det vill säga hur säker man är och hur man klarar av umgänge med andra. Detta kan vara aktuellt vid tandimplantat, kosmetiska operationer, synfelsoperationer, kontaktlinser, stomier.

Det gäller alltså att ha en klar bild av vad man vill mäta. Som framgår ovan är det ofta ett karakteristiskt (kännetecknande) drag hos en population man önskar belysa. Ett batteri av frågor kan ge en allsidigare belysning av detta än vad enkla ja- och nej-frågor kan.

Den engelska benämningen på vad man vill mäta hos patienterna är ”latent trait” och ibland förekommer uttrycket ”underliggande funktion” på svenska. Det är en abstrakt företeelse som vi saknar ett riktigt bra samlingsnamn för. I denna skrift

används uttrycket funktion för det vi vill mäta med ett sjukdomsspecifikt frågeformulär. Det avser t.ex. patientens självskattade synfunktion, rörelsefunktion, funktion pga. smärta etc.

2.2 Vilka respondenter är det avsett för?

Ett sjukdomsspecifikt frågeformulär skall ha en väl definierad population för vilken det skall användas. Det är viktigt att testpopulationen då formuläret tillverkas överensstämmer med den slutliga användarpopulationen. Med detta som utgångspunkt är det ändå oundvikligt att formuläret används på blandade populationer i den kliniska verkligheten. Det blir nästan alltid en blandning av patienter med och utan komorbiditet, behandlade och obehandlade etc. Ofta testas frågeformulär på närliggande tillstånd. Utgångsläget måste dock vara att utprovning och test av det sjukdomsspecifika frågeformuläret är gjort på en väl definierad population. Anledningen är förstås att det måste finnas en väl definierad referensgrupp för ett mätinstrument om det skall kunna användas seriöst inom vården. Nedan följer några frågor som man bör tänka igenom vid konstruktionen av ett sjukdomsspecifikt frågeformulär

– Vid vilka tidpunkter i en process skall frågeformuläret användas?

- Före och efter en åtgärd
- Med vissa tidsintervall i en kronisk process
- Efter vissa åtgärder i en kronisk process

– Hur ofta skall det användas?

- På alla patienter
- På ett urval (stickprov etc.) av patienter
- Hela året eller vissa begränsade tider
- Som begränsade projekt

2.3 Konstruktion av frågor

Vanligtvis försöker man att samla så många frågor (items) som möjligt i första skedet. Det gäller att samla en stor frågebank som på olika sätt belyser den underliggande funktion (karaktäristiska drag) hos en population som man vill mäta. I ett senare skede väljer man ut en begränsad mängd av de bästa frågorna för det slutliga frågeformuläret.

Det finns huvudsakligen tre olika principiella sätt att formulera frågor. Optimalt är att utnyttja alla tre metoderna:

1. Litteraturgenomgång

- Finns något redan etablerat frågeformulär?

Det första som bör göras är att undersöka om det redan finns ett etablerat och validerat frågeformulär för den funktion man vill undersöka. I så fall är det lämpligt att använda detta. Är det på svenska behövs inget annat än en kontroll av om det får användas fritt. Är det på ett annat språk krävs en översättning. Det finns en speciell procedur för översättning (fram- och tillbakaöversättning) som man bör följa¹⁻². Även i detta fall måste man kontrollera copyright och tillstånd.

- Vilka symptom/problem etc. finns beskrivna?

Finns inget frågeformulär etablerat kan man genom litteraturstudier göra en lista på de symptom eller problem som patienter beskrivit. Detta kan ligga till grund för frågor för att belysa den funktion man vill mäta.

2. Expertpanel

- Experter som hanterar sjukdomsområdet/tillståndet

Professionen har ofta en uppfattning om vilka frågor som kan vara relevanta för att belysa den funktion man vill mäta. Det är viktigt att inhämta synpunkter från experter som hanterar olika steg i en vårdprocess (olika personalkategorier t.ex.).

3. Patienter

- Intervjuer eller fokusgrupper

Självklart är patienten den bästa källan för att få fram frågor som belyser den funktion man vill mäta. Man brukar använda sig av intervjuer av enskilda patienter. Man kan använda strukturerade intervjuer varvid man ställer redan formulerade frågor för att se hur svaren blir eller semistrukturerade intervjuer där man t.ex. kan fråga patienten om vilka problem som vederbörande upplever. En annan metod är att använda s.k. fokusgrupper³. Vid dessa får en grupp patienter diskutera fritt vilka problem de upplever. En person leder diskussionen och en annan spelar in eller skriver ned vad som sägs. Ibland har man flera fokusgrupper som kan innehålla erfarenhet från olika nivåer av en sjukdomsprocess eller olika typer av åtgärd.

2.4 Reduktion av antalet frågor (items)

Man samlar så många frågor som möjligt för att allsidigt belysa den aktuella funktionen. Antalet frågor blir ofta så stort att man måste reducera antalet i nästa steg för att kunna skapa ett hanterligt frågeformulär³.

Reduktionen sker genom att en expertpanel går igenom frågorna och sorterar ut likvärdiga frågor (dubbletter) och de frågor man uppfattar som irrelevanta. Det reducerade frågeformuläret används på patienter i en pilotstudie. Patienterna måste vara representativa för den population frågeformuläret är avsett för. Pilotstudien bör optimalt omfatta ett par hundra patienter för ett ge tillräckligt stöd för den fortsatta processen. Generellt kan man säga att ju fler frågor formuläret innehåller, ju fler testpatienter bör ingå i pilotstudien.

