Tidsskriftet
DEN NORSKE LEGEFORENING

Unngå signifikanstester innen grupper i randomiserte studier

MEDISIN OG TALL

EVA SKOVLUND

E-post: eva.skovlund@ntnu.no
Eva Skovlund (f. 1959) er professor i medisinsk statistikk ved Institutt for samfunnsmedisin og sykepleie, Norges teknisk-naturvitenskapelige universitet, og seniorforsker ved Nasjonalt folkehelseinstitutt.
Forfatter har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

I kontrollerte studier randomiserer man pasienter mellom to (eller flere) behandlinger for å unngå systematiske skjevheter. En rimelig analysestrategi for å dokumentere effektforskjell er selvsagt en direkte sammenligning av de to gruppene. I stedet velger noen å legge vekt på endring over tid innen hver gruppe for seg.

Hensikten med å randomisere er å sikre en rettferdig sammenligning. Gruppene vil i gjennomsnitt være sammenlignbare med hensyn til alle relevante variabler, med unntak av hvilken behandling som er gitt. Hvis studien ellers er godt planlagt og gjennomført, kan en statistisk signifikant forskjell i utfall mellom de to gruppene tilskrives forskjellig effekt av intervensjonene.

Ikke sjelden ser man publikasjoner der det blir lagt vekt på endringer over tid innen hver gruppe i stedet for en direkte sammenligning av de to gruppene, på tross av at det dreier seg om en randomisert studie med to parallelle grupper (1, 2). Ofte skyldes nok en slik tilnærming at man ikke har funnet statistisk signifikant forskjell ved en direkte sammenligning av gruppene.

I én studie ble effekten på fysisk aktivitet med bruk av Fitbit skritteller sammenlignet med bruk av et vanlig pedometer (1). Det var ingen statistisk signifikant forskjell mellom de to gruppene etter 16 uker, og forfatterne fremhevet at det var en statistisk signifikant økning i gjennomsnittlig antall skritt per dag i Fitbit-gruppen ($p = 0{,}01$), men ikke i kontrollgruppen ($p = 0{,}17$). Hvorfor er dette en uhensiktsmessig strategi?

Når vi kjenner sann effekt

Tabell 1 viser simulerte data for 40 fiktive individer fordelt på to grupper. Alle observasjonene er trukket fra samme populasjon, og det er altså ingen forskjell på de to gruppene. De sanne utgangsverdiene hadde et gjennomsnitt på 5 800 skritt og et standardavvik på 1 900, og etter 16 uker var den sanne effekten et tillegg på 500 skritt i begge grupper, med et standardavvik på 1 000.

Tabell 1

Simulering av en randomisert studie med i alt 40 fiktive individer med data (antall skritt) trukket fra en fordeling med en sann endring 500 skritt etter 16 uker, men ingen forskjell mellom eksperimentell behandling og standardbehandling

	Intervensjonsgruppe	Kontrollgruppe
	122	-635
	-237	906
	1 526	-1 716
	2 202	2 554
	325	1 472
	712	-1 263
	1 382	601
	419	812
	204	310
	1 540	222
	-347	222
	2 806	-318
	1 389	-730
	427	-2
	-249	-158
	1 234	-985
	-258	1 065
	1 490	-112
	1 758	1 523
	700	2 045
Gjennomsnitt	857	290
Standardavvik (SD)	895	1 106

Det simulerte eksemplet ga i gjennomsnitt en økning på 857 skritt i den ene gruppen og på 290 i den andre. Denne observerte forskjellen vet vi skyldes tilfeldighet. En enkel analyse av endring ved hjelp av en toutvalgs t-test ga estimert effektforskjell 567 skritt med 95 % konfidensintervall (KI) -78–1 211, p = 0,083. Forskjellen er ikke statistisk signifikant på 5 %-nivå. Det er ikke noen overraskelse, siden vi vet at det ikke er noen forskjell på de to gruppene.

Hvis vi derimot analyserer endring fra utgangsverdi til målingen ved 16 uker i hver gruppe separat ved hjelp av en paret t-test, er det ingen statistisk signifikant endring i kontrollgruppen (290 skritt, 95 % KI -227–808, p = 0,25), mens det viser seg å være en statistisk signifikant økning i intervensjonsgruppen (857 skritt, 95 % KI 438–1 276, p < 0,001). Dette gir ikke grunnlag for å konkludere med at effekten er forskjellig. Spørsmålet er ikke om det er en endring fra utgangsverdien, men om endringen er større i den ene gruppen enn i den andre.

Simulering

Selv om eksemplet over er en god illustrasjon, er ett regneeksempel ikke tilstrekkelig. For å estimere sannsynligheten for å finne effekt over tid i én gruppe, men ikke i den andre, selv om nullhypotesen om ingen forskjell er sann, har jeg gjort 10 000 simuleringer av den samme modellen som over.

Resultatet viste at en toutvalgs t-test ledet til forkasting av nullhypotesen i 499 av de 10 000 testene, altså svært nær 5 %, nettopp som man forventer med et signifikansnivå på 5 %. Men hvordan ville det gå hvis man i stedet gjorde to separate tester av endring i antall skritt innen hver av de to gruppene? Da ville faktisk 50 % av de 10 000 gjentakelsene resultere i at det ble statistisk signifikant forskjell (p < 0,05) i den ene av de to gruppene, men ikke den

andre, selv om sannheten er at effekten er lik i de to gruppene. Med andre ord er det en svært høy risiko for å finne forskjell selv om nullhypotesen er sann (falskt positivt funn) hvis man benytter en slik strategi.

Risikoen for et falskt positivt funn vil avhenge av hvor stor endringen innen hver gruppe er samt variabiliteten til denne. Hvis den sanne effekten i begge gruppene hadde vært en økning på for eksempel 250 skritt, med uendret standardavvik, ville 30 % av de separate testene innen hver gruppe resultert i et statistisk signifikant resultat i den ene gruppen, men ikke i den andre. Dette er fortsatt en risiko som langt overstiger 5 %. Dersom vi hadde sammenlignet *mellom* gruppene, ville en toutvalgstest uansett gitt 5 %, slik den skal.

LITTERATUR:

1. Cadmus-Bertram LA, Marcus BH, Patterson RE et al. Randomized trial of a Fitbit-based physical activity intervention for women. Am J Prev Med 2015; 49: 414 - 8. [PubMed][CrossRef]

2. Eberhardson M, Karlén P, Linton L et al. Randomised, double-blind, placebo-controlled trial of CCR9-targeted leukapheresis treatment of ulcerative colitis patients. J Crohns Colitis 2017; 11: 534 - 42. [PubMed]

Publisert: 19. februar 2018. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.18.0094
Mottatt 25.1.2018, godkjent 29.1.2018.
© Tidsskrift for Den norske legeforening 2020. Lastet ned fra tidsskriftet.no