TOXTEST – ulkopuolinen tieteellinen arvio

Esa Läärä ja Olavi Pelkonen

Lähtökohta

Sosiaali- ja terveysministeriö on pyytänyt allekirjoittaneita (Prof Esa Läärä ja Prof emeritus Olavi Pelkonen, Oulun yliopisto) toteuttamaan riippumattoman asiantuntija-arvion THL:n, TTL:n, HY:n ja TY:n yhteistyössä laatimasta TOXTEST-loppuraportista sen sisältöön ja loppupäätelmiin liittyvistä ristiriitaisista arvioista. Arvioijilla on ollut käytössään loppuraportin lisäksi tulokset sisältävä Excel-tiedosto, alkuperäisen hakemuksen liite ja täydennys ja lisäksi loppuraportin julkistamistilaisuudessa pidettyjen esitysten PP-tiedostot.

Mitä tutkittiin ja mitä saavutettiin

Arvioitsijoiden lyhyt tiivistelmä hankkeesta ja sen tuloksista:
Tavoite:
kehittää **sisäympäristönäytteille soveltuva toksisuuden arviointimenetelmä**, jota voidaan käyttää terveysvalvonnan tukena **homevaurion vakavuuden arvioinnissa** ja **kosteusvauriokohteiden korjauksen priorisoinnissa**.
Suunnitelma:
Näytteenoton ja näytteen käsittelyn kehitys (vaihe 1)
Kenttäkäyttöön soveltuvan toksisuustestimenetelmän valinta (vaihe 2)
Valittujen toksisuustestien soveltaminen laajassa aineistossa (vaihe 3)
Tutkimusasetelma: Vaurioympäristö ("oireileva") vs kontrolliympäristö ("oireilematon")
Toksisuustestaus: Näytteiden toksisuuden testaus hankkeen eri vaiheissa kaikkiaan kuudella eri solutyypillä ja noin kahdellakymmenellä menetelmällä
Tulokset:
Analyysi ja tulkinta tapahtuivat näytekohtaisten, kohdekohtaisten ja testikohtaisten tulosten perusteella monenlaisia vertailuja tehden.
Johtopäätökset:
Loppuraportin perusteella hankkeen varsinainen tavoite jäi saavuttamatta: **"huonepölyuutoksen toksisuutta ei voida käyttää kosteusvauriokohteiden luokittelussa tai terveyshaitan arvioinnissa"**. Muutoinkin johtopäätökset ovat hyvin varovaisia. Hyödyllistä tietoa jatkotutkimuksia varten tuotettiin runsaasti.

Raportin muodollinen ja sisällöllinen laatu

Raportti liitteineen on varsin mittava ja monessa kohdassa yksityiskohtainen, joskin paikoitellen varsin huolimaton kielellisesti ja asiallisesti (esim lyhenteitä on jätetty selittämättä; liitteiden numerointia ei ole viety itse liitteisiin jne). Ilmeisesti hankkeelta on puuttunut ainakin raportin koostamisvaiheessa vastuullinen johtaja, joka olisi pitänyt huolen raportin muodollisesta laadusta.

Lukijan kannalta suurin yksittäinen puute on **taustatietouden niukkuus**; raportista ei löydy selkeää argumentointia, mikä on ihmisten kosteusvaurio-oireiden kausaalinen tai tilastollinen yhteys (tai

yleensä "uskottava" yhteys, jos se perusteltavissa tieteellisesti, ks alempana) erilaisiin kosteusvaurioympäristössä esiintyviin tekijöihin lyhyellä ja pitkällä aikajänteellä tai miten menetelmävalikoima valittiin ja perusteltiin. Tässä suhteessa lukija on pitkälti muun ja muualta hankkimansa tiedon varassa. Koska kysymyksessä on taloudellisesti valtavan mittaluokan ongelma ja tieteellisesti kompleksinen ja haasteellinen tutkimuskohde (ja melkeinpä päivittäinen uutisointiaihe julkisuudessa), olisimme odottaneet paljon huolellisemmin tehtyä, perusteellisesti taustoitettua ja kriittisesti argumentoitua raporttia.

TOXTEST-hankkeen tavoitteet

Kuten loppuraportissa todetaan, TOXTEST-hankkeen tavoitteena oli kehittää **kenttäkäyttöön soveltuva toksisuustesti**, jota voitaisiin hyödyntää kosteus- ja homevaurion vakavuuden arvioinnissa. Mielestämme on selvää ja epäilyksetöntä, että tämä **tavoite sinänsä on erittäin merkittävä** niin tieteellisesti, taloudellisesti kuin myös yhteiskunnan ja yksilön kannalta. Onko tämä "suuri tavoite" realistisesti toteutettavissa hankkeen käytössä olevilla resursseilla ja ajalla onkin jo varsinainen haaste, johon vastaaminen ratkaisee hankkeen onnistumisen. Voidaan myös kysyä, onko alan tutkimus ja tietämys edennyt sellaiseen vaiheeseen, että asetettu tavoite voidaan saavuttaa. Lienee syytä muistuttaa, että tieteellisessä(kin) tutkimuksessa "paholainen on yksityiskohdissa" ja siten arviointimme pääosan muodostaakin menetelmien, lähestymistapojen, tulosten esittämisen ja tulkinnan ja johtopäätösten varsin yksityiskohtainen tarkastelu.