En analys av pilotstudien görs och det slutliga antalet frågor avgörs. Den bästa metoden för att göra denna reduktion efter pilotstudien är genom en s.k. Rasch-analys (se appendix)⁴⁻⁶. Vid denna testas

- Patienternas förmåga att särskilja svarsalternativen
- Frågeformulärets förmåga att diskriminera olika patientgrupper
- Att alla frågor bidrar till resultatet utan att vara överflödiga eller utanför ämnet
- Att frågeformuläret mäter en och samma underliggande funktion
- Att frågornas svårighetsgrad överensstämmer med patienternas förmåga

Det kan också finnas skäl att utesluta en fråga om många respondenter väljer att inte svara på den, dvs. det uppstår mycket ”missing data”.

3. Vilken design skall svarsalternativen (skalorna) ha?

Att designa skalorna är en mycket viktig del av konstruktionen av ett frågeformulär. Det är en fördel om man kan använda så få skalalternativ som möjligt så att frågeformuläret ger ett enhetligt intryck, helst samma skalor genom hela frågeformuläret. Om man vill fråga om olika typer av problem eller besvär så bör svarsalternativen i alla frågor således innehålla samma skala som t.ex. inga, vissa, stora och mycket stora. Om frågeformuläret innehåller påståenden som skall bekräftas eller avvisas gäller samma princip, t.ex. stämmer inte alls, stämmer delvis, stämmer för det mesta, stämmer helt och hållet. Man bör sträva efter att hålla sig till antingen frågor eller påståenden. Man bör också sträva efter att ha samma utväxling av svaren, som t.ex. att börja med inga/stämmer inte och sluta med mycket stora/stämmer helt och hållet. Det är ingen vits med att testa vakenheten hos respondenten genom att ändra svarsföljden, det ökar risken för misstag.

Man önskar självklart att frågeformuläret skall ge så mycket information som möjligt. Det är dock ett dilemma, skapar man för många svarsalternativ kommer respondenten inte att kunna särskilja dem och svarar slumpartat. Det gäller att skapa klart urskiljbara svarsalternativ som respondenten inte tvekar om. Sju till tio svarsalternativ (inklusive termometer) är nästan alltid för många. Fyra till fem svarsalternativ brukar vara lagom då det ger ordnade svar. Då får man tänka på att vid fyra svarsalternativ finns ingen mittpunkt, vid fem finns en mittpunkt. Finns det en mittpunkt som i exempel 2 nedan riskerar man att många väljer mittalternativet och därmed inte tillför tillräcklig information. Vill man tvinga patienten att ta ställning bör man välja fyra. Exempel på detta kan vara fråga 1 nedan om nöjdhet:

Är Du nöjd med Din nuvarande xxx (funktion)?

1. *Fyra svarsalternativ: Ja, mycket nöjd; Ja, ganska nöjd; Nej, ganska missnöjd; Nej, mycket missnöjd.*
2. *Fem svarsalternativ: Ja, mycket nöjd; Ja, ganska nöjd; Varken nöjd eller missnöjd; Nej, ganska missnöjd; Nej, mycket missnöjd.*

4. Hur skall ett ifyllt frågeformulär poängbedömas?

I de flesta frågeformulär ger svarsalternativen poäng enligt en viss ordning. Vanligen ger man 0 eller 1 poäng till svarsalternativet som inte innebär några problem/besvär etc. Därefter stigande poäng då svårighetsgraden ökar.

Likert scoring

Poängen som varje fråga genererar adderas sedan till en totalpoäng för hela frågeformuläret. Denna teknik kallas ibland ”Likert scoring” och innebär att man summerar ordinala data som man satt på svarskategorierna^{7,8}. Vissa frågeformulär behandlar dessutom totalpoängen som frågeformuläret kan generera med en aritmetisk formel så att totalpoängen kan bli mellan 0 och 100. Denna metod att summera ordinala data är ålderdomlig och kan inte rekommenderas. Anledningen är att olika frågor har olika valör och bör inte ges exakt samma poäng.

Ett exempel på hur frågor kan betyda olika mycket är en enkät som används vid ätstörningar med en fråga om motion och en annan om kräkningar efter måltid. Ett annat exempel är i ett frågeformulär om synproblem frågor om bilkörning på natten och om att känna igen ansikten.

1. Frågeformulär vid ätstörningar

A. *Brukar Du motionera för att hålla vikten nere?*

Nej, aldrig (1); ja, ibland (2); ja, några gånger i veckan (3); ja, dagligen (4)

B. *Brukar Du kräkas upp maten efter måltid för att hålla vikten nere?*

Nej, aldrig (1); ja, ibland (2); ja, några gånger i veckan (3); ja, dagligen (4)

2. Frågeformulär vid nedsatt syn

A. *Upplever Du några problem på grund av synen vid bilkörning på natten?*

Nej, inga problem (1); Ja, vissa problem (2); Ja, stora problem (3); Ja, mycket stora problem (4)

B. *Upplever Du några problem med att känna igen ansikten på dem Du möter?*

Nej, inga problem (1); Ja, vissa problem (2); Ja, stora problem (3); Ja, mycket stora problem (4)

Uppenbarligen kräver dessa parvisa exempel olika mycket av respondenten. Dessutom är avståndet mellan svarsalternativen olika beroende på frågan och detta tar den gamla metoden inte hänsyn till. Ett exempel på detta är fråga 1: B i exemplet ovan. Avståndet mellan "aldrig (1)" och "ibland (2)" är betydligt större än mellan "ibland (2)" och "några gånger i veckan (3)". Men vid en Likert scoring är skillnaden i båda fallen 1 poäng. Dessutom måste ordinala data behandlas som sådana och inte som intervalldata. Vi önskar ju att frågeformuläret skall vara ett fullgott mätinstrument som andra instrument inom vården och då duger inte Likert scoring.

