Tilastoanalyysin avainsanat

TIETEELLINEN MENETELMÄ

Tutkimusartikkelin menetelmäosa saattaa sisältää maininnan tietokoneohjelmasta, jolla tilastolliset laskelmat ovat ajetut tai laitteesta, jolla laboratorioanalyysit ovat tehdyt. Tuollaisella teknisen välineen kuvauksella ei kuitenkaan voi korvata varsinaista tieteellistä menetelmää eli yleisiä sääntöjä, joilla harjoitetaan havaintoperäistä tutkimusta.

Perimmältään havaintoperäisen tieteen tavoite voidaan tiivistää joko entiteettien, olioiden, löytämiseen tai niitten välisten riippuvuuksien, kausaliteettien, todentamiseen. Luonnontieteen ideaali on rakentaa teoria muutaman olion varaan. Esimerkiksi Newtonin sanotaan päätyneen yhteen entiteettiin. Se on massapiste. Myös lääketieteessä tuollainen ideaali on joskus tavoitteena. Toisaalta esimerkiksi tautia voidaan kuvata geenin, potilasmerkityksen tai yhteisörasituksen perusteella tavallisen kliinisen kuvan sijasta tai ohella. Noita taudin dimensioita ei kuitenkaan voi korvata toisillaan, kuten jo WHO-määritelmäkin toteaa. Reduktioon pyrkivällä ajattelutavalla on laajoja seurauksia. Syntyy itseään ruokkiva kehä: luonnontieteellinen maailmankuva priorisoituu tai jopa dogmatisoituu lääketieteessä ja sen sovelluksena terveydenhuolto yksipuolistuu.

Entiteetti-ideaalista seuraa, tai on sen vastinpari, pyrkimys löytää välttämätön ja riittävä kausaalirelaatio. Tartuntataudit on perinteinen esimerkki, lääketieteen tutkimuksen esikuva. Kun perimmäinen syy, mikro-organismi, on löydetty, avautuu tie hoidolle ja ehkäisylle. Kun tuo syy määritellään myös taudiksi, syntyy kehäpäätelmä. Sen seurauksena saman kliinisen kuvan omaavat tilat määritellään eri taudiksi, jos mikro-organismit eivät ole samat. Ja muita tuon taudin syitä ei pidetä syinä tai ne luokitellaan vähempimerkityksellisiksi taipumuksiksi, alttiuksiksi, altistumisiksi, riskitekijöiksi, mitä eufemismia mikin lääketieteen ala suosiikaan.

Itse asiassa lääketiede olettaa – ja terveydenhuolto perustuu – useita entiteettejä ja monia riippuvuuksia niitten välille. Puhutaan monietiologisuudesta ja monipatogeenisuudesta. Taudilla voi olla monta syytä ja yksi syy voi aiheuttaa monta tautia. Tuo ajattelu johtaa tai on yhtäpitävä stokastisen kausaliteettikäsitteen kanssa: syy on tekijä, joka muuttaa seurauksen todennäköisyyttä. Näin määritellen syyt ovat samanarvoisia ja niitten prioriteetin määrää seurausten yleisyys, vakavuus, ehkäistävyys, tms, eikä dogmi esimerkiksi eri tieteenalojen arvojärjestyksestä.

Tieteelliseen tutkimukseen kuului siis sekä entiteettien keksiminen että kausaliteettien löytäminen. Muun muassa Rolf Nevanlinna ei pitänyt mahdollisena tieteellistä menetelmää, yleisiä sääntöjä, joilla tutkimusta harjoitetaan. Häntä lienee ohjannut ennenkaikkea olioitten reduktioon pyrkivä tiedekäsitys. Stokastiselle kausaalitutkimukselle on yleisiä sääntöjä, jotka tilastotiede tarjoaa. Lääketieteellisen kausaalitutkimuksen tiede puolestaan on epidemiologia, eli nuo kaksi tieteenalaa ovat syvällisesti toisiinsa liittyneitä ja muodostavat lääketieteessä tieteellisen menetelmän ytimen.

