VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Aleksi Reito, Lauri Raittio ja Olli Helminen

Tutkimustulokset eivät toistu – missä syy?

Tieteellisten havaintojen toistettavuus on keskeisin luotettavan tutkimuksen ominaisuus. Tutkimuslöydös on uskottava, jos se on toistettavissa aina uudelleen. Toistettavuusongelmalla tarkoitetaan, että aiemmin julkaistut tutkimuslöydökset eivät ole vahvistettavissa tai todennettavissa uusissa, riippumattomissa tutkimuksissa. Toistettavuusongelman tärkeimpiä syitä ovat tilastollisen merkitsevyyden tarkoitushakuinen etsiminen (”kalastelu”), virheelliset tilastolliset menetelmät ja toimintatavat, pieni tilastollinen voima sekä erilaiset harhakäsitykset. Tieteellisiä tutkimuksia lukevan on tärkeää ymmärtää, millaiset keinot toisaalta vahvistavat ja toisaalta heikentävät tutkimustulosten toistettavuutta.

Tilastotieteen pioneeri Douglas Altman kritisoi biolääketieteellisen tutkimuksen laatua pääkirjoituksessaan BMJ:ssä vuonna 1994 toteamalla, että ”tarvitsemme vähemmän tiedettä, parempaa tiedettä ja oikeista syistä tehtyä tiedettä” (1). Ongelmina hän mainitsi oikeiden menetelmien vääränlaisen käytön, tulosten virheellisen tulkinnan, tulosten valikoivan raportoinnin ja tulosten väärin ymmärtämisen. Altmanin pääkirjoituksen jälkeen hukkatutkimuksesta (research waste) on muodostunut yleinen puheenaihe, erityisesti kun keskustellaan lääketieteellisen tutkimuksen rahoittamisesta (2).

Hukkatutkimus nivoutuu läheisesti tieteen toistettavuusongelmaan, joka on viime vuosikymmenen aikana ollut yhä enemmän esillä (3–6). Toistettavuusongelmalla tarkoitetaan, että aiemmin julkaistut tutkimuslöydökset eivät ole vahvistettavissa tai todennettavissa uusissa, erillisissä tutkimuksissa. Se saattaa johtua siitä, että tulos oli sattuman seurausta, koeasetelma oli laadultaan heikko, käytetyt menetelmät olivat vääriä tai sinänsä oikeita menetelmiä käytettiin väärin. Asiaa on tutkittu muun muassa syöpäbiologiassa, lääketutkimuksissa, neurotieteissä ja psykologiassa, jossa puhutaan jopa toistettavuuskriisistä (6–9).

Hukkatutkimuksen ja toistettavuusongelman syyt ovat moninaiset, mutta ne nivoutuvat vahvasti toisiinsa (**TAULUKKO 1**) (17). Modernin tieteen toistettavuusongelman syyt rakentuvat tieteellisen tutkimuksen perusperiaatteille, joten näiden periaatteiden ymmärtäminen ja käsitteiden tunteminen on välttämätöntä jokaiselle, joka tuottaa, käsittelee tai käyttää tieteellistä tietoa.

Nollahypoteesin merkitsevyyden testaus tilastollisten menetelmien perustana

Tilastollisten menetelmien käyttö ja tilastollinen päättely on keskeistä modernissa biolääketieteellisessä tutkimuksessa. Nollahypoteesin (H_0) merkitsevyyden testaus (NHMT) muodostaa pohjan nykymuotoiselle tutkimustiedon käsittelylle eli frekventistiselle tilastotieteelle. NHMT:n keskiössä on p-arvon määrittäminen. Se kuvaa todennäköisyyttä havaita vähintään yhtä poikkeava aineisto, kun oletetaan, että nollahypoteesi ja muut taustaoletukset ovat voimassa. Toisin sanoen se vastaa kysymykseen, kuinka hyvin aineisto tukee asetettua nollahypoteesia (**KUVA 1**). P-arvo ei kuvaa, kuinka

Duodecim 2020;136:1155–63

TAULUKKO 1. Tieteen toistettavuutta uhkaavat tekijät ja niiden ehkäisyyn ehdotetut menetelmät. Mukailtu viitteestä (17).

