Måling af lægers kompetencer

Ebbe Thinggaard[1, 2], Ann Sofia Thomsen[1, 3], Lotte O'Neill[4] & Lars Konge[1]

STATUSARTIKEL

1) Copenhagen Academy for Medical Education and Simulation, Region Hovedstaden
2) Gynækologisk og Obstetrisk Afdeling, Hvidovre Hospital
3) Copenhagen Academy for Medical Education and Simulation, Center for HR
4) Universitets-pædagogik, Syddansk Universitet

Ugeskr Læger 2018;180:V01180059

Igennem en årrække har man i Danmark haft et ønske om at øge effektiviseringen og forbedre kvalitetssikringen af sundhedsvæsenet. Dette har ført til et øget antal målinger, hvor man har indsamlet data på forskellige parametre såsom patienttilfredshed, komplikationsrater, tidsforbrug etc. Målingerne bruges til at monitorere og evaluere hospitalsafdelingers produktivitet og kvalitet. F.eks. bruges målinger som knivtid ved en operation til at styre afdelingens udførelse af operationer. Der opstår dog et problem, når disse målinger tages ud af en kontekst og f.eks. bruges til at vurdere lægers færdigheder med. Den enkelte læges forbrug af tid kan ikke anvendes til at afgøre, om vedkommende er i stand til at varetage en given funktion, når der ikke eksisterer evidens for en sådan antagelse. Det er uforsvarligt og potentielt skadeligt at styre efter måleresultater, som har utilstrækkelig validitet [1, 2].

I den lægelige videreuddannelse har man indført kompetencebaseret medicinsk uddannelse, hvor krav om længde af ansættelse og antal udførte procedurer ikke står alene, når det skal vurderes, om læger er kompetente til at varetage deres arbejde. Kompetencebaseret medicinsk uddannelse beror på brugen af kompetencevurderinger, hvormed man beskriver eller måler en læges viden, færdigheder og/eller holdninger [3-5]. Kompetencevurderinger kan bruges til beskrivelse eller måling af lægens kompetencer og baseres typisk på et vurderingsværktøj [3, 6, 7]. Før at man kan sige noget meningsfyldt om en læges kompetencer, kræves der vurderinger, som er understøttet af evidens. Evidens, der er undersøgt systematisk og giver grundlag for en retfærdig vurdering [1, 8, 9]. Ofte anvendes målinger dog uden tanke på, om man egentligt måler det, man ønsker, og konklusioner om kompetencer drages i nogle tilfælde uden belæg [10, 11]. Der er en manglende forståelse af, hvordan man udvikler og undersøger målemetoder, og dette begrænser diskussionen om meningsfyldte målinger. Særligt når det drejer sig om fortolkning af målinger og spørgsmål om, hvilke konklusioner der kan drages på baggrund af en given måling [12, 13].

Denne artikel indeholder en beskrivelse af kompetencevurdering i den lægelige videreuddannelse, en gennemgang af berettigede validitetskrav til kompetencevurdering med udgangspunkt i moderne validitetsteori og eksempler på validitetsevidens. Formålet med artiklen er at nuancere og diskutere udførelsen og anvendelsen af kompetencevurderinger, øge forståelsen af kravet om validitetsevidens og forbedre lægers evne til at forholde sig kritisk til kvaliteten af kompetencevurderinger.

HOVEDBUDSKABER

- Kompetencevurderinger af lægers færdigheder beror på brugen af vurderingsværtøjer, hvormed man kan måle eller beskrive lægers viden, færdigheder og/eller holdninger.
- Vurderingsværktøjer bør være understøttet af validitetsevidens, så man sikrer, at man med vurderingsværtøjet giver en rimelig og retfærdig vurdering af en læges færdigheder.
- Undersøgelse af validitetsevidens er en kontinuerlig videnskabelig proces, hvor evidens undersøges systematisk, og der bruges en nutidig og accepteret teoretisk tilgang til validitet.