Rasch-analys

Rasch-analys^{5,6} är en typ av s.k. Item-Respons-Teori (IRT)⁹. Det innebär att man sorterar alla respondenter och alla frågor utefter en och samma skala (linjal). På ena sidan (vänster) av en lodrät skala sorteras respondenter från de svagast presterande till de starkast presterande och på andra sidan (höger) sorteras frågorna från de lättaste till de svåraste. Man utgår från de poängsiffror som är satta på frågornas svarsalternativ men analysen räknar sedan om dessa värden till en s.k. Rasch score. Enheten på skalan är logits och uttrycker logaritmen av odds ratio för att en respondent skall klara en fråga eller inte. Om en given respondent har en förmåga som placerar vederbörande på en plats längs skalan och frågan som skall besvaras har en svårighetsgrad som placerar den någon annanstans på skalan, då uttrycker avståndet mellan respondent och fråga på skalan sannolikheten för att respondenten skall klara frågan (t.ex. svara "inga problem"). Detta låter invecklat och det är det. Det finns dock datorprogram som sköter analysen och ger Rasch score för alla frågor och svarsalternativ. Denna Rasch score är intervalldata och kan behandlas med parametrisk statistik.

5. Vilka krav på validitet skall ställas?

Uttrycket "validerat frågeformulär" förekommer ofta och kan ha många betydelser. Det finns flera olika typer av validitet, och egentligen bör man upplysa om hur ett frågeformulär är validerat⁶. Språkbruket när det gäller validitet är engelskt men i många fall finns svenska motsvarigheter till validitetsuttrycket.

Några validitetsbegrepp

Face validity Klinikerns uppfattning om frågorna är relevanta för ändamålet.

Content validity Uttrycker till vilken grad frågorna speglar den funktion man vill mäta. Det finns ingen formell test för detta utan avgörande är att man använt korrekta metoder för att skapa frågeformuläret.

Construct validity I princip samma som **Content validity**.

Criterion validity Traditionell definition av validitet och innebär att det aktuella frågeformulärets resultat är korrelerat till existerande mått som mäter samma sak. Om man t.ex. vill mäta patientens självskattade synfunktion så vill man att det skall finnas ett visst samband mellan frågeformulärets resultat och synskärpan som ett kliniskt mått. Det skall inte vara en total överensstämmelse, i så fall är ett av testerna överflödigt. Det skall heller inte saknas samband, då är risken stor att man mäter något annat med frågeformuläret.

Convergent validity Att frågeformulärets resultat korrelerar med andra liknande frågeformulärs resultat.

Discriminant validity Innebär till vilken grad ett frågeformulär får resultat som skiljer sig ifrån andra frågeformulär som inte mäter samma sak.

Concurrent validity Ett frågeformulärs förmåga att skilja på grupper av respondenter som det är avsett att kunna skilja på. Mäter man t.ex. före och efter ett ingrepp eller annan åtgärd som har betydelse för funktionen skall frågeformuläret kunna påvisa skillnad i resultat. När man säger att ett frågeformulär är validerat brukar man inkludera att det testats mot någon viktig medicinsk variabel eller ett liknande frågeformulär och uppvisar korrelation. Dessutom brukar man inkludera en test på att alla frågor bidrar till att mäta samma funktion (intern konsistens) och att frågeformuläret ger stabila svar vid upprepade tester (test-retest reliability test). För dessa tester, se nedan.

6. Vilka krav på reliabilitet skall ställas?

Kravet på reliabilitet innebär krav på tillförlitlighet. Hit brukar man räkna:

- Vilken precision frågeformuläret har
- Vilken stabilitet frågeformuläret har vid upprepade test – test-retest-reliability test
- Hur frågeformuläret uppträder i olika presentationsformer
- ”Differential item functioning – DIF”
- “Intern consistency” – Chronbach’s α (alfa)
- Vilken precision frågeformuläret har.

Precision är en fundamental egenskap för alla typer av mätinstrument. Saknas precision blir mätningarna nonsens. En våg kan inte låta alla personer väga lika mycket, en linjal kan inte låta alla personer vara lika långa. På samma sätt måste ett frågeformulär kunna särskilja på olika nivåer av respondenter. Vid Rasch-analys mäts precisionen i "Person separation" och "separation reliability". Se vidare om **Rasch-analys**.

Vilken stabilitet frågeformuläret har vid upprepade test – **test-retest-reliability test**¹⁰. Denna test utförs på ett begränsat antal respondenter (50-100 brukar räcka). Respondenten får fylla i frågeformuläret två gånger med cirka 14 dagar emellan. Principen vid upprepat test är att tillståndet inte skall hinna ändras och att respondenten inte omedelbart skall minnas svaren från det första testet.