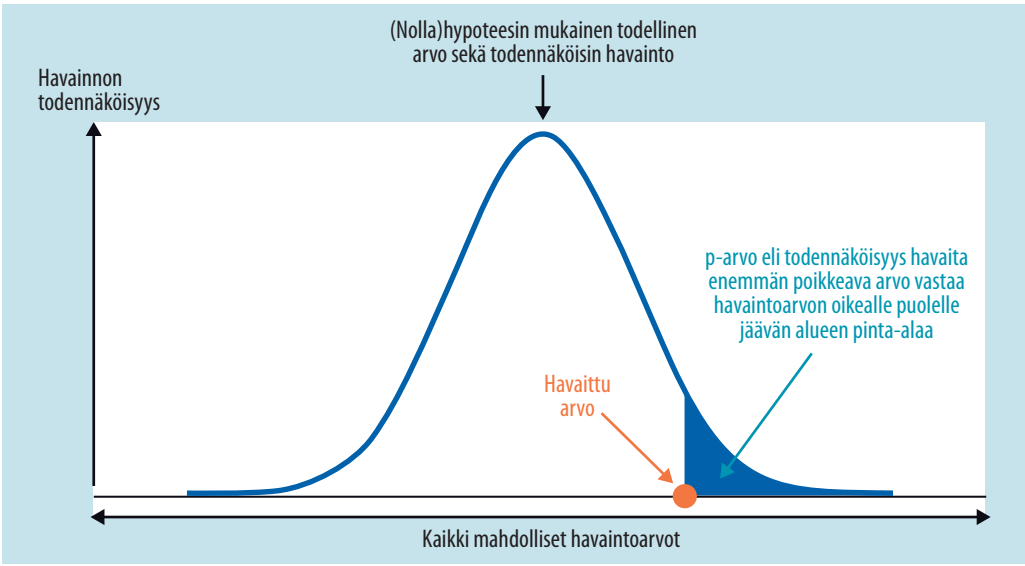
Toistettavuus-ongelman syy	Muita käsitteitä	Ilmentyminen kirjallisuudessa	Ehkäisevät toimenpiteet
Kalastelu (fishing) ja ruoppaus (dredging)	p-hakkerointi HARKing Tutkijan vapausasteet	Tilastollisesti merkitsevien tulosten suhteellinen ylimäärä Tyypin 1 virhe	Datan ja menetelmien avoimuus Ennakkorekisteröinti
Pieni tilastollinen voima	Negatiivisten tulosten väärä tulkinta	Laajat piste-estimaattien luottamusvälit Inflatoituneet vaikutuskoot Tyypin 2 virhe Voittajan kirous (winner's curse) Proteus-ilmiö	Ennakkorekisteröinti Tieteellinen yhteistyö[1]
Nollatulosten julkaisematta jättäminen	Pöytälaatikkoefekti Julkaisuharha Julkaisupaine	Tilastollisesti ei-merkitsevien tulosten suhteellinen pieni määrä Vääristyneet meta-analyysien tulokset	Datan ja menetelmien avoimuus
Suoranaiset virheet	–	–	Ennakkorekisteröinti Automaatio Tutkimusprotokollan julkaisu Julkaisusuositusten noudattaminen[2]
Vajaasti kuvattu metodologia	p-hakkerointi HARKing Tutkijan vapausasteet Väärät tilastolliset menetelmät	–	Ennakkorekisteröinti Tieteellinen yhteistyö Tutkimusprotokollan julkaisu Julkaisun jälkeinen vertaisarviointi Julkaisusuositusten noudattaminen
Heikko kokeellinen asetelma	Pieni signaali-kohinasuhde Tilastollinen kohina Jäännösharhan minimointi	–	Ennakkorekisteröinti Tieteellinen yhteistyö Tutkimusprotokollan julkaisu Julkaisun jälkeinen vertaisarviointi Julkaisusuositusten noudattaminen

[1]Tieteelliseen yhteistyöhön voidaan katsoa kuuluvan myös tiiviin työskentelyn tilastotieteen ammattilaisen kanssa.
[2] Julkaisusuosituksia ovat esimerkiksi CONSORT- ja STROBE-tarkistuslistat.
HARKing = hypoteesin asettaminen, kun tulokset ovat tiedossa (hypothesising after the results are known)

todennäköisesti nollahypoteesi on totta (10). Tutkimusta tekevien ja tulkitsevien on syytä ymmärtää käytetyimpien tilastomenetelmien filosofinen pohja, jotta niiden käyttöä voidaan kriittisesti arvioida. NHMT on kahden eri hypoteesitestausmenetelmän yhdistelmä (11).

Ronald Fisherin kehittämässä menetelmässä asetetaan nollahypoteesi, esimerkiksi kahden ryhmän keskiarvojen yhtäsuuruus. Tämän jälkeen kerätään aineisto. Aineiston perusteella lasketaan tilastollinen testisuure, jolle on määritetty vapausasteiden perusteella niin sanottu tarkka p-arvo eli tilastollinen merkitsevyys. Viitekehyksessä voidaan vain ja ainoastaan hylätä nollahypoteesi, eikä siihen kuulu vaihtoehtoisia hypoteeseja. Fisherin mukaan tarkka p-arvo antaa viitettä nollahypoteesin uskottavuudesta, ja sitä pitää arvioida yhdessä muun näytön, kuten vaikutuskoon (effect size) kanssa.

Fisherin mukaan 0,05 on p-arvo, jonka perusteella nollahypoteesin voisi yleensä hylätä. Mitään matemaattista perustelua tälle p-arvolle



KUVA 1. P-arvon merkitys ja tulkinta. Tutkija on tehnyt satunnaisotoksen väestöstä ja mitannut henkilöiden pituuden. Tutkija on asettanut (nolla)hypoteesin, jonka mukaan otoksen keskipituus on 175 cm keskihajonnalla 15 cm. Näin ollen otoksen keskiarvon voi katsoa noudattavan sinistä jakaumaa. Otoksesta mitattu pituuden keskiarvo on 182 cm (oranssi ympyrä). Tutkija tekee yksisuuntaisen testin, jolloin mittaustuloksen p-arvo vastaa tuloksen oikealla puolella olevan alueen pinta-alaa. Tässä tapauksessa p-arvo on 0,08 eli todennäköisyys havaita vielä enemmän poikkeava mittaustulos on 8 % kun oletetaan, että (nolla)hypoteesi on voimassa.

ei ollut. Kyseinen p-arvo voi olla mitä tahansa tutkijan arvomaailman mukaisesti. Fisherin viitekehyksen filosofinen ajatus ei ollut tuottaa tieteellisiä faktoja vaan enemmänkin korostaa ilmiöitä, jotka vaativat lisätutkimuksia (11).