H_0

Tilastoanalyysi on mielekäs sovellettavaksi vain silloin kun tutkimuksella on spesifioitu tavoite. Tilastotieteessä esiintyy termi nollahypoteesi, H_0, jota tavallisesti pidettäneen tilastotieteen teorian apuvälineenä ja havaintoperäiseen tutkimukseen kuulumattomana koristeena. Se on kuitenkin tieteellisen menetelmän tapa kertoa täsmällisessä muodossa, mikä on kyseisen tutkimuksen tavoite.

Nollahypoteesi määrittää verrokkien ideaalisen valinnan tai kääntäen: verrokit paljastavat tutkimuksen todellisen tavoitteen. Verrokkien puute tai huono valinta ovat tunnetusti lääketieteen tutkimuksen perusongelma. Tapaukset (potilaat, altistuneet, käsitellyt, uudistetut normit)

eli syy-positiiviset havainnot ovat kulloinenkin käsillä oleva lähtökohta ja verrokeista eli syy-negatiivisista saatetaan huolestua vasta kun tapausaineisto on kerätty tai terveydenhuolto uudistettu. Verrokit ovat kuitenkin verrattomasti tapauksia tärkeämmät. Itse asiassa tutkimuksen johtopäätös riippuu verrokeista, eli (nolla)hypoteesista enemmän kuin tapauksista.

Täydellisestä verrokittomuudesta on lääketieteessä suureksi osaksi päästy eroon. Aikakauslehdet eivät enää julkaise esimerkiksi tapausselostuksia. Kokonaan tuota lähestymistapaa ei ole haudattu, kausaalitutkimus ilman verrokkeja on löytänyt suojan erityisesti laadullisen tutkimuksen sisältä.

Tieteellisen artikkelin tausta tai lähtökohta on esimerkiksi seuraava: Tutkimme hoitoa A ja esitämme (nolla)hypoteesin, että A:n vaikutusero aikaisempaan tilanteeseen verrattuna on olematon, hoidontulosero on nolla. Tieteessä ei uskota ennen kuin tuollainen tasapainotila on vääräksi osoitettu. Käytännössä on toisin, tutkimusta tuskin lainkaan lähdettäisiin tekemään, jos oman rokotteen, seulonnan, hoidon tai peräti terveydenhuollon uudistuksen mahdollisuuksiin ei luotettaisi. Ennakkoasenteet ja muu subjektivismi ovat niitä tekijöitä, joita tieteellinen metodi pyrkii eliminoimaan.

Vertailutilanteen määrittelylle, vertailuarvolle ei ole yleispäteviä tieteellisiä kriteereitä. Se on aihealueen asiantuntemuskysymys, ja sen valintaan voidaan antaa yleisiä ohjenuoria. Jos olemme keksineet uuteen tautiin sen ensimmäisen hoidon, esimerkiksi koronavirusrokotteen, vertailu tapahtuu rokotettujen ja rokottamattomien välillä. Kun tilastoanalyysia kehitettiin, tuo oli tavallisin tausta ja lähtökohta. Nykysovellutukset ovat toisenlaiset. Esimerkiksi syöpähoitoja on monta vaihtoehtoa, joten on perusteltua verrata keksintöämme parhaaseen käytössä olevaan.

Toki vieläkin tapaa vertailun nollatilanteeseen sujahtaneen taustahypoteesiksi. Onhan helpompi osoittaa oman keksintömme paremmuus verrattuna neitseelliseen tai keskiarvoiseen tilanteeseen kuin parhaaseen olemassa olevaan vaihtoehtoon. Vertailuarvo saatetaan valita tarkoitushakuisesti tai valintaan vaikuttavat subjektiiviset ennakkoasenteet.

MERKITSEVÄ

Tieteellisen artikkelin tavallisimpia virkkeitä on ”tulos oli (tai ei ollut) tilastollisesti merkitsevä”. Virke perustuu tilastollisen testin tulokseen. Tilastotieteen testi p-arvoineen kohdistuu tutkimuksesta tehtävään johtopäätökseen, tutkimustavoitteeseen. Merkitsevä p-arvo ($p<0.05$) tulkitaan siten että ero on todellinen, esimerkiksi keksintömme on muita hoitoja parempi. Nuo muut hoidot oli siis saatettu valita tarkoitushakuisesti. P-arvo sinänsä ei kerro mitä verrattiin ja mihin.