Hanke eteni kolmessa vaiheessa, joille oli asetettu yksityiskohtaisemmat tavoitteet:

Tavoite 1. Luoda sisäympäristönäytteille soveltuva näytteiden uuttomenetelmä sekä määrittää varsinaisiin kokeisiin soveltuvat annokset ja altistusajat **("miten valmistaa edustavat näytteet")**.

THL:n ympäristötoksikologian laboratoriolla on pitkä kokemus erityyppisten näytteiden keruusta, valmistelusta ja tutkimisesta. Keräysmenetelminä käytettiin laskeutuneen huonepölyn keräämistä imuroimalla, pyyhkäisynäytteinä tai keräyslaatikkoon, sekä laskeumamaljoja, joihin kasvanut kasvusto kerättiin talteen. Uuttoliuoksena käytettiin etanolia, joka haihdutettiin ja sakka liuotettiin pieneen määrään metanolia. Metanoliliuosta käytettiin toksisuustesteihin.

Raportti sisältää tarpeeksi yksityiskohtaista informaatiota, jonka perusteella näytteiden keräystä, valmistelua ja käyttökelpoisuutta toksisuustestauksiin on mahdollisuus arvioida. Tulosten perusteella näytteiden säilytysaika ja säilytysmateriaali voivat vaikuttaa tuloksiin, mutta sinänsä näytekeräysvälineiden esikäsittely ja näytteiden keräys ja uutos uskoaksemme varmistavat, että näytteet edustavat lyhyen tai pidemmän ajan otosta sisäilmasta tai että esimerkiksi sekoittava kontaminaatioriski on pieni. Koska kysymyksessä kuitenkin olivat "realistiset" näytteet, niiden tarkempi ja yksityiskohtaisempi luonnehdinta olisi voinut tuoda lisäarvoa näytteiden edustavuuden ja kohdentuvuuden arviointiin.

Raportissa ei pohdita **pölynäytteiden edustavuutta ja koostumusta**, mikä on selkeä puute. Näytteenkeräysvälineen sijoittamisesta vauriolähteen lähelle ja/tai yleisempiin tiloihin löysimme viittauksia asianomaisissa protokollissa (liitteet 1 ja 4), mutta nähdäksemme tämänkaltaisen potentiaalisen paikkatekijän analysointia ei ole käytetty hyväksi tai sitä ei ole mainittu. Näytteenkeräyksen aika on toinen intuitiivisesti ajatellen hyvin merkittävä tekijä, mutta tämäkin jää varsin lyhyelle pohdinnalle raportissa.

Yleensä altistumistutkimuksissa pyritään tutkimaan altisteiden koostumusta ja ominaisuuksia, jotta voitaisiin löytää **altistumista kuvaavia vertailukelpoisia mittareita** (altistumisen markkerit). Epäilemättä sisäilmanäytteet ovat hyvin monimutkaisia seoksia ja hankalasti luonnehdittavissa, mutta siitä huolimatta tai ehkä juuri sen takia, **näytteiden fysikaalinen, fysikokemiallinen ja kemiallis-analyyttinen tutkiminen** olisi ollut tärkeää (mikrobiologista ja mikrobitoksiini-analytiikkaa tehtiin jossain määrin sen lisäksi että puhdasviljelmäuutteita tutkittiin 3. vaiheessa). Emme selvittäneet arvioinnin kuluessa, olivatko osallistuneet ryhmät julkaisseet analyyttisiä tutkimuksia, mutta jos työryhmillä oli tämänkaltaista tietoa olemassa, sitä olisi pitänyt käsitellä raportissa.

Yleensäkin, näytteiden edustavuuden ja koostumuksen pohdinta vauriokohteiden laadun ja vakavuuden ja altistuneiden akuuttien ja kroonisempien oireiden ja sairauksien yhteydessä (yhdistettynä asianmukaiseen kliiniseen diagnostiikkaan, huom. altistumisen ja oireiden biomarkkerit) olisi saattanut johtaa hankkeen kannalta hyödyllisiin uusiin muunneltuihin tutkimusasetelmiin.

1. vaiheen näytekeräyksen kehitystyöhön valitut kohde- ja kontrollinäytteet (2 + 2) antoivat samankaltaisia vasteita käytetyissä toksisuustesteissä, jonka takia keräys- ja säilytysohjeita muutettiin seuraavia vaiheita varten.