KOMPETENCEVURDERINGER

Kompetencebaseret medicinsk uddannelse har vundet indpas verden over siden offentliggørelsen af Flexner-rapporten fra 1910, hvor han beskrev kvaliteten af den lægelige uddannelse i USA [14]. Her fremgik det, at der var et behov for at fokusere på kompetencebaseret uddannelse for at sikre, at læger opnåede den nødvendige viden samt de nødvendige færdigheder og holdninger for at kunne bestride deres erhverv på et tilfredsstillende niveau [14, 15]. I Danmark har denne kompetencebaserede tilgang dannet grundlag for de syv lægeroller, som anvendes til at beskrive mål for lægers kompetencer [16]. De syv lægeroller stammer fra Canada, hvor man har udviklet CanMEDS, der beskriver de lægeroller, der er vigtigst for patienterne [17]. Til vurdering af målopfyldelse i de syv lægeroller anvendes kompetencevurderinger, der er baseret på brugen af vurderingsværktøjer. Moderne validitetsteori er den gennemgående teoretiske ramme, som bruges til at underbygge og forsvare anvendelsen af disse værktøjer. Den internationale standard for test i uddannelse og psykologi indledes med en beskrivelse af begrebet validitet (**Figur 1**) [9].

Der er med andre ord behov for tilstrækkelig evidens, før at man kan tage stilling til, om en given måling måler det, man antager, at den måler. Der eksisterer formelle krav til, hvilken validitetsevidens der bør foreligge, når man udvikler og undersøger vurderingsværktøjer [9]. Det påhviler de testansvarlige at sikre, at tilstrækkelig validitetsevidens for diverse bedømmelser foreligger.

VALIDITETSKRAV I KOMPETENCEVURDERINGER

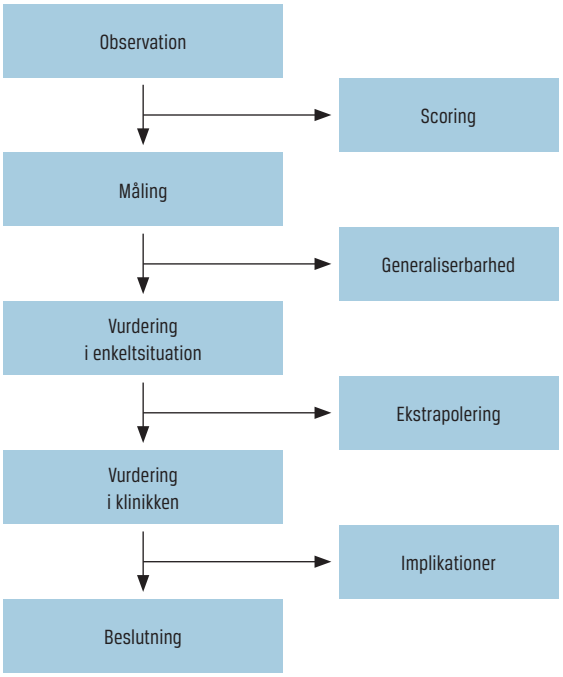
De fleste har en forståelse for validitet, der stammer fra brugen af diagnostiske test. Her er der ofte tale om et binært resultat, hvor patienter enten er syge eller raske. Resultatet af testen kan opstilles i en to gange to-tabel, og på baggrund af det kan sensitivitet og specificitet udregnes. Det er dog mere komplekst, når man taler om kompetencevurderinger. Viden, færdigheder og holdninger er kontekstafhængige, og der foreligger som regel ingen guldstandard. Der eksisterer ingen endelig sand værdi til afgørelse af, om en læge er kompetent. Der eksisterer heller ikke en entydig metode til måling af dette. Resultatet af kompetencevurdering er et konstrueret begreb og beskriver ikke en endelig sand værdi. Der er derfor et væsentligt behov for at sikre, at fortolkningen af testresultater – og beslutninger taget på baggrund af denne fortolkning – understøttes af evidens. Når man udvikler eller undersøger et vurderingsværktøj, er det vigtigt at sikre, at værktøjet ikke blot beskriver eller måler det, man ønsker, men at man kan retfærdiggøre sin fortolkning af resultatet [18]. Det vil sige, at ikke bare målingen, men den beslutning og de handlinger, man foretager på baggrund af en måling, er understøttet af evidens. Man starter med at begrunde kompetencevurderingen: Hvad skal målingerne bruges til? Hvilke fortolkninger forventer man at gøre på baggrund af prøveresultatet? Hvilke beslutninger skal der kunne tages? Efterfølgende udfærdiges undersøgelser af evidensen for de fremsatte fortolkninger, og man starter med at udfordre de dårligst underbyggede argumenter i fortolkningskæden i disse undersøgelser [19]. Validitetsevidensen, som undersøges for de fremsatte fortolkninger, kan vedrøre: 1) *Scoringsproces*: Hvad er evidensen for bedømmelsesprocessen? Dvs. det, der foregår fra observation af en læges præstation til registrering af en score for præstationen. Hvad er evidensen for indholdet af vurderingsværktøjet? Og bruges vurderingsværktøjet efter forskriften? 2) *Generaliserbarhed*: Måles lægen i tilstrækkeligt mange forskellige situationer med en tilstrækkelig blanding af patientcases og med det rette antal bedømmere, til at vurderingen (foretaget med det specifikke værktøj) er tilstrækkelig repræsentativ for lægens generelle præstationer? 3) *Ekstrapolering*: Er scoren associeret med det begreb, som man antager, at man tester? F.eks. kirurgisk kompetence. Er der i bedømmelsen konkurrerende forklarende begreber i spil, som kan underminere fortolkningen af scoren som antaget? 4) *Implikationer*: Er følgerne af beslutninger eller handlinger foretaget på baggrund af scoren tilstrækkeligt forsvarlige og tilsigtede? [1] (**Figur 2**).