Jerzy Neymanin ja Egon Pearsonin viitekehyksessä määritetään nollahypoteesin (H_0) lisäksi vaihtoehtoinen hypoteesi (H_1) sekä sallitut virhetasot alfa eli tyypin 1 virhe ja beeta eli tyypin 2 virhe. Yhden yksittäisen tutkimuksen tai hypoteesin hylkäämisen sijasta viitekehyksessä tarkastellaan lukuisia perättäisiä päätöksiä, ja pyrkimyksenä on minimoida niin väärät positiiviset (alfa) kuin väärät negatiiviset (beeta).

Neyman–Pearsonin viitekehyksessä ei määritetä tarkkaa p-arvoa, vaan hypoteesin hylkääminen perustuu yksittäisen kokeen perusteella saadun testisuureen ennalta asetettuun raja-arvoon (C), jonka perusteella joko pidetään voimassa nollahypoteesi tai se hylätään ja oletetaan vaihtoehtoisen hypoteesin voimassaolo. Luonnollisesti testisuureen raja-arvolle on määritettävissä myös sitä vastaava p-arvo. Neyman–Pearsonin viitekehyksessä vain p-arvolle määritetty raja-arvo merkitsee, eikä sitä tulkita näytönasteena kuten Fisherin viitekehyksessä.

Nykymuotoinen käytäntö eli NHMT on kehittynyt näistä kahdesta erilaisesta viitekehyksestä, mitä on kritisoitu jo 1940-luvulta alkaen, ja kritiikki on edelleen lisääntynyt viime vuosina (12–14). Hukkatutkimuksen perimmäiset syyt ovat juuri NHMT-pohjaisessa tilastotieteen menetelmien käytössä. NHMT:n viitekehyksessä asetetaan nollahypoteesi ja vaihtoehtoinen hypoteesi. Aineiston perusteella lasketaan testisuure ja sitä vastaava p-arvo. Sen alittaessa lähes universaalisesti hyväksytyn raja-arvon 0,05 hylätään nollahypoteesi ja oletetaan aineiston selittyvän vaihtoehtoisella hypoteesilla. Mikäli p-arvo ei alita sovittua raja-arvoa, nollahypoteesia ei voida hylätä ja usein oletetaan virheellisesti, että se on voimassa aineistossa. **TAULUKOSSA 2** esitetään NHMT:n keskeisimmät ongelmat (14).

Kalastelu, p-hakkerointi ja HARKing

Yksi suurimmista NHMT:n ongelmista liittyy usein kiinteään p-arvorajaan 0,05 (4,15,16). Tutkimusrahoituksesta kilpailtaessa ja julkai-

TAULUKKO 2. Nollahypoteesin merkitsevyyden testauksen ongelmia ja niiden seurauksia. Mukailtu viitteestä (14).