Hur frågeformuläret uppträder i olika presentationsformer. Ett frågeformulär kan presenteras som en intervju, ett papper att fylla i, ett dokument på webben, ett dokument på en pekskärm eller i en mobiltelefon. Vet man att det skall användas flera presentationsformer är det lämpligt att testa hur enskilda individer svarar vid de aktuella presentationsformerna.

"Differential item functioning – DIF". Detta är ett test som utförs i samband med en Rasch-analys⁹. Testet visar om subgrupper i populationen (t.ex. kvinnor, respondenter från en viss trakt, äldre etc.) svarar signifikant olika på enskilda av frågeformulärets frågor. DIF bör vara obetydlig så att svaren kan analyseras likvärdigt oavsett gruppstillhörighet.

Intern konsistens (egentligen borde det heta intern konsekvens, men det engelska internal consistency har kommit att översättas så) – Chronbach's alfa⁸. En form av reliabilitet avser hur väl olika delar av måttet mäter samma sak. Till exempel kan man önska att de olika frågorna ("items") i ett sjukdomsspecifikt frågeformulär mäter den undersökta funktionen på samma sätt. En metod för att mäta detta är split-half-metoden, som innebär att man delar upp frågorna i två hälften - till exempel varannan fråga eller första och andra hälften av testet - och beräknar sedan korrelationen mellan de båda hälfterna. Ett annat mått på den interna konsistensen är Cronbach's α (alfa), som är ett sammanfattande korrelationsmått mellan samtliga items. I teorin är Cronbach's α ett mått på alla split-half utförda på en gång. Cronbach's α skall helst anta ett värde mellan 0,80 och 0,90. Blir värdet högre talar det för att någon eller några frågor är överflödiga. Blir värdet för lågt speglar frågorna olika saker.

7. Vilka krav på känslighet (för förändringar) skall ställas?

Det finns några begrepp som uttrycker olika former av känslighet.

Minimum Clinically Important Difference – MID¹¹

Detta begrepp avser den minsta score i frågeformuläret som respondenten kan uppfatta som av värde. Man testar detta ofta genom att ställa extra frågor utanför frågeformuläret och försöker sedan korrelera svaret med förändringen i score. Exempel: För en patient som genomgått en behandling så mäter frågeformuläret en skillnad mellan före och efter behandlingen i en viss score. Samtidigt frågar man samma patient vilken effekt behandlingen haft: sämre, status quo, lite bättre, mycket bättre etc.

Responsiveness – effect size¹²

På svenska används uttrycket effektstorlek. Det finns olika tekniker för att mäta effektstorlek. Den vanligaste beräkningen är att effektstorleken uttrycker kvoten mellan medelförändringen i score (före-efter) dividerat med standard deviationen

av alla frågeformulärs svar (både före och efter tillsammans). En effektstorlek på 0 – 0,20 bedöms som liten, 0,20 – 0,80 som medel och över 0,80 som stor.

8. Hur skall utvärderingen göras begriplig?

Man får relatera frågeformulärets score till begripliga storheter. Man kan beskriva vad som är meningsfulla skillnader i score och man kan beskriva vilken score som motsvarar t.ex. vissa besvär och stora besvär.

Exempel: I ett Rasch-analyserat frågeformulär kan en viss fråga se ut som 2B ovan. Poängen 1-4 omvandlas vid analysen till -3,54; -0,51; 1,15; 3,14 (vid denna analys innebär negativa värden mindre problem och positiva värden större problem. Enheten är logits). Om den undersökta populationens genomsnitt är 1,08 före operation och -3,20 efter operation så har motsvarande svarsalternativ gått från ”stora problem” före operation till ”inga problem” efter operation.

Ytterligare ett sätt är att beskriva typiska skillnader efter en intervention eller demografibaserade skillnader.

Exempel: Ett annat sätt är att beskriva en representativ population och vilken score denna får i genomsnitt. Ser man till frågeformulärets alla frågor så kan ett genomsnitt före operation på 0,5 ändrats till -3,2 efter operation. Dessa nivåer för hela populationen kan enskilda vårdgivare jämföra sig med. I Nationella Kataraktregistrets analys av PROM så fann man att en privat operationsklinik i Stockholm opererar patienter med en genomsnittlig score på -2,28 medan motsvarande score på en landsortsklinik med låg operationsvolym sedan många år var 1,04. Det betyder att på den ena kliniken opereras patienter med ”vissa problem” – ”inga problem” i genomsnitt medan den andra kliniken opererar patienter med ”stora problem” – ”mycket stora problem” i genomsnitt.

9. Hur skall man länka samman kliniska utfallsdata och patientrapporterade utfallsdata?

Om man vill använda ett sjukdomsspecifikt frågeformulär till att förbättra vården så är denna fråga naturligtvis väsentlig. I princip kan man tänka sig tre olika analyser¹³:

1. Vilka kliniska parametrar är relaterade till resultatet av frågeformuläret?
2. Vilka kliniska parametrar är relaterade till en förbättring eller försämring av frågeformulärets score?
3. Vad kännetecknar den situation då kliniskt utfall och patientrapporterat utfall (frågeformulärets score) divergerar? Särskilt intressant är då kliniska parametrar förbättras men frågeformulärets score försämras.