Tilastollisilla testeillä saatetaan verrata myös muita kuin tutkimustavoitteen mukaisia eroja. Sellaisella tilastoanalyysilla ei ole mieltä, vaan p-arvot ovat lähinnä tutkimuksen tieteelliseksi naamioitua koristelua.

Tilastollisesti merkitsevän p-arvon saatetaan tulkita suojanneen kaikilta virhepäätelmiltä eli tutkitun vaikutuksen (eron) olevan todellisen. Tilastollinen testi mittaa pelkästään sattumaa, ei esimerkiksi tutkimuksen yleisten puutteitten tai muitten tulokseen vaikuttavien eli sekoittavien tekijöitten vaikutusta. Havaintoaineistosta laskettu vaikutusero, estimaatti, sisältää sekä todellisen vaikutuksen että systemaattisen virheen eli harhan ja sattumavaikutuksen. Kaksi ensin mainittua ovat vakioita, mutta tuntemattomia. Kolmannella on tunnettu satunnaisjakauma, johon tilastollinen testi perustuu. Esimerkiksi, jos testataan vakioimattomia, karkeita lukuja, mahdollinen merkitsevä testitulos saattaa johtua siitä, että harha oli suuri mutta varsinainen efekti olematon. Tilastollisella vakioinnilla pyritään arvioimaan harhan suuruutta ja eliminoimaan se. On siis mielekästä soveltaa tilastollista testiä vain eroihin, jotka perustuvat vakioituihin vaikutuslukuihin.

Harhan eliminoiminen onkin yksi tilastoanalyysin tehtävistä. Yhden parhaista vakiointikeinoista tarjoavat regressiomenetelmät – pätevästi sovellettuina. Erityisesti datan käsittelyn helppous tietokoneitten avulla on tuonut uusia sudenkuoppia, harhaa eliminoitaessa saatetaan samalla ja vahingossa eliminoida myös itse vaikutus. Jos kaksi syytä tai syy ja harhatekijä ovat tiiviisti korreloituneita, regressiomenetelmän sovellutus saattaa lähes sattumanvaraisesti eliminoida yhden syyn eli leimata sen harhatekijäksi. Yleisellä tasolla tilastoanalyysia tulee edeltää analyysi tutkimuksen entiteettien olemuksesta ja yleissääntönä mikään tilastoanalyysi ei voi korvata tutkimuksen tavoitteen ja asetelman puutteita.

Myös luottamusvälit eli ”virhemarginaalit” ovat keino tehdä havainnoista yleistäviä johtopäätöksiä. Käytännössä testien ja luottamusvälien antama informaatio tulkitaan samalla tapaa. Jos tehtäisiin sata kertaa samanlainen (esimerkiksi hoito-) tutkimus, niin 95 kertaa tuntematon todellinen (esimerkiksi kuolleisuus-) ero osuisi luottamusvälille ja 5 kertaa sen ulkopuolelle. Uskomme siis, että meidän oma ainutkertainen tutkimuksemme on yksi 95:stä eikä yksi viidestä. Usein näkee tulkittavan, että havaintomme asettuu 95 kertaa sadasta luottamusvälille. Toki havainto on aina luottamusvälillä, päättely koskee todellista, tuntematonta vaikutuseroa.

Testitulos voi olla myös ”ei tilastollisesti merkitsevä” tai vastaavasti luottamusväli saattaa peittää nollaeron. Tuollainen tulos tulkitaan usein siten, että esimerkiksi hoidot olivat käytännön tarpeisiin yhtä hyvät, niillä ei ollut vaikutuseroa. Tuo johtopäätös on virheellinen. Esimerkkinä olkoon havaittu vaikutusero -10 prosenttia eli oma keksintömme on 10 prosenttia vertailuarvoa huonompi, mikä siis ei ollut tilastollisesti merkitsevä. Tällöin nollavaikutuksen kanssa yhtä uskottavaa on, että todellinen vaikutus onkin huomattava eli -20 prosenttia. Havaintoperäisellä tutkimuksella hypoteesi voidaan joskus osoittaa vääräksi, mutta ei milloinkaan oikeaksi. Ei merkitsevä tulos tarkoittaa, että emme tee lainkaan tieteellisiä johtopäätöksiä tutkimushoidon ja vertailuhoidon paremmuussuhteista. Kun käytännössä potilas pitää joka tapauksessa hoitaa, tieteellisessä mielessä avoin tilanne ohjaa meidät käytännössä pitäytymään havaitun -10 prosentin perusteella yhä vertailuhoidossa ainakin siihen saakka, kun pitävä näyttö toisin osoittaa.