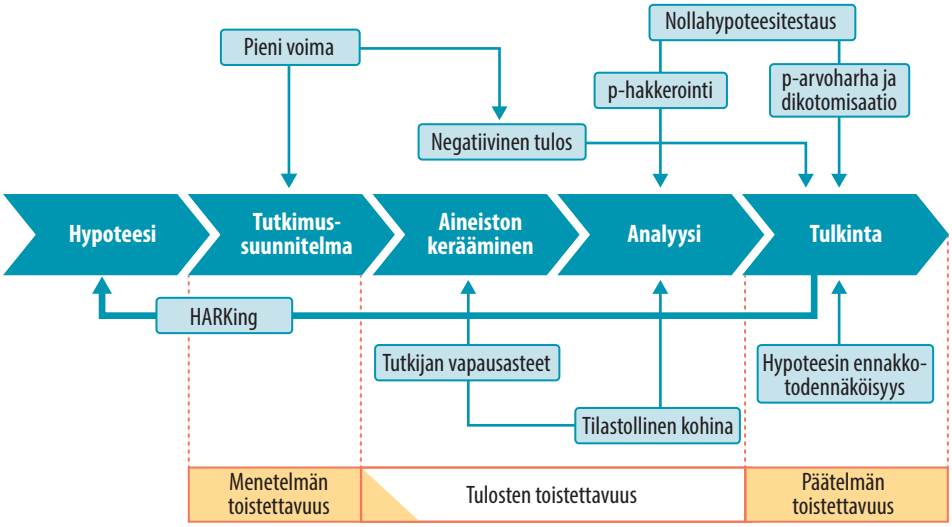
Ongelma	Kuvaus	Seuraus	Esimerkki
Viitekehyksen ainoa oikea päätelmä on H_0:n hylkääminen	H_0:aa ei voi koskaan hyväksyä. Suurempi p-arvo kuin ennalta sovittu raja-arvo ei tarkoita, että H_0 olisi totta.	Usein p-arvoa suurempi raja-arvo tulkitaan niin, että H_0 jää voimaan ja päätelmänä on, ettei ryhmien välillä ole eroa.	Kahden ryhmän välisen keskiarvon erotusta vastaava p-arvo on 0,07. Oikea tulkinta on, ettei ryhmien välille voitu osoittaa eroa. Se on ehkä olemassa, mutta otoskoko ei riittänyt sen osoittamiseen.
Viitekehyksessä ei oteta huomioon hypoteesin ennakkotodennäköisyyttä	Virhetasot α ja β kuvastavat päätöksentekoa pitkällä aikavälillä. Jos H_0 hylätään ja oletetaan H_1, ei α kuvasta lainkaan H_1:n mahdollista todenmukaisuutta.	Jotta todellinen väärän positiivisen tuloksen mahdollisuus voidaan laskea, pitää arvioida hypoteesin ennakkotodennäköisyyttä.	Tutkimuslöydöksen ennakkotodennäköisyys on 10 %. Jos tutkimuksen voima on 80 % ja tutkimustulos tilastollisesti merkitsevä, on väärän positiivisen tuloksen mahdollisuus jopa 36 %.
Viitekehys ei sovi suurten aineistojen analyysiin	Vain hyvin harvoin, jos koskaan, H_0 kuten ryhmien välisen keskiarvon erotus on todella nolla.	Mitä suuremmaksi otoskoko kasvaa, sitä todennäköisemmin mikä tahansa triviaali korrelaatio tai ryhmien välinen erotus voi saavuttaa tilastollisen merkitsevyyden.	Erittäin laajasta aineistosta voidaan osoittaa, että selittävän eli lähtömuuttujan ja selitettävän eli päätemuuttujan välinen korrelaatiokerroin on 0,05. Tämä tarkoittaa, että lähtömuuttuja selittää $0{,}05^2 = 0{,}25$ % päätemuuttujan vaihtelusta. Tällaisen löydöksen kliininen merkitsevyys on todennäköisesti hyvin vähäinen.
Nollahypoteesin hylkääminen tapahtuu ennen pitkää	Mitä useampi hypoteesin testaus tehdään, sitä todennäköisempää on, että jossain vaiheessa saadaan tilastollisesti merkitsevä tulos	Väärien positiivisten tulosten määrä lisääntyy.	Tutkimuksessa voidaan testata kymmenen eri muuttujan vaikutusta päätemuuttujaan. Vaikka muuttujilla ei olisi mitään selkeää yhteyttä päätemuuttujaan, niin 50 %:n todennäköisyydellä ainakin yksi muuttuja osoittautuu tilastollisesti merkitseväksi.
Viitekehys ei edistä systemaattista tiedon integrointia	Nollahypoteesitestauksen tärkein tulos on p-arvo, jota ei voi vertailla eri tutkimusten välillä.	Jokainen tutkimus on itsenäinen eikä aikaisempien tutkimusten löydöksiä voida järjestelmällisesti ottaa huomioon tilastollisessa testauksessa.	Jos kahdessa tutkimuksessa tulokset raportoidaan esimerkiksi käyttämällä mediaaneja ja vertailu tehdään käyttämällä järjestyssummia, pelkkä p-arvon raportointi ei mahdollista tulosten vertailua.

H_0 = nollahypoteesi; H_1 = vaihtoehtoinen hypoteesi; α = tyypin 1 virhe; β = tyypin 2 virhe

supaineen alla tilastolliseen merkitsevyyteen pyrkimisestä on tullut yleinen tapa. Esimerkiksi Nature-lehden kirjoitusohjeissa mainitaan, että tutkimukset, joiden tuloksissa on jotakin uutta tai yllättävää, lähetetään vertaisarvioon. Samaan aikaan tutkimuksia, joissa on saatu tilastollisesti ei-merkitseviä löydöksiä, kutsutaan negatiivissävytteisesti "nollatutkimuksiksi". Siten on ymmärrettävää, että tutkijoiden on tarkoituksenmukaista saada tilastollisesti merkitseviä löydöksiä.

Tietoista tai tiedostamatonta tilastollisen merkitsevyyden metsästämistä kutsutaan termillä p-hakkerointi tai aineiston kalastelu (3,17). Valikoimalla aineistoa ja tutkimusjoukkoa sekä vaihtelemalla tilastollista testiä tai muokkaamalla aineistoon ajettavaa tilastollista regressiomallia voidaan yrittää "kalastaa" aineistosta tilastollisesti merkitseviä löydöksiä ja pyrkiä tieteelliseen "uutuusarvoon".

Kalasteluun ja p-hakkerointiin liittyy läheisesti englanninkielinen termi HARKing (hypothesising after the results are known) eli hypoteesin asettaminen, kun tulokset ovat tiedossa (18). HARKing on p-hakkeroinnin ja kalastelun seuraava askel: kun aineistosta on tehty analyysi ja tulokset ovat tiedossa, laaditaan hypoteesi, jota tutkimuslöydöksellä pyritään selittämään. Ihanteellinen tieteellinen prosessi noudattaa hypoteettis-deduktiivista prosessia, jossa ensin asetetaan hypoteesi (**KUVA 2**). Tästä tehdään ennustuksia deduktion perusteella



KUVA 2. Hypoteettis-deduktiivinen prosessi ja siihen vaikuttavat toistettavuutta heikentävät tekijät. Mukailtu viitteestä (3). HARKing = hypoteesin asettaminen, kun tulokset ovat tiedossa (hypothesising after the results are known)

ja testataan, toimivatko ennustukset kerätyssä aineistossa. HARKing rikkoo tätä tieteellistä prosessia. Konkreettinen seuraus HARKingista on tyypin 1 virheiden lisääntyminen. Vaikka aineisto sisältäisi muuttujia, jotka todellisuudessa eivät korreloi tai liity toisiinsa ja joiden keskiarvot eivät eroa, satunnaisvaihtelu saa aikaan sen, että tietty määrä testeistä on kuitenkin tilastollisesti merkitseviä.