Avslutande kommentarer

1. Finns ett färdigt validerat frågeformulär som passar Dina syften: använd det! Men se till att det är modernt och testat med s.k. Rasch-analys eller med annan item-respons teknik.
2. Är frågeformuläret på ett annat språk, gör en översättning enligt standardprocedur. Kontrollera först att det inte begränsas av licenser eller copyright.
3. Vill Du skapa ett helt nytt frågeformulär, tag gärna kontakt med specialister på att konstruera sjukdomsspecifika frågeformulär.
4. Se till att syftet är helt klart, det är ett villkor för att komma rätt med content validity.
5. För validitetstest av ett nytt frågeformulär rekommenderas s.k. Rasch-analys.

Referenslista

1. White M, Elander G. Translation of an instrument. The US-Nordic Family Dynamics Nursing Research Project. *Scand J Caring Sci.* 1992;6(3):161-4.
2. Su CT, Parham LD. Generating a valid questionnaire translation for cross-cultural use. *Am J Occup Ther.* 2002 Sep-Oct;56(5):581-5.
3. Richardson JR, Peacock SJ, Hawthorne G, Iezzi A, Elsworth G, Day NA. Construction of the descriptive system for the Assessment of Quality of Life AQoL-6D utility instrument. *Health Qual Life Outcomes.* 2012 Apr 17;10:38. doi: 10.1186/1477-7525-10-38.
4. Rasch BG. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Denmark, Danmarks Paedagogiske Institut, 1960
5. Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Med Care* 2004; 42(suppl 1):I25–I36
6. Pesudovs K, Burr JM, Harley C, Elliott DB. The development, assessment, and selection of questionnaires. *Optom Vis Sci* 2007; 84:663–674
7. Wolfle D, Likert R, et al. Standards for appraising psychological research. *Am Psychol.* 1949 Aug;4(8):320-8.
8. Risser NL. Development of an instrument to measure patient satisfaction with nurses and nursing care in primary care settings. *Nurs Res.* 1975 Jan-Feb;24(1):45-52.
9. Orlando Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. Application to the Mini-Mental State Examination. *Med Care.* 2006 Nov;44(11 Suppl 3):S134-42.
10. Yen M, Lo LH. Examining test-retest reliability: an intra-class correlation approach. *Nurs Res.* 2002 Jan-Feb;51(1):59-62.
11. Todd KH, Funk JP. The minimum clinically important difference in physician-assigned visual analog pain scores. *Acad Emerg Med.* 1996 Feb;3(2):142-6.
12. Bird KD, Hall W. Statistical power in psychiatric research. *Aust N Z J Psychiatry.* 1986 Jun;20(2):189-200.
13. Lundström M, Stenevi U. Analyzing patient-reported outcomes to improve cataract care. *Optom Vis Sci.* 2013 Jun 6. [Epub ahead of print].

APPENDIX I

Arbetsmodell för analys av frågeformulär med Rasch-analys

I nedanstående exempel har använts WINSTEPS Version 3.75.0

(www.winsteps.com)

Bearbetning av grunddata

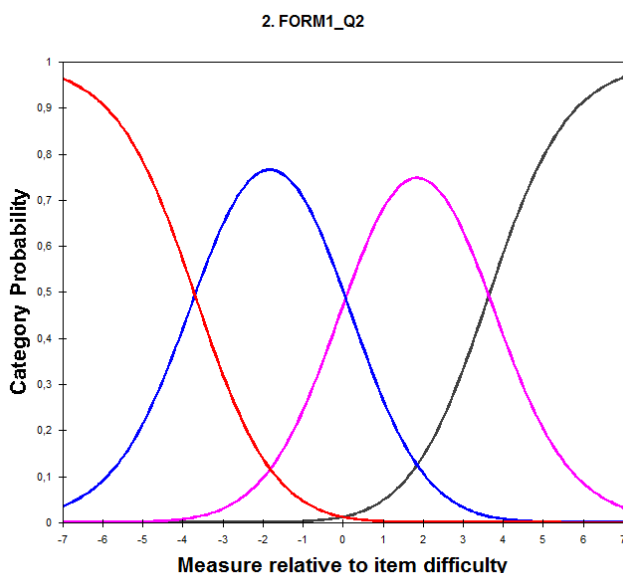
I regel finns en excelfil med data. I denna kan fall och enkätsvar vara arrangerade olika. För en Rasch-analys måste varje enkätsvar stå på en egen rad. Om det finns flera enkätsvar (t.ex. före och efter en intervention) på samma rad motsvarande patienten får enkätsvar 2 flyttas till en helt ny rad med tillhörande ID-data. Det betyder att en individ kan förekomma som flera "cases" i Rasch-filen med olika enkätsvar (t.ex. före och efter en intervention).

För själva Rasch-analysen behövs förutom enkätsvaren ett antal personvariabler. Det kan handla om ID i form av löpnummer, kön, ålder och om det är enkät 1, 2 etc. samt någon enstaka viktig variabel av typen komorbiditet eller klinik/nationalitet om det är en multicenterstudie. Dessa personuppgifter används i analysen för att se om någon undergrupp svarar systematiskt annorlunda än resten, s.k. DIF.

RASCH-ANALYSEN

Steg 1, Används svarsalternativen på ett ordnat sätt?

Det första man ser på är om svarsalternativen är ordnade, d.v.s. att de besvaras på ett logiskt sätt som man tänkt och inte helt tillfälligt. Man ser på två saker, dels de s.k. **category probability curves**, dels hur mycket varje svarsalternativ används (tak- eller golveffekt). Om svarsalternativen inte används på ett ordnat sätt fungerar inte frågeformuläret och det är ingen idé att fortsätta analysen förrän man slagit samman svarsalternativ och provat igen om svaren är ordnade. När detta är klart övergår man till steg 2.



