Tidsskriftet
DEN NORSKE LEGEFORENING

Statistisk styrke – før, men ikke etter!

MEDISIN OG TALL

STIAN LYDERSEN

E-post: stian.lydersen@ntnu.no
Stian Lydersen er dr.ing. og professor i medisinsk statistikk ved Regionalt kunnskapssenter for barn og unge – psykisk helse og barnevern (RKBU Midt-Norge) ved Institutt for psykisk helse, NTNU.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

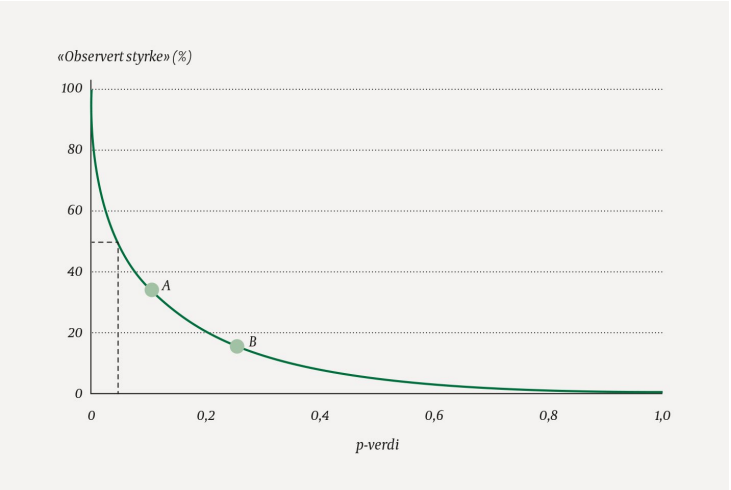
Enkelte tidsskrift og enkelte fagfeller etterlyser beregning av statistisk styrke med den observerte effektstørrelsen etter at en studie er gjennomført. Dette er fundamentalt feil.

Statistisk styrke er sannsynligheten for å forkaste nullhypotesen i en fremtidig studie. Etter at studien er gjennomført, er denne sannsynligheten enten 100 % (hvis nullhypotesen ble forkastet) eller 0 % (hvis nullhypotesen ikke ble forkastet).

Før man setter i gang en studie, bør man som regel gjennomføre en styrkeberegning eller antallsberegning. Man anslår på forhånd hvor stor en effekt enten kan forventes å være, eller hvor stor den bør være for at den skal ha klinisk betydning. Så beregner man sannsynligheten for at resultatet av en studie med et gitt antall deltakere blir statistisk signifikant (1). Denne sannsynligheten kalles statistisk styrke. Alternativt beregner man det antallet deltakere som trengs for å oppnå en bestemt statistisk styrke, for eksempel 90 %, for denne effektstørrelsen. Beregningene gjøres for et gitt signifikansnivå, som oftest lik 0,05.

«Observert styrke» og p-verdi

Etter at studien er gjennomført, rapporteres som regel et estimat og et 95 % konfidensintervall for effekten samt en p-verdi. Noen ganger opplever man imidlertid at et tidsskrift eller en fagfelle ber om en «observert styrke» i tillegg (2). Det vil si at man skal beregne statistisk styrke basert på den estimerte effekten som om det skulle dreie seg om en fremtidig studie, og rapportere dette som om det gir tilleggsinformasjon om den gjennomførte studien. Ikke bare er dette fundamentalt feil, det gir heller ingen tilleggsinformasjon til den rapporterte p-verdien: For enhver statistisk hypotesetest finnes det en entydig sammenheng mellom p-verdien og «observert styrke». For en ensidig test med normalfordelt utfallsvariabel og kjent varians er denne sammenhengen spesielt enkel (2). Dette er illustrert i figur 1. Der er «observert styrke» over 50 % hvis p-verdien er under 0,05, og under 50 % hvis p-verdien er større enn 0,05.



Figur 1 *«Observert styrke» for en ensidig test med normalfordelt utfallsvariabel og kjent varians ved signifikansnivå 0,05. Ved p-verdi 0,05 blir «observert styrke» 50 % (2). Studie A har høyere «observert styrke» enn studie B, men det betyr ikke at studie A gir sterkest evidens i favør av nullhypotesen. Tvert imot har A lavere p-verdi, altså sterkest evidens mot nullhypotesen.*

«Observert styrke» gir altså ingen tilleggsinformasjon utover p-verdien. Det kan derimot være direkte misvisende, noe mange synes ikke å være klar over. La oss tenke oss to studier, A og B, der nullhypotesen ikke ble forkastet og p-verdien er henholdsvis 0,10 og 0,25 (figur 1). Studie A har høyest «observert styrke» av de to. Noen ville tolke dette slik at studie A gir sterkest evidens i favør av nullhypotesen, som ikke ble forkastet, men dette er en feilslutning. Studie A har lavest p-verdi av de to studiene og slik sett sterkest evidens *mot* nullhypotesen. Denne typen feilslutning kalles «the power approach paradox» (2).

«Observert styrke» og konfidensintervall

Andre former for retrospektiv styrkeberegning har også vært foreslått, blant annet slik: «Anta at en studie ikke medførte at nullhypotesen ble forkastet. Med den observerte variabiliteten i studien, hvor stor måtte en hypotetisk effektstørrelse ha vært for å gi en bestemt statistisk styrke, for eksempel 90 %?» Dette er imidlertid også logisk feilaktig og kan føre til en annen versjon av «the power approach paradox», som nærmere beskrevet i (2). Et 95 % konfidensintervall, derimot, angir de verdiene av effektstørrelsen som er forenlige med observerte data.

Rapportering av statistisk styrke

Man bør rapportere hvordan en eventuell styrke- eller antallsberegning ble gjort før en studie ble iverksatt. Dette anbefales blant annet i CONSORT-retningslinjene (Consolidated Standards Of Reporting Trials) for randomiserte studier (3), og bidrar til å dokumentere at studien var godt planlagt. Etter at studien er gjennomført, vil konfidensintervall og p-verdi være egnede mål på usikkerhet. Beregning av «observert styrke» etter at studien er gjennomført, er både overflødig og misvisende.

LITTERATUR:

1. Pripp AH. Antalls- og styrkeberegninger i medisinske studier. Tidsskr Nor Legeforen 2017; 137. doi: 10.4045/tidsskr.17.0414. [PubMed][CrossRef]

2. Hoenig JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. Am Stat 2001; 55: 19–24. [CrossRef]

3. The CONSORT 2010 Statement. 7a. Sample size. http://www.consort-statement.org/checklists/view/32—consort-2010/83-sample-size (30.10.2018).

Publisert: 28. januar 2019. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.18.0847