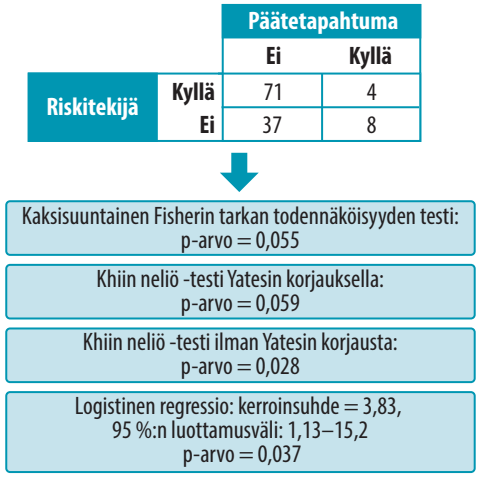
Kaiken sen joustavuuden, joka liittyy aineiston määrittelyyn, sen rajaamiseen ja tulosten käsittelyyn, kuvaamiseen on käytetty termejä ”tutkijan vapausasteet” sekä ”haarautuvien polkujen puutarha” (garden of forking paths) (**KUVA 3**) (19,20). Näillä ei suoraan tarkoiteta kalastelua vaan tilanteita, joissa tilastollinen menetelmä ja testi valitaan, kun tutkija on nähnyt aineiston. Erilaisella aineistolla tutkija olisi valinnut erilaiset menetelmät. Nämä kuvastavat hienosti sitä, kuinka tutkijalla on mahdollisuus päätyä erilaisiin tuloksiin vaikuttamalla aineiston määrittelyyn ja käytettyihin menetelmiin.

Päätelmät ja tutkimuslöydösten dikotomisaatio

Päätelmätoistettavuus (inferential reproducubility) tarkoittaa, että kahdessa tutkimuksessa voi olla identtiset tulokset, mutta niiden tulkinta voi olla erilainen (21). Varsinaisten numeeristen tutkimustulosten lisäksi tutkimukseen kuuluu aina niiden tulkinta (15). Dikotomisaatio tarkoittaa tutkimuslöydösten jaottelua tiukasti niin sanottuihin negatiivisiin ja positiivisiin tuloksiin p-arvorajan 0,05 mukaisesti (**KUVA 3**) (13,15). Tämä tarkoittaa erityisesti tilannetta, jossa ei oteta lainkaan laskennallisesti huomioon esimerkiksi aiempaa tutkimustietoa tai tuloksen epätarkkuutta (15). Vaikka tutkimuksen tulos muuttuu tilastollisesti merkitsevästä (esimerkiksi 0,049) tilastollisesti ei-merkitseväksi (esimerkiksi 0,051), voi muutos itsessään olla täysin merkityksetön (16).

P-arvorajan alittava löydös julistetaan todelliseksi, ja tutkimuksen päätelmänä todetaan, että ryhmien välillä on ero tai hoidolla on vaikutus. Piste-estimaatin arvosta riippuu, onko ero kliinisesti merkitsevä. Kun p-arvoraja ylittyy, tulkinta on päinvastainen: ryhmien välillä ei ole eroa tai hoito ei vaikuta. Tieteelliset teoriat sekä biokemialliset ilmiöt ja selityssuhteet ovat harvoin näin karkeita, vaan aineiston suhde hypoteesiin on aina liukuva (13,22).

Fysiologia ja biokemialliset ilmiöt ovat monimutkaisia, ja kahden muuttujan välillä vallitsee monisyisten korrelaatioiden maailmassa jokin yhteys. Ääripäässä on myös näkemys, ettei mikään vaste tai korrelaatio koskaan voi



KUVA 3. Tutkijan vapausasteet. Tutkijalla on 120 potilaan aineisto, jonka avulla hän selvittää riskitekijän vaikutusta päätetapahtuman esiintyvyyteen. Kyseinen yksimuuttuja-analyysi voidaan tehdä ainakin neljällä eri tavalla, joista jokainen on periaatteessa oikein. Kaksi testeistä antaa nimellisesti tilastollisesti merkitsevän löydöksen ja kaksi taas ei. Analyysiin liittyy useita sudenkuoppia. Jos testit tehdään ilman suunnitelmaa sovitusta testistä, tutkija syyllistyy kalasteluun, jos hän valitsee nimenomaan testin, joka antaa tilastollisesti merkitsevän tuloksen. Toisaalta tuloksen luokittelemista dikotomisesti (”riskitekijä vaikuttaa” tai ”riskitekijä ei vaikuta”) ei suositella, koska tulos on erittäin raja-arvoinen kumpaankin suuntaan. Kerroinsuhteen (odds ratio, OR) luottamusvälien tulkinta on mielekkäintä. Kerroinsuhteen luottamusväli on hyvin leveä, mikä tarkoittaa, että aineisto on yhtenevä erittäin suuren arvojoukon kanssa. Tällöin kannattaa tulkita, miten esimerkiksi kliinisesti merkitsevät kerroinsuhteen arvot sijaitsevat piste-estimaatin luottamusvälillä. Kokonaisuutena rehellisin tulkinta on, että aineistosta voidaan päätellä melko vähän. Yatesin korjausta käytetään, kun jonkin solun lukumäärä lähenee nollaa. Sillä pyritään välttämään pienten lukujen aiheuttamaa virhettä määritettäessä testin p-arvoa khiin neliö -jakauman pohjalta. Yksiselitteistä ohjetta sen käytöstä ei ole.