Figur 1. Category probability curves. 4 svarsalternativ, 3 trösklar mellan svaren.

Steg 2, Frågeformulärets precision

Nu har man ett frågeformulär med ordnade svarsalternativ. I detta steg ser man på vilken diskrimineringsförmåga (precision) frågeformuläret har. Precision är en fundamental aspekt på mätning; huruvida en mätmetod kan särskilja objekt längs sin mätskala. Detta betyder i praktiken för frågeformulär hur många nivåer av

respondenternas förmåga som instrumentet kan särskilja. Man brukar kräva åtminstone 3 nivåer. I Rasch-analysen så mäts diskrimineringsförmågan i s.k. person separation och separation reliability coefficient. Person separation bör vara över 2,0; över 2,5 är utmärkt. Separation reliability coefficient bör vara minst 0,8. I steg 2 får man alltså reda på om instrumentet är meningsfullt eller inte.

Steg 3, ”Item fit statistics”

Nästa steg innebär att man ser på s.k. item fit statistics, d.v.s. om varje fråga ”passar in” i frågeformuläret. Item fit statistics indikerar om frågeformuläret mäter ett enda koncept (en enda underliggande funktion), dvs. är endimensionellt. Ett frågeformulär är i regel tänkt att spegla en underliggande funktion genom att olika frågor belyser funktionen från olika vinklar. Det kan handla om t.ex. förmågan att klara dagliga aktiviteter med en viss funktionsnedsättning eller tryggheten att klara sociala situationer med tandimplantat eller kontaktlinser etc., osv. Om alla frågor speglar samma underliggande funktion sägs instrumentet vara endimensionellt. Om frågorna speglar flera funktioner, t.ex. smärta och social trygghet är instrumentet tvådimensionellt. Det finns generiska instrument som är flerdimensionella.

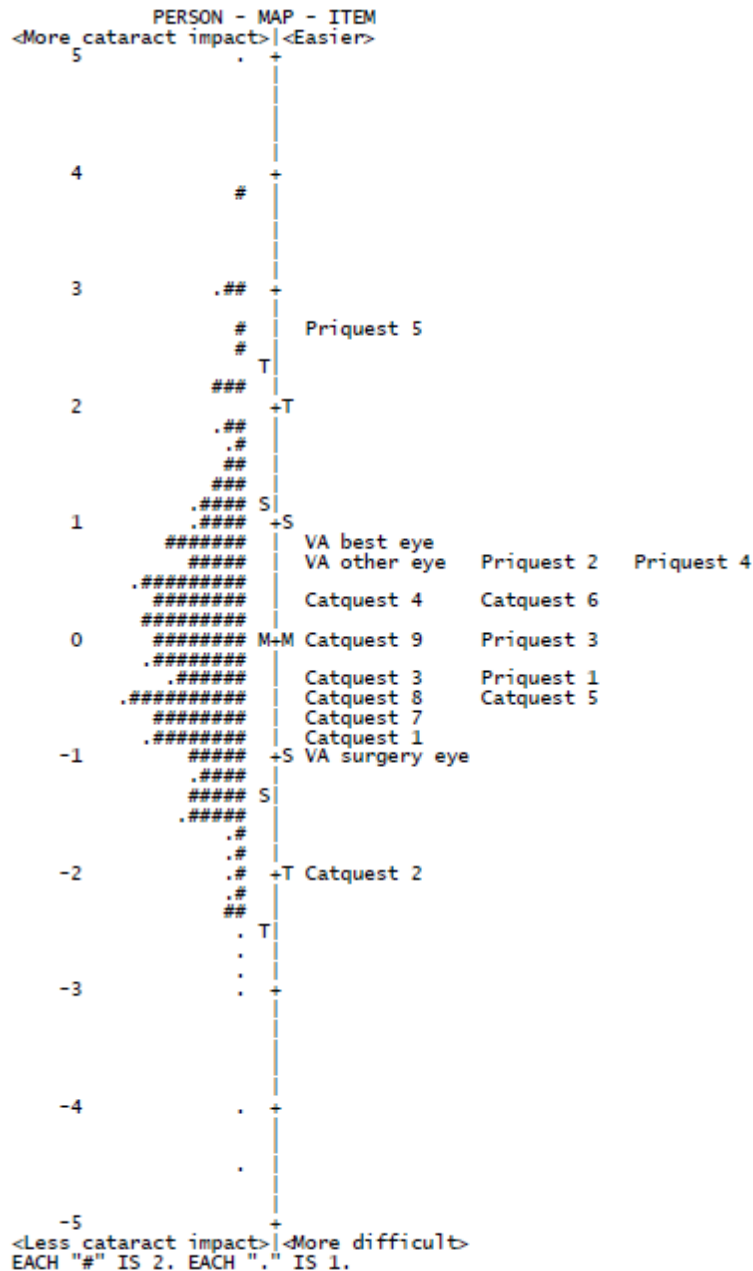
Item fit statistics handlar om ”Infit statistics” (informations-viktad fit statistics) vilket innebär huruvida de observerade svaren passar in i Rasch-modellen och ”outfit statistics” (outlier sensitive statistics) som visar om det föreligger många oväntade svar. Båda dessa värden skall vara 1 för perfekt passande items. Man sätter gränser för vilka avvikelser som tolereras. I vissa sammanhang hävdar man att infit och outfit skall ligga mellan 0,8 och 1,2 för att ett item skall godkännas, inom sjukvård brukar man acceptera 0,7 – 1,3 eller t.o.m. 0,6 – 1,4. Ibland sätter man striktare gränser för infit (0,8-1,2) än för outfit (0,7-1,3); värden visas som exempel. Låga värden (t.ex. under 0,7) tyder på att frågan (item) tillför överflödigt information, dvs. information som redan ges av övriga item. Höga värden (t.ex. värden över 1,3) tyder på att frågan mäter något annat än vad övriga frågor gör.