Tavoite 2. Osoittaa ne toksikologiset tutkimusmenetelmät, jotka ovat käyttökelpoisimpia sisäympäristönäytteiden toksisuuden testaukseen, sekä osoittaa näiden menetelmien luotettavuus useilla rinnakkaisilla analyyseillä **("Miten valita osuvat toksisuustestit")**.

Näytteiden toksisuutta testattiin eri vaiheissa kaikkiaan **kuudella eri solutyypillä ja noin kahdellakymmenellä menetelmällä ("potentiaalisella biomarkkerilla",** ks jäljempänä**)**, jotka kaikki olivat jo käytössä osallistuvissa laitoksissa. Raportissa tai taustamateriaalissa ei käynyt ilmi, oliko käytettyjen toksisuusmenetelmien valinnassa muita kriteerejä kuin se, että niitä oli käytetty vastaavantyyppisissä tai muihin tarkoituksiin tehdyissä tutkimuksissa jo aiemmin ja/tai ne kuuluivat laitosten menetelmävalikoimiin. Alempana esitämme perusteellisemman arvion testien valinnasta ja tutkimusasetelmasta.

Oleellisin tulos 2. vaiheessa oli se, että vaikka useimpien testien tulokset korreloivat toistensa kanssa varsin hyvin, ts. ne ilmeisesti mittasivat samankaltaisia tekijöitä, testit eivät kyenneet luotettavasti erottelemaan vaurioasuntoja kontrolliasunnoista (toisen vaiheen 4 + 4 kohdetta).

Koska toksisuustestit eivät kyenneet luotettavasti erottelemaan vaurioasuntoja kontrolliasunnoista, ensimmäinen alustava johtopäätös mielestämme on se, että

1) kerättävät näytteet eivät edusta niitä tekijöitä, jotka johtavat ihmisillä kosteusvaurio-oireisiin ja sairauksiin,
2) käytetyt toksisuustestit eivät mittaa terveyshaittojen esiintymisen kannalta tärkeitä tekijöitä, ja/tai
3) mittausten toistettavuus on epätyydyttävä.

Emme tiedä, käytiinkö hankkeessa tämänkaltaisia keskusteluja tai pohdiskeltiinko muita mahdollisia näytteenottotapoja tai toksisuustestejä, mutta lopputulos kuitenkin oli, että jatkossa lisättiin ainoastaan kohteita ja näytteitä ja supistettiin menetelmävalikoimaa.

Hankkeen toksikologista taustaa, lähestymistapoja ja menetelmien valintaa käsitellään alempana perusteellisemmin.

Tavoite 3. Testata käyttökelpoisimmilla toksisuustesteillä eri näytteenkeräysmenetelmiä käyttäen, voidaanko oireilevat vauriokohteet erottaa oireilemattomista kohteista **("Miten kehitetty ja valittu tutkimusasetelma/menetelmävalikoima toimii tositilanteessa").**

Hankkeen kolmannessa vaiheessa valitut toksisuustestit pantiin varsinaiseen kokeeseen: näytteitä kerättiin kuntoarvioinnin ja terveyskyselyn perusteella valituista vaurio- ja kontrollirakennuksista ja testattiin sokkoutetusti (myös muissa vaiheissa sokkoutus oli protokollassa) erityisesti sian siittiötestillä ja bakteeritestillä. Koska tämä vaihe oli hankkeen tavoitteiden kannalta ratkaiseva, käsittelemme sen tuloksia ja niiden tulkintaa laajemmin alempana.

Toksikologisen taustan arviointi: TOXTEST-hanke biomarkkeritutkimuksena

Arvioijien käsityksen mukaan tutkimuksen toksikologinen lähestymistapa kuuluu ns biomarkkeritutkimuksen alueelle; tällaisesta tutkimuksesta on olemassa valtava määrä tieteellistä kirjallisuutta erityisesti lääkekehityksen alueella (erityisesti ns. translationaalinen tutkimus), mutta myös ympäristötoksikologisessa tutkimuksessa.

Sairaus/oire-suuntautuva biomarkkeri (kosteusvaurio → haitalliset tekijät → sairaus/oire) mittaa sellaista haitallista tekijää, joka on riittävän hyvin osoitettu olevan **terveyshaitan kausaalinen tekijä**. Kausaalisuus edellyttää taustalla olevien biologisten ja mekanististen prosessien tuntemusta tai ainakin sitä, että yhteys haitallisen tekijän ja sairauden tai oireen välillä on tieteellisesti "uskottava". Tällöin tarvitaan ns. weight-of-evidence –tyyppistä lähestymistapaa, jolloin kaiken relevantin tutkimustietouden valossa punnitaan yhteyden uskottavuutta. Yhteyden "uskottavuus" edellyttää, että tunnistetaan ja luonnehditaan ainakin tärkeimmät haitalliset tekijät (synteettiset kemikaalit, mikro-organismien tuottamat toksiinit, mikrobit itsessään, muut haittatekijät jne) ja vieläpä niin, että tärkeimmistä on olemassa tutkimustietoa, miten ja kuinka paljon ne altistavat elimistöä ja mitkä ovat niiden vaiheet ja potentiaaliset vasteet elimistössä niin lyhyellä kuin pitkälläkin tähtäimellä.