olla täysin nolla (23). Onkin ehdotettu, että tyypin 1 ja 2 virheet korvattaisiin tyypin S ja M virheillä. Näillä tarkoitettaisiin sitä, onko raportoitu vaikutuksen estimaatti oikeansuuntainen (S, sign) ja oikeankokoinen (M, magnitude) (24).

Tutkimuslöydöksen ennakkotodennäköisyys

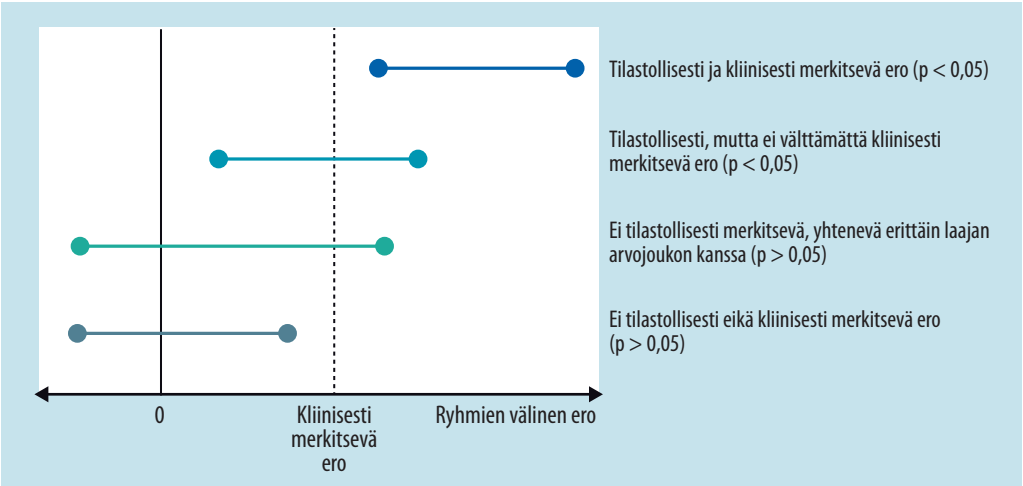
Yksi merkittävä syy tutkimuslöydösten huonoon toistettavuuteen on, ettei NHMT:ssä oteta huomioon mahdollisen tutkimuslöydöksen ennakkotodennäköisyyttä (13,14,25). Aivan kuten kliinisten testien tarkkuudessa ja herkkyydessä myös NHMT:ssä pitäisi huomioida tutkimuslöydöksen ennakkotodennäköisyys. Herkkyys ja tarkkuus ovat matemaattisesti sidottuja tutkittavan ilmiön esiintyvyyteen populaatiossa. Esimerkiksi CRP:n toimivuus keuhkokuumeen selvittelyssä riippuu siitä, koostuuko aineisto ylähengitystieoireisista vai muusta paikallisoireesta kärsivistä.

Tutkimuslöydöksen ennakkotodennäköisyys tarkoittaa käytännössä teoreettista arviota siitä, kuinka moni kaikista mahdollisista tutkimusideoista (hypoteesi) käytännössä on todellisia ja merkityksellisiä. Vaihtoehtoisesti voidaan ajatella, että jos tutkitaan hoidon A vaikutusta muuttujaan B ja yhdeksässä tutkimuksessa ei ole havaittu yhteyttä näiden välillä mutta yhdessä on, olisi tutkimushypoteesin ennakkotodennäköisyys seuraavassa tutkimuksessa 10 %. Vasta ennakkotodennäköisyyden huomioon ottamisen jälkeen on mahdollista laskea todellinen väärän positiivisen tuloksen mahdollisuus tutkimuslöydökselle (13,14,25).

Tilastollinen voima

Pieni tilastollinen voima on myös tärkeä syy tieteen toistettavuusongelmaan (4,8,26). Se tarkoittaa liian pientä otoskokoa suhteessa todelliseen vaikutuksen suuruuteen. Selkein pienen otoskoon seuraus on suuri tyypin 2 virheen eli väärän negatiivisen tuloksen mahdollisuus. Tämä tarkoittaa, että tutkittava ilmiö on olemassa, mutta otoskoko ei ollut riittävä sen toteamiseen eli nollahypoteesin hylkäämiseen.