FIGUR 1

Begrebet validitet dækker over, i hvilken grad evidens og teori understøtter fortolkningen af testresultater i forhold til brugen af testen. Overvejelser om validitet er derfor selve fundamentet for udvikling og evaluering af tests. Valideringsprocessen består i at indsamle relevant evidens, der kan give en tilstrækkelig videnskabelig basis for den ønskede fortolkning af testresultater. Det er fortolkningen af et testresultat, der evalueres, ikke testen i sig selv. Når en test fortolkes på mere end en måde … skal hver fortolkning evalueres for sig.

Definition af begrebet validitet [9].

FIGUR 2



Oversigt over antagelser.

EKSEMPEL PÅ KOMPETENCEVURDERING AF LÆGER

I dette afsnit vil vi prøve at konkretisere en kompetencevurderingssituation med en læge som et eksempel på den teoretiske tilgang til validitet, som er beskrevet ovenfor. Et sådant eksempel kunne være kompetencevurdering af en uddannelseslæges kirurgiske færdigheder i videoassisteret torakoskopisk (VATS) lobektomi, som kan være indiceret i behandlingen af lungecancer. Udarbejdelsen af et bedømmelsesredskab for proceduren vil kræve evidens i alle fire ovennævnte kategorier.

Som validitetsevidens for *scoringsprocessen* kunne man f.eks. forestille sig, at man fremskaffede dokumentation for vurderingsredskabets indhold (testitems) ved at lave et systematisk review af den eksisterende forskningslitteratur eller ved at gennemføre en delphiprocedure med eksperter i VATS-lobektomi. Her ville det også være en fordel at kunne dokumentere træningen af bedømmerne og resultaterne af en pilottest af redskabet. Ved en pilottest af vurderingsredskabet tjekker man, om instrumentet også i praksis anvendes som foreskrevet.