On syytä painottaa, että **aikajänne toksikologisissa ilmiöissä** on eräs tärkeimpiä pohdittavia tekijöitä niin altistumisen (akuutti – pitkäaikainen) kuin toksisten vasteiden (akuutti – subakuutti – krooninen – viivästynyt) suhteen ja aikajännettä olisi hyvä tarkastella liittyen potentiaalisiin altistaviin tekijöihin. Eräs merkittävä puute loppuraportissa on se, ettei näytteiden koostumuksesta löydy mitään mainintaa lukuun ottamatta mainintoja toksiineista ja mikrobeista (ks liite 5), vaikka oletettavasti ryhmillä tällaista tietoa on joko itse tuotettuna tai kirjallisuudesta kerättynä.

Koska mahdollisia haitallisia tekijöitä on todennäköisesti lukuisia, "sisäilmaongelmien selvittämisessä olisi erityisen hyödyllistä käyttää menetelmää, joka ottaa huomioon eri altisteiden yhteisvaikutukset yksittäisen altisteen määrän mittaamisen sijasta." Sinänsä tämä perustelu on järkevä ja myös tieteellisesti perusteltavissa, mutta se ei voi mitätöidä yksittäisten altisteiden mittaamisen hyötyä ja merkitystä. Yleensäkin kompleksisten seosten tutkimuksissa on tutkittava sekä seosta että sen yksittäisiä toksikologisesti merkityksellisiä komponentteja ottaen huomioon mahdolliset yhteisvaikutukset.

Valittu tutkimusasetelma

Ilmeisesti kuitenkin ylläkuvattu prosessi kosteusvauriosta lukuisten haitallisten tekijöiden kautta terveyshaittoihin on niin monimutkainen ja suurelta osin huonosti tunnettu, että tutkimuskonsortio painottaa päinvastaista lähestymistapaa: **voitaisiinko valittujen potentiaalisten biomarkkerien joukosta osoittaa ne testit, jotka selvimmin ja luotettavimmin erottaisivat oireettomat kontrollikohteet ja oireilevat kosteusvauriokohteet**. Tällaisessa lähestymistavassa kausaalisten yhteyksien ja mekanismien osoittaminen ei ole periaatteessa edes tarpeellista, jos kartoittavien tutkimusten perusteella tunnistetaan spesifinen ja sensitiivinen testi, joka riittävällä luotettavuudella erottelee ongelmakohteet ongelmattomista. Tässä tapauksessa **tieteellinen arviointi pelkistyy aika pitkälle tilastolliseen arviointiin: kuinka hyvä erotuskyky on tilastollisesti**. Kuitenkin biomarkkerin käyttö olisi paljon varmemmalla pohjalla, jos tieteellinen relevanssi olisi riittävässä määrin tutkittu ja tunnettu.

Regulatorisen biomarkkerin kehittäminen

Tavoitellun arviointimenetelmän kehittäminen on analoginen regulatorisen biomarkkerin kehittelyn kanssa ja sisältää tekstilaatikossa kuvatut vaiheet soveltuvin osin. Arvioijien käsityksen mukaan taustan käsittely oheisen vaiheluettelon (TextBox) tai vastaavan rakenteen avulla olisi hyvin hyödyllistä ja kertoisi eksplisiittisesti, minkälaista tutkimustietoutta on olemassa kunkin valitun potentiaalisen biomarkkerin taustalla. Tämä liittyy biomarkkerin tieteelliseen relevanssiin (kohdat 1-3(4) oheisessa tekstilaatikossa), joka on selkeästi eri asia kuin biomarkkerin kvalifikaatio (luotettavuus, toistettavuus jne, kohdat (4)5-7). Nyt tieteellisen relevanssin pohdinta puuttuu lähes

TextBox. *Crucial steps to be taken into consideration when creating biomarkers for regulatory purposes (based on the biomarker qualification exercise as described in Dieterle, F., Sistare, F., Goodsaid, F., Papaluca, M., Ozer, J.S., Webb, C.P., et al., 2010. Renal biomarker qualification submission: a dialog between the FDA-EMEA and Predictive Safety Testing Consortium. Nature Biotechnology 28, 455–462.)*

1) Identification of biomarkers from non-clinical or clinical setting
2) Prioritization of identified biomarkers (on the basis of potential biomarker sets created in point 1)
3) Characterization of biomarkers
4) Validation of biomarkers
5) Preliminary exercise with prioritized biomarkers
- Linking to non-clinical or clinical endpoint or biological process
6) Qualification/verification of biomarkers
- Reproducibility and transferability of assays between laboratories and processes
- Suitability of an assay: accuracy, precision, limit of quantification (LoQ), stability of an analyte, etc
7) Acceptance of biomarkers for regulatory use

kokonaan raportista.

