Riittääkö otoskoko?

Harry Scheinin

Riittämätön otoskoko on edelleenkin valitettavan yleinen virhe kliinisissä tutkimuksissa. Suuria, jopa *kliinisesti merkittäviä* eroja ei saada *tilastollisesti merkitseviksi*, jolloin vaarana on tehdä ns. tyypin II virhe. Tällöin esim. kahta hoitoa pidetään virheellisesti samanarvoisina eli ei havaita todellista eroa. Suuremmalla potilasmäärällä samansuuruinen ero ryhmien välillä olisi johtanut tilastollisesti merkitsevän eron löytymiseen ja siten aivan päinvastaiseen johtopäätökseen. Tyypin II virhettä eli väärän negatiivisen tuloksen riskiä kutsutaan myös β-virheeksi. Erotus 1-β kuvaa käytetyn testin voimaa (engl. power) eli todennäköisyyttä löytää todellinen ero. Väärän positiivisen tuloksen (todellisuudessa ei eroa) riskiä kutsutaan vastaavasti tyypin I eli α-virheeksi. Tämän virheen hyväksyttävyysrajana pidetään yleisesti 5 %:n tasoa. Usein käytetään ilmaisua $p < 0.05$, mikä siis tarkoittaa alle 5 %:n mahdollisuutta, että ero johtuu sattumasta. α-virheen hallintaan on perinteisesti kiinnitetty huomattavasti enemmän huomiota kuin β-virheen.

Kuvassa 1 on esitetty totuuden ja tuloksen välisen yhteyden neljä mahdollisuutta. Sarakkeiden sisältämien todennäköisyyksien (todellisuudessa ei eroa tai ero on) summa on aina yksi. Tutkimuksen voima ilmoitetaan usein prosentteina. Mitä suurempi voima tutkimuksella on, sitä pienempi todennäköisyys on tehdä β-virhe. Esim. jos β on 0.1, voima on 1-0.1=0.9 eli 90 %. 80-90 %:n voimaa pidetään yleisesti hyväksyttävänä.

Tutkimuksen voimaa voi ja tulee arvioida jo ennakolta. Käytännössä tämä tapahtuu tekemällä ns. *voima-analyysi* tutkimuksen suunnitteluvaiheessa. GCP-ohjeistot, viranomaismääräykset ja eettiset toimikunnat edellyttävätkin otoskoon tarkkaa perustelemista tutkimussuunnitelmassa. Tutkimusta, joka ei kykene vastaamaan asetettuun kysymykseen liian pienen otoskoon takia, on pidettävänä jopa epäeettisenä; potilaita altistetaan turhaan kokeelliselle hoidolle ja lisäksi rajallisia tutkimusresursseja mme tuhlataan. Myös liian suurta tutkimusta voidaan pitää epäeettisenä, koska siinä satunnaistetaan tarpeettoman moni potilas saamaan potentiaalisesti huonompaa hoitoa. Voima-analyysi voidaan tehdä yksinkertaisten kaavojen ja /tai valmiiden taulukoiden avulla. Jatkuvalle datalle (esim. sydämen syke) tai luokittelevalle datalle (kyllä tai ei, vaste ilmoitetaan prosentteina) on olemassa ovat kaavansa (yleisesti puhutaan ”kvantitatiivisesta” ja ”kvalitatiivisesta” voima-analyysistä). Otoskoon laskemiseksi tulee olla jonkinlainen ennakkokäsitys hoidon tehosta ja jatkuvan muuttujan kyseessä ollen myös vasteen hajonnasta. Usein ainakin referenssihoidon teho tunnetaan esim. kirjallisuuden perusteella.

Jo tehdyn tutkimuksen otoskoon riittävyyden arvioinnissa ei kannata turvautua voima-analyysiin (vaikka refereet tätä usein pyytävätkin!), vaan sen tulisi perustua saavutettujen tulosten *luottamusvälitarkasteluun*. Luottamusväli (confidence interval) ilmoittaa haarukan, jonka sisällä todellinen keskiarvo (tai ryhmien välinen erotus) tietyllä todennäköisyydellä sijaitsee. Tutkimustuloksemmehan perustuvat otokseen koko populaatiosta ja sisältävät

		Totuus	
		A <>B	A=B
Tulos	A <>B	ok	α-virhe
	A=B	β-virhe	ok

Kuva 1.

siten sattuman aiheuttaman virheen. Luottamusvälin laajuuteen vaikuttaa kolme eri tekijää: otoskoko (mitä pienempi potilasmäärä sitä leveämpi luottamusväli), hajonta (mitä suurempi hajonta sitä leveämpi luottamusväli) sekä haluttu luottamuksen aste (mitä suurempi luottamus sitä leveämpi luottamusväli). Yleisesti käytetään 95 % luottamusvälejä, mikä palautuu normaaliin α-virhetasoon 5 % (eli $p < 0.05$). Mikäli kahden keskiarvon erotuksen luottamusväli sisältää nollan tai riskisuhteen luottamusväli luvun yksi, eivät ryhmät eroa tilastollisesti merkitsevästi toisistaan eli $p > 0.05$.

Miten luottamusvälejä sitten käytetään? *"Negatiivisessa tutkimuksessa"* (eli ryhmien välillä ei ole tilastollisesti merkitsevää eroa) katsomme ryhmien välisen erotuksen luottamusvälin ylärajaa (oikeastaan eniten nollasta poikkeavan pään itseisarvoa), siis todettua ryhmien välistä eroa suurempaa arvoa. Vain jos tälläkään erolla ei katsota olevan kliinistä merkitystä, voidaan hoitoja pitää samanarvoisina ja otoskokoa riittävänä negatiivisen johtopäätöksen tekemiseksi. Pelkkä tilastollisesti merkitsevän eron puuttuminen esim. kahden hoidon vertailututkimuksessa ei siis vielä oikeuta pitämään hoitoja samanarvoisina. Vastaavasti *"positiivisessa tutkimuksessa"* (ryhmien välillä on tilastollisesti merkitsevä ero) katsomme luottamusvälin alarajaa. Jos tällä (keskiarvojen erotusta pienemmällä) arvolla on vielä kliinistä merkitystä, voimme kiistattomasti pitää eroa myös kliinisesti merkittävänä. Luottamusvälitarkastelu auttaa siten, paitsi otoskoon riittävyyden arvioinnissa, niin myös tilastollisen merkitsevyyden ja kliinisen merkittävyyden suhteuttamisessa. Erityisen huono, mutta edelleen valitettavan yleinen tapa tieteellisissä julkaisuissa on ilmoittaa tutkimuksen tulokset vain p:n arvoina ilman mainintaa absoluuttisesta tai suhteellisesta erosta ryhmien välillä. Varsinkin pragmaattisissa tutkimuksissa tämä on ehdottomasti väärin, ja luottamusvälitarkastelu tulisikin tehdä aina perinteisten tilastollisten analyysien lisäksi.

Suunnitteluvaiheessa tehty voima-analyysi ei "vapauta" tekemästä tulosten luottamusvälilaskelmia. Mikäli tuloksissa on trendi referenssihoidon eduksi (esim. sattuman takia), voi luottamusvälitarkastelu osoittaa otoskoon riittämättömyyden vaikka tutkimuksen voima sinänsä olisikin ollut asianmukainen ja oikein laskettu. Vastaavasti hyvällä tuurilla (trendi uuden hoidon eduksi) alhaisen voiman omaavalla tutkimuksella voidaan onnistua osoittamaan, että uusi hoito ei ole ainakaan huonompi kuin referenssihoito.

Harry Scheinin, prof.
TYKS ja PET-keskus