I detta steg tar man bort items som inte passar in av ovanstående skäl, s.k. misfitting. Man tar bort ett item i taget och gör om analysen eftersom fit statistics kan ändras om en fråga tas bort. Man brukar börja ta bort det item som är mest misfitting.

Steg 4, Huruvida frågeformuläret passar för målgruppen (Targeting)

Targeting syftar på i vilken utsträckning frågornas svårighetsgrad matchar respondenternas förmåga. I idealfallet ligger medelvärdet för patienternas förmåga och frågornas svårighetsgrad på samma nivå. Man kan tolerera en skillnad på en logit. Targeting studeras på en s.k. person-item map. Denna visar förutom medelvärdet också om flera items hamnar på samma svårighetsgrad eller om det finns stora luckor eller tak- respektive golveffekter.

Figur 2 på nästa sida visar en idealisk person-item map.



Figur 2. Person-item map visande en god överensstämmelse mellan frågornas svårighetsgrad (höger) och respondenternas förmåga (vänster). Till vänster visas alltså respondenterna och varje # innebär 2 personer. Nedtill på "linjalen" är respondenter som svarat att de inte har speciellt stora problem att utföra aktiviteterna och upptill är sådana som uttrycker större problem. Frågornas placering till höger om "linjalen" uttrycker hur lätt (upptill) eller svårt (nedtill) som populationen av respondenter haft för att utföra respektive aktivitet.

APPENDIX II

Analysmodeller för att åstadkomma bättre vård baserat på patientrapporterat utfall

Rent generellt kan sambandet mellan kliniska utfallsdata och patientrapporterat utfall analysera på tre olika sätt. I nedanstående beskrivning används uttrycken åtgärd, behandling och terapi. De är alla synonymer för en aktivitet riktad mot patienten för att förbättra dennes hälsa.

1. **Analys av sambandet mellan kliniska data och patientrapporterat utfall (PROM).**

Denna analys gäller vilken påverkan enskilda kliniska parametrar har på patientrapporterat utfall. Inom verksamheter där man studerar åtgärder (kirurgi, medicinsk behandling etc.) kan PROM påverkas av utgångsstatus, åtgärdsrelaterade faktorer och sena resultat relaterade till åtgärden. Kliniska parametrar kan analyseras i relation till förändringen av patientrapporterat utfall (från före till efter en åtgärd) eller den postterapeutiska nivån på patientskattad funktion. I nedanstående exempel fokuseras på påverkan av kliniska parametrar på förändringen av patientrapporterad synfunktion efter kataraktoperation.

2. **Analys av vilka kliniska parametrar som är relaterade till ett bra eller dåligt patientrapporterat utfall.**

Det samband som är intressant är mellan kliniska parametrar och patientrapporterat dåligt utfall. I typiska fall görs en riskfaktor analys som resulterar i odds ratio för ett dåligt patientrapporterat utfall vid närvaro av en given faktor (t.ex. komorbiditet). Inom kataraktkirurgin inkluderar potentiella riskfaktorer för ett patientrapporterat dåligt utfall ögats preoperativa tillstånd, förekomst av andra ögonsjukdomar, vissa demografiska egenskaper och komplikationer.

3. **Analys av vilka faktorer som är relaterade till om kliniskt utfall (COM) och patientrapporterat utfall (PROM) överensstämmer eller inte.**

Detta kan göras på två olika sätt: att analysera förbättringar efter en åtgärd jämfört med före eller genom att definiera ett bra kliniskt resultat och ett bra patientrapporterat resultat och huruvida utfallet motsvarar denna standard. Oavsett om vi analyserar förbättring eller ett gott postterapeutiskt status kan fyra olika kombinationer av utfall modelleras (+ = förbättring/bra utfall, - = försämring/dåligt utfall):

COM+PROM+	COM-PROM+
COM+PROM-	COM-PROM-

Om vi jämför förbättringen efter en åtgärd så kan en positiv COM betyda bättre funktion efter en åtgärd än före och en positiv PROM kan betyda en bättre frågeformulärpoäng efter en åtgärd än före (=färre upplevda besvär att utföra aktiviteter i det dagliga livet). Om vi jämför status efter åtgärden så kan en bra COM betyda en postoperativ funktion motsvarande ett visst mått, och en bra PROM kan betyda få problem att utföra dagliga aktiviteter. I vårt exempel vill vi undersöka skillnaden i COM och PROM mellan före en åtgärd och efter en åtgärd. Skälet till att vi väljer denna modell är att individuella skillnader i språkbruk neutraliseras genom att vi jämför en persons svar före och efter en åtgärd. Detta gör varje respondent till sin egen kontroll.

Tabell 1. Exempel på en analys av sambandet mellan kliniska data och PROM.