Tieteellisen relevanssin pohdinnan puute tulee erityisesti ilmi solumallien ja toksisuusindikaattorien valinnassa. Ilmeisesti tutkijat ovat olettaneet *a priori*, että valitut solut ja päätevasteet ovat asianmukaisia ja merkittäviä kosteusvaurionäytteiden toksikologisten ominaisuuksien mittaamisessa ja oikeastaan tärkein tavoite oli tutkimusten perusteella valita

"toksisuustesteistä edullisimmat ja nopeimmat". Jos taas *a priori* oletus on puutteellinen tai virheellinen, ei "edullisuudella ja nopeudella" ole kovin suurta tieteellistä relevanssia.

Toksisuustestien tulokset ja tulkinta

Toksiset vasteet (niin kuin biologiset vasteet yleensäkin) ovat **kvantitatiivisia** ja ilmaistavissa havainnollisimmin annos/pitoisuus–vaste –suhteina. Se että regulatorisissa yhteyksissä käytetään usein erilaisia **raja-arvoja** ja toksisten aineiden **luokittelua**, ei vähimmässäkään määrin mitätöi annos-vaste-suhteiden selvittämistä. Säätelyn tavoite on riskinhallinta ja raja-arvot toksisille aineille ja altistumisille kuvaavat riskinarvioinnille perustuvia "turvallisiksi uskottuja" annoksia tai altistumisia, jotka taas perustuvat annos-pitoisuus-vaste-suhteitten selvittelyyn.

Useimmat hankkeessa käytetyt toksisuustestit ovat perimmältään kvantitatiivisia eivätkä dikotomisia ja perustuvat ainakin muutamaan pitoisuus-vaste-pisteeseen. Tällaisten tulosten tarkistaminen ja arviointi eivät aiheuttaneet arvioijille ongelmia (paitsi aineiston tilastollisen käsittelyn osalta, ks alempana). Ehkä eniten meitä hämmensi siittiötestin tulosten käsittely ja tulkinta, jossa esiintyi epäjohdonmukaisuuksia ja puutteita, joiden taustasta oli vaikea saada selkoa (ks alempana). Myös siittiötesti on kvantitatiivinen ja tulosten perusteella voitiin laskea esim. EC50-arvot näytteille. Ryhmä käytti raportissa myös luokittelua, jossa merkittävin raja-arvo oli 40 µg/ml, jonka alapuolella olevat arvot edustavat "merkittävää toksisuutta". Miksi tämä nimenomainen arvo oli valittu luokituksen perustaksi, jäi meille epäselväksi. Lisäksi tämän raja-arvon määrittelyssä esiintyi epäjohdonmukaisuutta, koska raportista löytyy kaksi raja-arvoa: <40 ja ≤40. Kovin suurta merkitystä lopputulokset kannalta raja-arvon valinnalla ei varmaankaan ole, mutta osaltaan tämäkin epäjohdonmukaisuus osoittaa loppuraportin viimeistelyn puutetta.

Pölynäytteiden toksisuuden luokittelu on erityisen ongelmallista, koska ne ovat kompleksisia seoksia, joiden komponentit ovat useimmiten tuntemattomia ja koska ei ole mitään yhteismitallista toksisuusasteikkoa johon kyseinen arvo voitaisiin asettaa. Kyseessä on lisäksi *in vitro* –menetelmä, jonka tulosten ekstrapolaatio *in vivo* –tilanteeseen tai ihmiselle aiheutuvan vaaran ja riskin arviointiin on nykytiedon valossa hyvin vaikeaa tai jopa mahdotonta, eikä tällä hetkellä ole näkyvissä sellaista tutkimuksellista perustaa, että validaatio olisi mahdollista. Tässä yhteydessä on syytä painottaa, että in vitro –testitulos osoittaa harvoin muuta kuin terveydelle aiheutuvaa potentiaalista "hypoteettista" vaaraa (hazard), jonka todellista suuruutta on vaikea arvioida ja joka nimenomaan ei ole riski; tuloksen "jalostaminen" määrälliseksi riskiksi (ihmiselle) edellyttää in vitro-in vivo –ekstrapolaation validaatiota, joka taas on yleensä vuosia kestävä kansainvälinen tutkimushanke (esim. ECVAM:in koordinoimat validaatiohankkeet).