I evidenskategorien *generaliserbarhed* kan en generaliserbarhedskoefficient bruges som evidens. En ge-

Tid bruges ofte til måling af lægers effektivitet, men det er en uhensigtsmæssig måde at måle lægers kompetencer på.

neraliserbarhedskoefficient som tager højde for alle vigtige og samtidigt virkende kilder til varians. I eksemplet her ville de forskellige items i redskabet, de forskellige patientcases (forskellige sværhedsgrader) og de forskellige bedømmere alle være samtidigt virkende kilder til varians i scorerne og derfor faktorer, som burde indgå i en og samme generaliserbarhedskoefficient. Ved udredningen af generaliserbarheden har man også mulighed for at bestemme den sammensætning af patientcases og bedømmere, som er nødvendig for, at scoren kan siges at være tilstrækkeligt repræsentativ for uddannelseslægens svingende præstationer i proceduren.

Til støtte for antagelsen om, at man med scoren kan *ekstrapolere* til uddannelseslægens færdighedsniveau, kan man f.eks. fremskaffe evidens for, at man med redskabet kan differentiere mellem novicer og eksperter som antaget. Eller endnu bedre: Man kan forsøge at fremskaffe evidens for, at patienter, som er opereret af kirurger, der generelt scorede højt på redskabets items, klarer sig bedre postoperativt eller har færre komplikationer end patienter, som er opereret af kirurger, der scorede lavt [20-22]. I analyserne af sådanne sammenhænge er det vigtigt at kontrollere for konkurrerende forklarende faktorer i forhold til patienternes efterfølgende sygdomsforløb.

Hvis scorerne f.eks. skal bruges til at afgøre, om uddannelseslægens kompetenceniveau er tilstrækkeligt eller ej, skal man kunne dokumentere, at det i det hele taget er muligt at afgøre dette spørgsmål. Man kan f.eks. udregne en beståelsesscore ved brug af en anerkendt metode og herefter undersøge, om andelen af falsk positive og falsk negative bedømmelser er acceptabel eller ej. Den slags evidens vil hjælpe med at støtte antagelsen om, at *implikationerne* eller konsekvenserne af beslutninger, der er baseret på scorerne, er forsvarlige og tilsigtede. Principielt bør der være evidens fra alle fire kategorier, som samlet støtter validitetsargumenterne for brugen af bedømmelsesredskabet i tilstrækkelig grad.

DISKUSSION

Manglende ressourcer bliver ofte brugt som argument for at parkere spørgsmål om validitet, og manglende evidens for validitet vil ofte kunne forklares, men aldrig forsvares, med manglende ressourcer. Hvis man ikke har tilstrækkelige ressourcer, er det måske mest reelt, at man helt afstår fra at foretage kompetencevurderinger. De testansvarlige risikerer nemlig derved ikke blot at spilde de i forvejen knappe ressourcer, men også at gøre individer uret og tage dårlige beslutninger for organisationen og patienterne. Som regel foretages vurderinger i den kliniske hverdag af den person, der superviserer lægen, og vurderingen behøver ikke at være mere ressourcekrævende end den feedback, man ellers ville give til en yngre kollega. En særlig problemstilling knytter sig til kompetencevurdering af team. Optimal patientbehandling kræver oftest et godt teamsamarbejde blandt sundhedspersonalet, hvilket inkluderer mange svært målbare parametre som en god og klar kommunikation. Der er publiceret adskillige metoder beregnet til kompetencevurdering af team, men validitetsbeviserne for dem er fortsat mangelfulde [23].

KONKLUSION

Vi har i denne artikel beskrevet kompetencevurdering i den lægelige videreuddannelse, gennemgået validitetskrav til kompetencevurdering af læger samt givet et eksempel på anvendelse af validitetsevidens i en kompetencevurderingssituation. Vi har gjort det klart, at udvikling og undersøgelse af et vurderingsværktøj er en kontinuerlig videnskabelig proces, hvor der ikke eksisterer en endelig sand vurdering af kompetencer, og at en test ikke kan være universelt valid. Validitet handler ikke kun om, at man med en test måler det, man ønsker at måle, men også om fortolkning af testresultater og de beslutninger, man tager på baggrund af fortolkningerne. Når man bruger et vurderingsværktøj, er det vigtigt, at man først gør klart, hvad formålet med brugen af værktøjet er, hvilke antagelser man har, og hvilken beslutning man vil skulle træffe. På dette grundlag kan der tages stilling til, om vurderingsværktøjet giver en retfærdig og tilstrækkelig vurdering af kompetencer.