Tarkasteltaessa otoksesta määritettyjä piste-estimaatteja, esimerkiksi ryhmien välistä keskiarvoa, pieni tilastollinen voima näkyy leveinä luottamusväleinä. Tämä on erityinen ongelma niin sanotuissa negatiivisissa tutkimuksissa. Pienen otoskoon seurauksena piste-estimaatille lasketut luottamusvälit ovat leveät ja saattavat sisältää esimerkiksi kuolleisuudelle määritetyt kliinisesti merkitsevät muutokset kahden tutkitun intervention välillä molempien interventioiden eduksi (27,28). Tällöin tutkimus ei juuri ohjaa tutkimushypoteesien oletettua ennak-



KUVA 4. Luottamusvälien tulkinta. Ylin luottamusväli sulkee pois 5 %:n virhetasolla niin erotuksen yhtäsuuruuden kuin kliinisesti merkitsevän eron, kun oletetaan, että nollahypoteesina on ryhmien keskiarvojen yhtäsuuruus. Toiseksi ylin luottamusväli sulkee pois yhtäsuuruuden, mutta ei kliinisesti merkitsevää eroa. Toiseksi alin luottamusväli on yhtenevä erittäin suuren arvojoukon kanssa eikä sulje pois 5 %:n virhetasolla ryhmien välistä yhtäsuuruutta tai kliinisesti merkitsevää eroa. Näin ollen luottamusvälin perusteella ei voida päätellä mitään. Alin luottamusväli ei sulje pois ryhmien yhtäsuuruutta mutta sulkee pois kliinisesti merkitsevän eron 5 %:n virhetasolla. Kahden alimman luottamusvälin osalta on väärin todeta, että "ryhmien välillä ei ole eroa". Luottamusvälin sekä p-arvon perusteella ei voida todeta muuta kuin että ryhmien välistä eroa ei voitu osoittaa.

kotodennäköisyyttä kummankaan hypoteesin suuntaan.

Toinen pieneen otoskokoon liittyvä ilmiö on nimeltään vaikutuksen inflaatio (29). Pienellä otoskoolla vain hyvin suuri vaikutus, kuten ryhmien välinen ero, on riittävän suuri ollakseen tilastollisesti merkitsevä. Raportoitu vaikutuskoko on tällöin inflatoitunut. Tähän ilmiöön liittyvät myös termit voittajan kirous (winner's curse) ja Proteus-ilmiö: ensimmäinen tutkimus, joka raportoi uuden löydöksen, on liian suuri vaikutuskooltaan, ja isommalla otoskoolla tehdyissä jatkotutkimuksissa todellinen vaikutus on paljon pienempi (8,30).

Negatiivinen tulos

Ainoa mahdollinen toimenpide, jonka tutkija voi tehdä NHMT:ssä, on nollahypoteesin hylkääminen. Se, että nollahypoteesia ei hylätä vaan se jätetään voimaan, ei tarkoita sitä, etteikö ryhmien välillä voisi olla eroa. Tämä on kuitenkin lähes universaalinen, väärä tulkinta. Kun p-arvo jää yli ennalta sovitun raja-arvon, ainoa oikea päätelmä on, että nollahypoteesia ei voida hylätä.

Oikea tapa tulkita negatiivista tutkimusta on katsoa, mitkä arvot sisältyvät piste-estimaatin luottamusväliin eli toisin sanoen mitkä arvot voidaan ainakin hylätä luottamusvälin mukaisesti suunnitellulla virhetasolla (**KUVA 4**) (27,28). Pieni tilastollinen voima suurentaa riskiä saada negatiivisia tuloksia. Niiden oikeaoppinen tulkinta on keskeinen osa päätelmätoistettavuutta (21).

Signaali-kohinasuhde ja tilastollinen kohina

Signaali-kohinasuhteella tarkoitetaan, kuinka tarkasti haluttua ilmiötä tai asiaa voidaan mitata olemassa olevilla menetelmillä (19). Tutkimuksessa, jossa arvioidaan leikkausvuotoa mittaamalla hemoglobiinipitoisuus ennen ja jälkeen leikkauksen, on suuri signaali-kohinasuhde, koska hemoglobiinimuutos kuvaa tarkasti leikkausvuodon määrää. Pieni signaali-kohinasuhde taas on usein ongelma mitattaessa hyvinkin käsitteellisiä asioita, kuten onnellisuutta tai mielialaa, tai kun mitataan erilaisten biomerkkiaineiden yhteyttä kliinisesti merkitseviin muuttujiin.

Ydinasiat

- Tieteessä keskustellaan lisääntyvästi toistettavuusongelmasta eli siitä, että aiemmin julkaistut tutkimuslöydökset eivät ole toistettavissa uusissa, erillisissä tutkimuksissa.
- Toistettavuusongelman syyt rakentuvat tieteellisen tutkimuksen perusperiaatteille, joten näiden periaatteiden ymmärtäminen ja käsitteiden tunteminen on erittäin tärkeää.
- Niin tiedostamaton kuin tahallinenkin epätieteellinen toiminta on sitä epätodennäköisempää, mitä enemmän aiheesta on tietoa ja ymmärrystä.
- Tutkijan ja kliinikon on tärkeää ymmärtää, millaiset keinot vahvistavat tai heikentävät tutkimustulosten toistettavuutta, jotta osaan tutkimuksista osataan suhtautua varauksella.