TESTI VAI LUOTTAMUSVÄLI VAI MOLEMMAT

Luottamusväli kertoo sen alueen, jolla tutkimuksen todellinen kohde-ero voisi piillä. Testi taas on luonteeltaan päätössääntö, joka antaa mahdollisuuden hylätä yhden luottamusvälin ulkopuolella piilevistä hypoteeseista. Luottamusvälit ovat siten informatiivisempia kuin tilastolliset testit.

Saman eron merkitsevyyttä ei tule analysoida sekä luottamusvälin että testin avulla. Havaintoaineisto on tuollaisessa tilanteessa sama ja mahdolliset johtopäätöserot ovat siten teknisiä ilman sisällöllistä merkitystä. Tutkijan tuleekin itse päättää joko testi- tai luottamusväliesitystavasta.

Ilmeisesti tietokoneohjelmistojen helpon käytettävyyden seurauksena joskus esiintyy päällekkäisiä testejä tavallisesti kehittäjänsä mukaan ristittyinä. Tai vaihtoehtoisesti voidaan testata tilastoanalyysin eri vaiheitten tuloksia. Tai esitetään eri kombinaatioin testejä ja luottamusvälejä. Tuollainen esitystapa on tarpeeton ja harhaanjohtava.

Testi tai luottamusväli koskee tutkimuksesta tehtävää johtopäätöstä. Erityisesti sekoittuneisuuden muodollinen testaaminen on virhe. Harha tulee eliminoida riippumatta sen suuruudesta satunnaisvirheeseen verrattuna. Harhan eliminaatiotarve arvioidaan vertaamalla vakioimatonta ja harhatekijän suhteen vakioitua vaikutusestimaattia.

Testi vai luottamusväli -ongelma palautuu tutkimuksen tavoitteeseen. Jos se kohdistuu yleiseen luonnonominaisuuteen, kuten kysymykseen onko syöpä periytyvä, on perusteltua raportoida luottamisväli eli testiä monipuolisempi informaatio. Sovellutuskysymyksessä, tuleeko syöpähoidot korvata geenihoidolla, tavoitellaan johtopäätöstä, joka antaa hoito-ohjeen. Sen puolestaan tarjoaa tilastollinen testi, joka tiivistää havaintoaineiston tuollaiseksi päätössäännöksi. Erityisesti laajoissa sosiaalilääketieteen ja kansanterveystieteen kysymyksissä, esimerkiksi sote-uudistuksessa, perätään näyttöön perustuvaa toimintaohjetta. Sellaisen siis tarjoaa tieteellisen menetelmän kalupakista tilastollinen testi. Toki sillä rajoituksella, että tilastolliseen testiin perustuva johtopäätös koskee vain yhtä ja rajallista entiteettiä, kuten vain rahaa tai pelkästään potilastyytyväisyyttä tai rajautua biologisen taudin paranemistodennäköisyyteen tms.

VOIMAKKUUS

Tutkimusartikkelin metodiosasta voi löytää maininnan: ”tutkimus suunniteltiin löytämään 25 prosentin kuolleisuusero riskitasolla 5% ja voimakkuudella 80%”. Virke on tuossa yhteydessä tarpeeton ja lisäksi virheellinen.