Raja-arvo näyttää olevan merkittävässä osassa tutkijoiden kehittämässä "toisen tason" riskinarviointiluokittelussa, joka nojaa elintarviketurvallisuuden ajatteluun (liite 11, kohdat 11.2.4 ja 11.3.1). Tässä "toisen tason" luokittelu perustuu paitsi yllämainittuun raja-arvoon myös kustakin kohteesta määritettyjen näytteiden luokitteluun raja-arvojen suhteen (ks. alempana). Luokittelulla ei ole mitään syvällisempää tieteellistä perustaa, lisäksi sen perusteella hylätään merkittävää aineistoa (vain kolmen näytteen kohteet hyväksyttiin, ks. alempana) ja lopuksi se näyttää kuuluvan enemmän riskinhallinnan puolelle. Viimeinen kohta on hyvin merkittävä, koska riskinhallinnassa astutaan tieteellisen tutkimuksen ulkopuolelle ja mukaan tulevat monet yhteiskunnalliset, poliittiset ja arvostukselliset asiat. Mielestämme riskinhallinnan perustelu kyseenalaisilla tieteellisillä argumenteilla ei aja sen enempää riskinarvioinnin kuin riskinhallinnankaan asiaa.

TOXTEST-hankkeen tilastotieteellinen arviointi

Siirrymme nyt tarkastelemaan raportin tilastollista osuutta varsin yksityiskohtaisesti, koska valittu tutkimusasetelma edellyttää asianmukaisia tilastollisia työkaluja ja niiden pätevää käyttämistä, ja itse asiassa hankkeen tulokset ja johtopäätökset perustuvat (lähes) täysin tilastolliseen käsittelyyn.

Toksisuustestien tilastollisen arvioinnin periaatteista

Toksisuustestien hyvyyden tilastollinen arviointi nojautuu pääosin samoihin periaatteisiin ja menetelmiin kuin diagnostisten ja seulontatestien ominaisuuksien arviointi lääketieteen piirissä. Perusteellisen johdatuksen aihepiiriin tarjoaa esim. Margaret Sullivan Pepen monografia (Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, Oxford, 2003.)

Arvioinnissa käytettävän aineiston havaintoyksiköt (kuten ihmiset tai talot) muodostavat parhaimmillaan edustavat osajoukot niistä testattavan kohdehäiriön (tietty sairaus tai kosteusvaurio) esiintymisen suhteen erilaisiksi luokiteltavista kohdepopulaatioista, joiden erotteluun arvioitavaa testimenetelmää suunnitellaan käytettäväksi. Testimenetelmän alkuvaiheen tutkimuksissa on kuitenkin usein hyödyllistä valita havaintoyksiköt yhtäältä "sairaista sairaimpien" ja toisaalta "terveistä terveimpien" populaatioista. Jatkotutkimuksiin ja -kehittelyihin kannattaa kelpuuttaa vain ne menetelmät, joiden erottelukyky näin selvästi poikkeavien ryhmien kesken on lähes täydellinen.

Yksittäisen testimenetelmän erottelukykyä mitataan asianmukaisilla parametreilla eli määrällisillä tunnusluvuilla. Keskeisille parametreille lasketaan ja raportoidaan paitsi piste-estimaatit niin myös estimaattien tilastollista epävarmuutta kuvaavat virhemarginaalit eli **luottamusvälit**.

Kun kohdehäiriö on dikotominen ("sairas" vs. "terve"), ja testin antamat tulokset on selvästi jaettavissa positiivisiin ja negatiivisiin, niin erottelukykyä on tapana kuvata sellaisilla parametreilla kuin **sensitiivisyys** (Se), eli oikean positiivisen tuloksen todennäköisyys, ja **spesifisyys** (Sp), eli oikean negatiivisen todennäköisyys. Spesifisyyden asemesta voidaan yhtäpitävästi tarkastella **väärän positiivisen riskiä** (*false positive risk*, FPR = 1 − Sp). On yleisesti perusteltua odottaa, että niin sensitiivisyys kuin spesifisyys ovat kumpikin vähintään 50 % ja mieluummin kauempana tästä arvosta kuin ideaalitasosta, joka on 100 %. Mielekkäällä testillä sensitiivisyys on joka tapauksessa huomattavasti suurempi kuin väärän positiivisen riski, ja väärän negatiivisen riski on reilusti pienempi kuin spesifisyys. Sen sijaan ei ole mahdollista asettaa yleispäteviä kynnysarvoja sille, millaiset sensitiivisyyden ja spesifisyyden arvot ovat tarpeeksi suuret. Kussakin sovelluksessa riittävyyttä koskeva harkinta perustuu saatavilla olevaan informaatioon ao. päätöksentekotilanteen muista piirteistä, kuten (a) kuinka suuri on testattavan kohdehäiriön todellinen yleisyys eli **vallitsevuus** (*prevalence*, Pr), ja (b) kuinka vakavia haitallisia seurauksia on yhtäältä vääristä negatiivisista ja toisaalta vääristä positiivisista tuloksista.