SUMMARY

Ebbe Thinggaard, Ann Sofia Thomsen, Lotte O'Neill & Lars Konge:
Measuring doctors competencies
Ugeskr Læger 2018;180:V01180059

Competency-based medical education relies on the use of assessment tools, which can describe and/or measure medical competencies and are supported by validity evidence. The exploration of validity evidence ensures, that the assessment tool not only measures what is intended, but also that the interpretation and decisions made are fair and just. In this review, we have described a contemporary

approach to validity used in Kane's framework. We have also used an example to illuminate, how the exploration of validity evidence can be done in a scientific and systematic manner.

KORRESPONDANCE: *Ebbe Thinggaard*.
E-mail: ebbe.thinggaard@gmail.com
ANTAGET: 18. maj 2018
PUBLICERET PÅ UGESKRIFTET.DK: 6. august 2018
INTERESSEKONFLIKTER: ingen. Forfatternes ICMJE-formularer er tilgængelige sammen med artiklen på Ugeskriftet.dk

LITTERATUR

1. Kane MT, Brennan RL. Educational measurement. 4th ed. Westport Am Counc Educ, 2006:17-64.
2. Brennan RL. Generalizability theory: statistics for social science and public policy. N Y Springer-Verl, 2013:30.
3. Kompetencevurderingsmetoder – en oversigt. Sundhedsstyrelsen, 2013.
4. Eraut M, Du Boulay B. Developing the attributes of medical professional judgement and competence: a review of the literature. Cogn Sci Res 2000 http://users.sussex.ac.uk/~bend/doh/reporthtml.html (15. dec 2017).
5. Wass V, van der Vleuten C, Shatzer et al. Assessment of clinical competence. Lancet 2001;357:945-9.
6. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul 2016;1:31.
7. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006;119:166.e7-16.
8. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ 2004;38:327-33.
9. Standards for educational and psychological testing. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing (US), 2014.
10. van der Vleuten C, Verhoeven B. In-training assessment developments in postgraduate education in Europe: in-training assessment in Europe. ANZ J Surg 2013;83:454-9.
11. Cook DA, Brydges R, Zendejas B et al. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acad Med 2013;88:872-83.
12. Korndorffer JR, Kasten SJ, Downing SM. A call for the utilization of consensus standards in the surgical education literature. Am J Surg 2010;199:99-104.
13. Cook DA, Kuper A, Hatala R et al. When assessment data are words: validity evidence for qualitative educational assessments. Acad Med 2016;91:1359-69.
14. Flexner A. Medical education in The United States and Canada. 1910. http://archive.carnegiefoundation.org/pdfs/elibrary/Carnegie_Flexner_Report.pdf (2. nov 2017).
15. Frank JR, Snell LS, Cate OT et al. Competency-based medical education: theory to practice. Med Teach 2010;32:638-45.
16. De syv lægeroller. Sundhedsstyrelsen, 2013.
17. Andreassen P, Pedersen K, Jensen RD et al. Optagelsessamtaler på medicin-studiet ved Aarhus Universitet: en pilottest af et multiple mini interview (MMI). Center for Sundhedsvidenskabelige Uddannelser, Aarhus Universitet, 2016.
18. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas 2013;50:1-73.
19. Cook DA, Brydges R, Ginsburg S et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ 2015;49:560-75.
20. Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. Adv Health Sci Educ 2015;20:829-34.
21. Cook DA, Zendejas B, Hamstra SJ et al. What counts as validity evidence? Adv Health Sci Educ 2014;19:233-50.
22. Birkmeyer JD, Finks JF, O'Reilly A et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434-42.
23. Rehim SA, DeMoor S, Olmsted R et al. Tools for assessment of communication skills of hospital action teams: a systematic review. J Surg Educ 2017;74:341-51.