Tieteellisen metodin kannalta ja havaintoaineiston perusteella tehtävä tutkimuksen johtopäätös (esimerkiksi hoito A on parempi kuin hoito B) riippuu neljästä tekijästä, jotka kukin mittaavat omalta osaltaan tutkimukseen kätkeytyvää epävarmuutta. Ne ovat esimerkissämme vertailtavien hoitojen todellinen vaikutusero eli tuntematon hoidontulosero (25 prosenttia), riski

virheellisesti hylätä se (nolla)hypoteesi, että eroa ei todellisuudessa ole (5 prosenttia), virheellisesti hyväksyä sama hypoteesi, kun nollahypoteesi ei ollutkaan tosi (100-voimakkuus, eli 20 prosenttia) ja potilaitten lukumäärä (n). Nuo neljä muuttujaa eivät ole edes teknisesti riippumattomia: jos kolme kiinnitetään (siis oletetaan), neljäs voidaan laskea noitten kolmen kiinnepisteen avulla. Edellisen lainauksen sana ”suunniteltiin” tarkoittikin, että tutkimuksen minimipotilasmäärä oli suunnitteluvaiheessa laskettu annettujen kolmen oletuksen varassa.

Voimalaskelmien esittäminen varsinaisen tutkimusartikkelin menetelmäosassa on tarpeetonta. Aineistohan on silloin jo olemassa ja siitä on laskettu p-arvo ja muut tilastolliset tunnusluvut, joitten perusteella johtopäätös on tehty, eli annettu hoito-ohje. Voimalaskelman avulla arvattiin tuo johtopäätöstulos, kun toimitaan tiettyjen rajoittavien olettamusten puitteissa, eli oletetaan numeeriset raja-arvot noille kolmelle tieteellisen menetelmän kiintopisteelle.

Lainattu voimalaskelman muotoilu on myös virheellinen. Tutkimus ei nimittäin löydä 25 prosentin todellista eroa. Mahdollisia todellisen eron kandidaatteja on yhä suuri joukko. Edes tutkimusaineistosta laskettu, havaittu ero (eron estimaatti) ei ole 25 prosenttia. Suunnittelun ainoa lupaus on, että (mainittujen erhetodennäköisyyksien puitteissa) havaittu ero tulisi olemaan tilastollisesti merkitsevä (kun hypoteesina on nollaero) siinä teoreettisessa tilanteessa, että todellinen, mutta tuntemattomaksi jäävä, vaikuttavuusero sattuisi olemaan 25%.

Edellinen saattaa tuntua vaikeaselkoiselta, ja sitähän se onkin. Selvennykset ja helpotukset eivät kuitenkaan synny virheitten ja tulkinnanvaraisuuksia lisäävien yksinkertaistuksien kautta.

Voimalaskelmilla on toki oma tehtävänsä. Suunniteltaessa tutkimusta saadaan niitten avulla osviittaa tutkimusaineiston koosta, esimerkiksi haastattelujen vähimmäismäärästä, tai vaihtoehtoisesti, kuinka suuria vaikutuseroja on käytettävissä olevilla resursseilla mahdollista todentaa. Voimalaskelman tärkein merkitys onkin ehkäistä liian pienen tutkimuksen tekeminen, koska esimerkiksi liian harvoihin potilaisiin perustuva kliininen koe jättää johtopäätöksen niin puoleen jos toiseen avoimeksi. Tuo koskee erityisesti (kansainvälisiä) monikeskustutkimuksia, joissa korostuvat rekrytointi-, ajoitus- ja vertailtavuusongelmat. Avoimeksi jäävässä tilanteessa sekä hoito-ohje että hoitokäytännöt jäävät perustumaan ennakkokäsityksille. Ne ovat alttiina virheille ja vakiintuessaan estävät myöhemmän tutkimusnäytön, jos vakiintuneeseen hoitoon kohdistuvaa satunnaistettua koetta pidetään epäeettisenä.

MERKITTÄVÄ

Empiirisesti, havaintoperäisesti ei voida (nolla)-hypoteesia todistaa oikeaksi. Eli on mahdotonta esimerkiksi osoittaa, että vertailtavien hoitojen ennuste-ero on nolla. Ongelma on kierretty kliinisessä tutkimuksessa määrittelemällä pienin kliinisesti merkittävä ero. Jos todellinen hoidontulosero alittaa tuon rajan, ajatellaan olevan käytännön kannalta merkityksetöntä kumpiko hoito valitaan. Käytännössä oma keksintömme legitimoituisi, jos ennuste-ero pysyisi tuon raja-arvon alla riippumatta eron suunnasta.