Sensitiivisyys ja väärän positiivisen riski voidaan tiivistää yhteen parametriin esim. laskemalla niiden pohjalta **Youdenin indeksi**: $J = \mathrm{Se} + \mathrm{Sp} - 1 = \mathrm{Se} - \mathrm{FPR}$. Youdenin indeksin viitepisteenä on 0, joka kuvastaa täydellistä erottelukyvyn puutetta (vrt. lantin- tai nopanheitto). Käyttökelpoiselta testiltä on perusteltua odottaa, että Youdenin indeksi on selvästi lähempänä 1:tä kuin 0:aa. Jos arviointiaineistosta laskettu Youdenin indeksin luottamusväli kattaa sekä negatiivisia että positiivisia arvoja, niin aineisto ei anna ainakaan riittävästi näyttöä sen puolesta, että testillä olisi minkäänlaista erottelukykyä.

Kaksiarvoisen testin erottelukykyä voidaan mitata vaihtoehtoisesti myös uskottavuusosamäärillä. **Positiivisen tuloksen uskottavuusosamäärä** LR+ ja **negatiivisen tuloksen uskottavuusosamäärä** LR− määritellään sensitiivisyyden ja väärän positiivisen riskin pohjalta seuraavasti: LR+ = Se/FPR ja LR− = (1−Se)/(1− FPR). Aiemman tarkastelun perusteella LR+:n pitää olla reilusti 1:tä suurempi ja LR−:n selvästi alle 1:n. Jos näiden arvot olisivat LR+ = LR− = 1, niin testi antaisi kohteen todellisesta tilasta täysin riippumattomia tuloksia (kuten lantin- tai nopanheitto). Jos taas LR-luvun luottamusvälin alaraja on ykkösen alapuolella ja yläraja 1:n yläpuolella, voidaan sanoa, että toksisuustestin erottelukyvystä ei ole mitään näyttöä. Uskottavuusosamäärien etuna on mm. se, että ne yleistyvät moniluokkaiselle tai jatkuvalle testimuuttujalle.

Testin **positiivinen ennustearvo** PV+, eli ehdollinen todennäköisyys sille, että positiivisen testituloksen saanut havaintoyksikkö on todella sairas tai vaurioitunut, riippuu vallitsevuudesta ja positiivisen tuloksen uskottavuusosamäärästä Bayesin kaavan mukaan. Samoin **negatiivinen ennustearvo** PV− riippuu vallitsevuudesta ja negatiivisen tuloksen uskottavuusosamäärästä.

Kun testimenetelmän erilaiset mahdolliset tulokset muodostavat joko moniluokkaisen järjestysasteikollisen muuttujan tai mittayksiköllä varustetun kvantitatiivisen muuttujan T, niin sensitiivisyys Se(t) ja väärän positiivisen riski FPR(t) on periaatteessa määriteltävissä kullekin mahdolliselle kynnysarvolle tai katkaisupisteelle t, joka sisältyy testimuuttujan T vaihtelualueeseen. Tällaiseen testimenetelmän kokonaisvaltaista erottelukykyä voidaan arvioida ja verrata muiden samaan tarkoitukseen ajateltujen testien kanssa, joiden asteikot ja mittayksiköt voivat olla toisistaan hyvinkin poikkeavia, käyttäen **ROC-käyränä** (*Receiver Operating Characteristic curve*) tunnettua kuviota. Se konstruoidaan piirtämällä murtoviiva eri katkaisupisteiden t kohdalla estimoitujen FPR(t) ja Se(t)-lukujen muodostamien x-y-koordinaattien kautta. Käyrä lähtee yksikköneliön vasemmasta alakulmasta ja päätyy neliön oikeaan yläkulmaan, ja hyvällä testillä se kulkee läheltä neliön vasenta yläkulmaa. Käyrän alle jäävä pinta-ala AUC (*area under curve*) on suoraviivaisin tapa mitata testin erottelukykyä yhdellä parametrilla, ja sen estimointiin liittyvän virhemarginaalin suuruutta arvioidaan asianmukaisen luottamusvälin avulla. Täydellisellä testillä AUC = 1, kun taas täysin erottelukyvyttömän testin käyrä kulkee pitkin neliön lävistävää diagonaalia, jossa Se(t) = FPR(t) jokaisella t, ja sen AUC = 0.5.