Lineär regressionsanalys, stegvis. Beroende variabel: person Rasch score förbättring efter kataraktoperation. B = b koefficient i regressionen.

Oberoende variabel	B	Standardiserad koeff. (Beta)	p	95% konfidens intervall för B
Preop. Rasch score	0.698	0.564	<0.001	0.677 – 0.718
Postop. synskärpa, operationsögat	3.375	0.298	<0.001	3.168 – 3.581
Preop. synskärpa, bästa ögat	0.802	0.076	<0.001	0.610 – 0.993
Preop. synskärpa, operationsögat	-0.866	-0.069	<0.001	-1.088 – -0.644
Okulär komorbiditet	-0.366	-0.067	<0.001	-0.459 - -0.273

R square: 0.365.

Exkluderade i analysen:

Ålder	0.284
Kön (kvinna=1)	0.117
Kapselkomplikation	0.079
“Biometry prediction error”	0.780

Konklusion: Flera kliniska variabler är relaterade till det patientrapporterade utfallet och Beta visar ordningen från starkast till svagast samband.

Tabell 2. Exempel på en analys av vilka kliniska parametrar som är relaterade till ett dåligt patientrapporterat utfall.

Logistisk regressionsanalys, (enter). Beroende variabel: försämrad synfunktion efter operation (1) eller inte (0) (sämre person Rasch score efter operation eller inte). B = b koefficient i regressionen, SE = standard error, Sig. = p-värde för signifikans, Exp(B) = exponentiering av b (odds ratio), 95 % CI = 95 % konfidens intervall, BPE = ”biometry prediction error”.

Variabel	B	SE	Sig.	Exp(B)	95% CI för Exp(B)
Preop. Rasch score	-.405	0.021	<0.001	.667	.641 – .695
Preop. synskärpa, bästa öga	-1.453	0.173	<0.001	.234	.167 – .328
Kön (kvinna = 1)	0.269	0.082	0.001	1.309	1.115 – 1.537
Okulär komorbiditet	0.588	0.080	<0.001	1.800	1.537 – 2.106
Kapsel komplikation	0.624	0.306	0.041	1.866	1.024 – 3.399
BPE ≥ 3D	0.976	0.387	0.012	2.653	1.242 – 5.665

Exkluderad i modellen

Anisometropia ≥ 3D	0.144
Ålder	0.825

Konklusion: Flera kliniska variabler är relaterade till dåligt utfall och Exp. (B) visar odds ratio för vilken risk enskilda variabler innebär. I detta exempel är ett lågt värde på preoperativt Rasch score och ett lågt värde på preoperativ synskärpa bästa ögat (båda är kontinuerliga variabler) riskfaktorer, men tabellen visar att höga värden på dessa variabler utgör ett skydd (därav odds ratio under 1,0).

Tabell 3. Analys av enskilda item för den grupp där det kliniska utfallet var gott och det patientrapporterade utfallet dåligt. Medelvärde av item Rasch score före och efter operation. Två grupper visas: COM+PROM- och alla övriga. Item som representerar aktiviteter som kräver syn på nära håll är märkta med fetstil. COM = clinical outcome measure, PROM = patient reported outcome measure.

Item	Före operation		Efter operation	
	COM+ PROM-	All övriga	COM+ PROM-	All övriga
Problem i allmänhet	-1.68	-0.43	-1.36	-3.77
Nöjdhet med synen	-1.58	-0.23	-1.77	-4.03
Läsa tidningar	-2.26	-0.85	-0.87	-3.4
Känna igen ansikten	-2.31	-1.04	-2.1	-3.16
Se prislappar	-2.41	-0.93	-0.83	-3.36
Se att gå på ojämn mark	-2.13	-1.1	-1.6	-3.24
Se att handarbete	-2.2	-0.78	-0.79	-3.61
Läsa text på TV	-2.11	-0.71	-1.89	-3.64
Se att utföra en önskad aktivitet eller hobby	-2.33	-0.77	-1.26	-3.63

Konklusion: Tabellen visar att enskilda item för aktiviteter som kräver syn på nära håll innebär föga besvär före operation men stora besvär efter för gruppen COM+PROM-. För alla övriga är det tvärtom. Således ser patienter i COM+PROM- gruppen bra på nära håll före operation, men dåligt efter. För att förbättra utfallet måste man tillförsäkra att dessa patienter får bra läskorrektion efter operationen.

Patientens synpunkter och erfarenheter spelar en central roll i utveckling av vården. Det räcker dock inte att gissa sig till vad patienterna tänker, tycker och känner. Ett sätt är att via frågeformulär fråga patienten om sin upplevelse av hur man klarar dagliga aktiviteter och subjektivt uppfattar resultatet av vård eller behandling.

En handbok för att utveckla ett sjukdomsspecifikt frågeformulär är sammanställt av professor Mats Lundström i samarbete med;
Registercentrum Syd, EyeNet Sweden
Blekingesjukhuset
371 85 KARLSKRONA

Frågor om handboken besvaras av
Professor Mats Lundström
mats.lundstrom@karlskrona.mail.telia.com
eller RC Syd Karlskrona, EyeNet Sweden.



Beställ eller ladda ner handboken från
www.eyenetsweden.se

Kontaktinformation

Thomas Troëng
thomas.troeng@ltblekinge.se

Susanne Albrecht
Susanne.albrecht@ltblekinge.se

Eva Wendel
eva.wendel@ltblekinge.se