Tilastollinen kohina tarkoittaa aineistossa esiintyvää selittämätöntä muuttujien vaihtelua. Ongelmaksi se muodostuu silloin, kun kohinaa raportoidaan tilastollisesti merkitsevänä löydöksenä. Mitä enemmän aineistossa on mitattuja muuttujia, sitä todennäköisempää on löytää aineistosta kohinan seasta jokin tilastollisesti merkitsevä löydös. Ymmärrettävästi tällaisten löydösten toistettavuus on huono.

Lopuksi

Tieteellisen tutkimuksen tekeminen ja tulkinta ovat täynnä sudenkuoppia. Tutkija voi tahallaan tai tiedostamattaan syyllistyä sellaisten epätieteellisten menetelmien käyttöön, jotka heikentävät tieteellisen tutkimuksen toistettavuutta. Tiedostamaton epätieteellinen toiminta on sitä epätodennäköisempää, mitä enemmän aiheesta on tietoa ja ymmärrystä. Tieteellisiä tutkimuksia lukevan tutkijan tai kliinikon on tärkeää ymmärtää, millaiset keinot toisaalta vahvistavat ja toisaalta heikentävät tutkimusten toistettavuutta. ■

ALEKSI REITO, LT, dosentti
Twitter: @AleksiReito
LAURI RAITTIO, LK
Tampereen yliopisto, lääketieteen ja terveysteknologian tiedekunta
OLLI HELMINEN, LT, dosentti
Keski-Suomen keskussairaala, kirurgian klinikka
Oulun yliopistollinen sairaala, kirurgian klinikka

SIDONNAISUUDET
Aleksi Reito: Luentopalkkio/asiantuntijapalkkio (Orion)
Lauri Raittio: Ei sidonnaisuuksia
Olli Helminen: Ei sidonnaisuuksia

VASTUUTOIMITTAJA
Seppo Meri

SUMMARY
Reproducibility of scientific studies and problems associated with it: from understanding to better science
Reproducibility of scientific findings and results is in the core of modern science. Issue of irreproducibility is raised when previous finding cannot be replicated or reproduced in a new, separate experiment. Major issues resulting in problems with reproducibility are fishing, flawed statistical methods and procedures, low statistical power and misconceptions in general. Clinicians and researchers who read and assimilate scientific research papers must understand which factors are crucial for reproducibility of the studies and which factors should be a cause for concern.

KIRJALLISUUTTA

1. Altman DG. The scandal of poor medical research. BMJ 1994;308:283–4.
2. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374:86–9.
3. Munafò MR, Nosek BA, Bishop DVM, ym. A manifesto for reproducible science. Nat Hum Behav 2017;1:21.
4. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. Peer J, julkaistu verkossa 7.7.2017. DOI:10.7717/peerj.3544.
5. Baker M. 1,500 scientists lift the lid on reproducibility. Nature 2016;533:452–4.
6. Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature 2012; 483:531–3.
7. Errington TM, Iorns E, Gunn W, ym. An open investigation of the reproducibility of cancer biology research. Elife, julkaistu verkossa 10.12.2014. DOI: 10.7554/eLife.04333.
8. Button KS, Ioannidis JP, Mokrysz C, ym. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev 2013;14:365–76.
9. PSYCHOLOGY. Estimating the reproducibility of psychological science. Open Science Collaboration. Science 2015, julkaistu verkossa 28.9.2015. DOI: 10.1126/science.aac4716.
10. Greenland S, Senn SJ, Rothman KJ, ym. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337–50.
11. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front Psychol 2015;6:223.
12. Gigerenzer G. Mindless statistics. J Socio Econ 2004;33:587–606.
13. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond " p < 0.05." Am Stat 2019;73:1–19.
14. Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. Front Hum Neurosci 2017;11:390.
15. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999;130:995–1004.
16. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. Am Stat 2006;60:328–31.
17. Reproducibility and reliability of biomedical research: improving research practice. Symposium report. The Academy of Medical Sciences 2015. https://acmedsci.ac.uk/file-download/38189-56531416e2949.pdf
18. Kerr NL. HARKing: Hypothesizing after the results are known. Personal Soc Psychol Rev 1998;2:196–217.
19. Loken E, Gelman A. Measurement error and the replication crisis. Science 2017; 355:584–5.
20. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology. Psychol Sci 2011;22:1359–66.
21. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med 2016;8:341ps12.
22. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016;70:129–33.
23. Meehl PE. Why summaries of research on psychological theories are often uninterpretable. Psychol Rep 1990;66:195–244.
24. Gelman A, Carlin J. Beyond power calculations. Perspect Psychol Sci 2014;9:641–51.
25. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 2014; 1:140216.
26. Halsey LG, Curran-Everett D, Vowler SL, ym. The fickle P value generates irreproducible results. Nat Methods 2015; 12:179–85.
27. Anderson AA. Assessing statistical results: magnitude, precision, and model uncertainty. Am Stat 2019;73:118–21.
28. Gewandter JS, McDermott MP, Kitt RA, ym. Interpretation of CIs in clinical trials with non-significant results: systematic review and recommendations. BMJ Open, julkaistu verkossa 18.7.2017. DOI: 10.1136/bmjopen-2017-017288.
29. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008;19:640–8.
30. Ioannidis JPA, Trikalinos TA. Early extreme contradictory estimates may appear in published research: the proteus phenomenon in molecular genetics research and randomized trials. J Clin Epidemiol 2005;58:543–9.