Erityisesti kliinisessä hoitotutkimuksessa on paitsi mahdotonta, myös tarpeetonta yrittää havaintoperäisesti osoittaa, että hoitojen vaikutusero on nolla, että hoidot olisivat yhtä hyvät. Potilas tullaan joka tapauksessa hoitamaan vain yhdellä noista vertailtavista hoitomuodoista. Jos molempien hoitojen vaikutus on täsmälleen sama, ei ole väliä kumpiko hoito valitaan. Hoito-ohje on virheellinen ja vaarallinen, jos valitaan huonompi hoito.

Kliinisesti merkittävä ero voidaan tieteellisessä menetelmässä korvata huonomman hoidon valintariskillä. Huonomman hoidon valintaa tulee välttää riippumatta vaikutuseron suuruudesta. Valintariskille, todennäköisyydelle että valitaankin huonompi hoito, voidaan kiinnittää samanlainen raja-arvo kuin esimerkiksi p-arvon 5 %. Näin päästään eroon vaikeasti hahmottuvasta käsitteestä kliinisesti merkittävä ero ja tutkimus voidaan suunnitella siten, että järkevissä rajoissa vältetään tuollainen virheellinen ja vaarallinen hoito-ohje. On tuntematonta, miksi tämä ajattelutapa ei ole saanut kannatusta. Olisiko syynä se, että ajatuksen isät Schwartz ja Lellouch ovat kaksi ranskalaista – eikä angloamerikkalaista – tutkijaa.

JOHTOPÄÄTÖS

Tutkimuksen johtopäätös on vastaus sen tavoitteeseen, täsmällisemmin (nolla)hypoteesin uskot-

tavuuteen, krebiliteettiin. Tieteellinen metodi formalisoi sattuman vaikutuksen tutkimustuloksiin p-arvojen, luottamusvälien, testin voimakkuuksien ja muitten tilastollisten tunnuslukujen avulla. Niillä on satunnaisjakauma, jonka määräarvo, kuten p=0.05, on muotoutunut tieteellisen johtopäätöksen kriteeriksi. Raja- arvojen ajatellaan olevan tieteellisesti perusteltuja optimiarvoja. Noitten määräarvojen asettamiselle (kuten p-arvon 5 prosenttia) ei kuitenkaan ole objektiivista, tieteeseen pohjaavaa perustetta. Ne perustuvat käytäntöihin, joista muutama on noussut lähes dogmin asemaan. Mainituilla numeerisilla raja-arvoilla on ratkaiseva vaikutus tutkimuksen johtopäätökseen ja siten lääketieteen kokonaisuuteen. Se puolestaan on näyttöpohja terveydenhuollolle, kuten varsinaiselle lääkinnälle, eli diagnostiikalle ja hoidolle. Toki vain siltä osin kuin käytännön toiminta on näyttöön eli tutkimukseen perustuvaa.

Vuosien mittaan raja-arvot ovat löystyneet. Esimerkiksi pienin ”kliinisesti merkittävä ero” on lähes poikkeuksetta – ja joskus huomattavasti – suurempi kuin ne hoidontuloserot, joitten perusteella käydään asiantuntijadebattia. Tilastollisen merkitsevyyden raja on puolestaan parin tutkijapolven aikana löystynyt yhdestä viiteen prosenttiin. Tällöin esimerkiksi olematon ennuste-ero tulkitaan todelliseksi viisi kertaa aiempaa useammin siinäkin tapauksessa, että todellisuus olisi pysynyt muuttumattomana. Tuo sopii hyvin yhteen sen kanssa, että tutkimuksen julkaistavuuden (virheellisenä) edellytyksenä pidetään ”positiivista” eli tilastollisesti merkitsevää tulosta.

Tieteellinen menetelmä asettaa lääketieteen tutkimuksen kausaalitutkimuksen kehikkoon. Sattuman suvaitseva metodi kattaa mm. diagnostiikan, hoidon ja terveydenhuollon rakenneratkaisut. Käytännön terveydenhuollossa on aina vääriä diagnooseja, vääriä hoidonvalintoja, vääriä sairaalainvestointeja jne. siinäkin tapauksessa, että ne olisivat näyttöön, lääketieteelliseen tutkimukseen perustuvia. Itse tutkimus, parhaimmillaankin, kun on altis virhepäätelmille.

Matti Hakama