Tidsskriftet
DEN NORSKE LEGEFORENING

Validitet

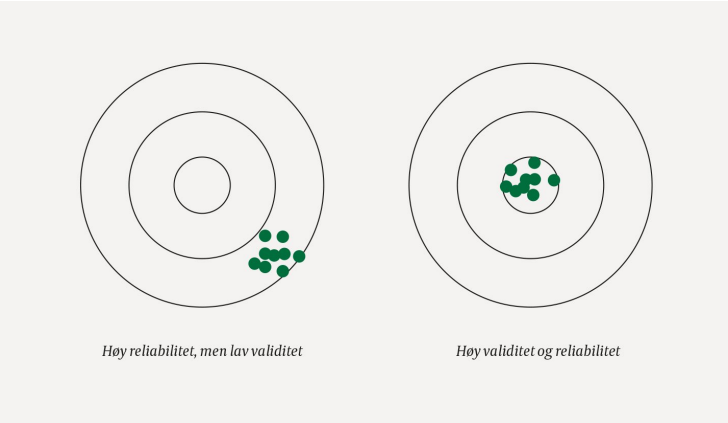
MEDISIN OG TALL

ARE HUGO PRIPP

E-post: apripp@ous-hf.no
Are Hugo Pripp er forsker og biostatistiker ved Oslo senter for biostatistikk og epidemiologi, Forskningsstøtteavdelingen, Oslo universitetssykehus og professor II ved Fakultet for helsevitenskap, OsloMet – storbyuniversitetet.
Forfatter har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Statistikere elsker variasjon. Analyser av variasjon gir oss p-verdier, standardavvik og konfidensintervaller. Men hvilken nytte har dette hvis målingen ikke er valid? Nytteverdien forutsetter validitet, og uten validitet vet vi ikke om målingen måler det vi ønsker å måle.

Hvis vi gjentar en måling ved like betingelser, er det betryggende å få like resultater. Målingen viser høy repeterbarhet eller reproduserbarhet, noe vi uttrykker som reliabilitet (1). God reliabilitet er likevel til liten hjelp hvis målingen er systematisk feil eller måler noe annet enn det vi ønsker. En vekt kan gi identiske resultater ved gjentatte målinger, men hvis den alltid viser for lav verdi, er det likevel en dårlig måling. Tilsvarende er et spørreskjema med reproduserbare resultater lite nyttig hvis spørsmålene ikke er relevante for det vi ønsker å måle (fig 1).



Figur 1 *Vurdering av validitet. Høy reliabilitet er ikke nok. Uten høy validitet vet vi ikke om målingen måler det vi ønsker å måle*

Validitet av måleinstrumenter

Vi har statistiske prosedyrer for å teste validiteten av tekniske måleinstrumenter. Målinger som høyde, vekt, glukose i blod og blodtrykk kan vi sammenligne og teste opp mot standarder, referanseverdier og referansemetoder. Et systematisk avvik fra standarden angir mangel på validitet. Avvik og variasjon for tekniske måleinstrumenter skal rutinemessig undersøkes med fastlagte prosedyrer og statistiske metoder.

Validitet av kliniske fenomener

Mange kliniske fenomener kan ikke måles med en enkel laboratorietest. Smerte, kvalme, depresjon eller utmattelse er eksempler på slike fenomener. Foruten klinisk skjønn, kunnskap og erfaring, kan vi bruke standardiserte metoder som strukturerte intervjuer og spørreskjemaer for å tallfeste slike fenomener.

En systematisk kartlegging av validiteten av et spørreskjema er viktig (2). Dette gjelder spesielt hvis vi oversetter et spørreskjema fra et annet språk, bruker et eksisterende spørreskjema på en annen pasientgruppe eller lager et helt nytt et. Hvert spørsmål skal måle et spesifikt fenomen. For spørreskjemaer kaller vi ofte slike fenomener for et konstrukt. Ideelt sett bør flere spørsmål uttrykke nyanser av et spesifikt konstrukt. Er vi interessert i en graderingsskala for utmattelse, bør det være flere spørsmål relatert til ulike nyanser av utmattelse. Vi bruker resultatene fra disse spørsmålene til å estimere en graderingsskala for utmattelse.

Det at spørsmålene virker rimelige og fornuftige, er en god begynnelse, men sjelden et tilstrekkelig bevis for validitet. Tre sentrale vurderinger av validiteten av spørreskjemaer er innholds-, konstrukt- og kriterievaliditeten. Innholdsvaliditeten (content validity) uttrykker i hvilken grad utvalget av spørsmål dekker alle dimensjoner av det fenomenet vi ønsker å måle. Den kan til dels vurderes uten formelle statistiske analyser (face validity). Konstruktvaliditeten (construct validity) uttrykker om spørreskjemaet måler det det er ment å måle. En relevant undersøkelse av konstruktvaliditeten er sammenhengen med assosierte fenomener, for eksempel om depresjonsspørsmålene viser assosiasjon med tretthet og hodepine. Kriterievaliditeten (criterion validity) uttrykker hvor godt målingen korrelerer med eller predikerer en annen valid og observerbar variabel (ofte kalt en kriterievariabel), for eksempel hvorvidt resultater fra et spørreskjema om smerte er relatert til forventet smerte ved ulike kliniske tilstander (3).

Indre og ytre validitet

Indre validitet uttrykker at resultatene er korrekte og gyldige for det studerte utvalget. Randomiserte studier har ofte høy indre validitet på grunn av randomiseringen og definerte inklusjons- og eksklusjonskriterier. Derimot er det usikkert om resultatene er gyldige for andre pasientgrupper. Ytre validitet angir i hvilken grad resultatene er gyldige under andre betingelser og for andre utvalg – altså generaliserbarheten. Valg av studiedesign kan bli en avveining ut ifra ønsket om indre eller ytre validitet (4).

Høy reliabilitet er ikke nok

Høy reliabilitet er bra, men målinger med lav reliabilitet kan likevel gi verdifull informasjon. Vi trenger bare et større utvalg for å få statistisk sikre estimater. Lav validitet er mer kritisk. Hvis målingen er systematisk feil eller måler noe annet enn det vi tror den måler, gjør større utvalg situasjonen verre. Vi blir statistisk sikrere på et ikke gyldig resultat.

LITTERATUR:

1. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008; 31: 466 - 75. [PubMed][CrossRef]

2. Bolarinwa OA. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. Niger Postgrad Med J 2015; 22: 195 - 201. [PubMed][CrossRef]

3. Fletcher RH, Fletcher SW. Performance of measurements. I: Clinical Epidemiology: The Essentials. 4. utg. Baltimore, MD: Lippincott Williams & Wilkins, 2005: 19–20.

4. Godwin M, Ruhland L, Casson I et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003; 3: 28. [PubMed][CrossRef]

Publisert: 3. september 2018. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.18.0398