Uskottavuusosamäärät yleistyvät moniluokkaisen tai kvantitatiivisen testimuuttujan T asetelmaan seuraavasti. Olkoon nyt $p(t \mid D)$ arvon $T = t$ todennäköisyys tai todennäköisyystiheys sairaiden yksiköiden osapopulaatiossa D, ja vastaavasti $p(t \mid H)$ yhtä kuin arvon t todennäköisyy(stihey)s terveiden osapopulaatiossa H. Minkä tahansa mahdollisen testituloksen t uskottavuusosamäärä määritellään näistä: $\mathrm{LR}(t) = p(t \mid D) \,/\, p(t \mid H)$. Uskottavuusosamäärän voidaan odottaa olevan testimuuttujan arvojen t monotoninen funktio. Se on kasvava, jos suuret testimuuttujan T arvot ovat tavallisempia sairailla kuin terveillä, ja vähenevä päinvastaisessa tapauksessa. Uskottavuusosamäärä on hyödyllinen yksittäistä kohdetta koskevassa diagnostisessa päätöstilanteessa. Se kuvaa testituloksen t antaman suhteellisen näytön voimakkuutta diagnostisen hypoteesin D = "kohde on sairas" puolesta hypoteesia H = "kohde on terve" vastaan. Kun on saatu testitulos $T = t$, niin kohteen todellista tilaa koskevan **posterioritodennäköisyyden**

PV(t) = "todennäköisyys sille, että kohde on sairas, kun testin T tulos on t"

arvo määräytyy prioritodennäköisyydestä Pr ja uskottavuusosamäärästä LR(t) Bayesin kaavan yleistyksen nojalla.

Diagnostisten ja seulontatestien yksi tärkeä ominaisuus on myös **konsistenssi** eli **toistettavuus**. On toivottavaa, että samasta kohteesta tehdyt rinnakkaiset mittaukset antaisivat mahdollisimman samanlaiset eli konsistentit tulokset. Toistettavuus on huono, jos näiden toistojen välinen hajonta

olisi suuri. Tämä nimittäin vaikuttaisi negatiivisesti kaikkiin erottelukykyä kuvaaviin parametreihin, jos lopullinen testi perustuisi vain yhteen mittauskertaan. Näissä olosuhteissa toistoista saatavien tulosten sopivalla yhdistelmällä päästäisiin kohentamaan konsistenssia ja sitä kautta myös erottelukykyä, edellyttäen että useamman rinnakkaismittauksen suorittaminen ei käy liian työlääksi ja/tai kalliiksi. Kategorisen testimuuttujan konsistenssia mitataan kappakertoimen tapaisilla tunnusluvuilla. Jatkuvan muuttujan konsistenssia arvioidaan varianssianalyysin menetelmin erottamalla kohteiden väliset ja kohteiden sisäiset vaihtelukomponentit toisistaan. Kohteiden sisäisen varianssin suhde mittausten kokonaisvaihteluun antaa reliabiliteettikertoimena tunnetun tunnusluvun testimuuttujan suhteelliselle konsistenssille.

Tilastolliset menetelmät ja niiden käyttö TOXTEST-hankkeessa

Hankesuunnitelmassa (28.10.2010) kommentoidaan tutkittavien toksisuustestien tilastollisen arvioinnin toteutusta:

- "... yhteydet analysoidaan tilastollisin menetelmin",
- "Käyttökelpoisuuden kriteerinä on mm. se, kuinka hyvin mitattava parametri korreloi raportoitujen terveyshaittojen ja rakennusta koskevan tiedon kanssa."

Tästä saa sen vaikutelman, että tutkijoilla ei vielä siinä vaiheessa olisi ollut mitään täsmällistä käsitystä siitä, millä erityisillä tilastomenetelmillä näitä yhteyksiä ja korreloitumisia pitää analysoida.

Kiinnitämme huomiota myös siihen, että hankkeessa ei ole lainkaan ollut mukana ammattitaitoista tilastotieteilijää, joka olisi aktiivisesti osallistunut jo tietojenkeruun asetelman ja aineistonhankinnan suunnitteluun sen lisäksi, että hän olisi vastannut saatujen tulosten analyysin pätevyydestä. Ketään alan ammattilaista ei myöskään liene missään vaiheessa edes konsultoitu, vaikka tilastotieteellistä asiantuntemusta olisi tarvittaessa mitä todennäköisimmin löytynyt jokaisen hankeorganisaatiossa edustettuna olevan taustayhteisön piiristä.

Hankkeen kolmannen vaiheen tutkimusaineiston muodostamisessa noudatettiin sitä edellä mainittua periaatetta, jonka mukaan vauriokohteet valittiin "sairaista sairaimpien" joukosta ja vertailukohteen puolestaan "terveistä terveimpien" kohteiden populaatiosta. Tämä menettelytapa on hyvin perusteltu ottaen huomioon tutkittavien menetelmien arvioinnin vaihe. Vaurio- ja vertailukohteiden lukumäärä näyttää määräytyneen ensisijaisesti muista näkökohdista kuin esimerkiksi tilastollisen tarkkuuden vaatimuksista.

Tulosten analyysissä sovellettujen tilastollisten menetelmien dokumentaatio on käytännössä olematon loppuraportin metodisissa osissa. Tämä puute on huomattavassa kontrastissa mm. sen kanssa, että raportissa ja sen liitteissä on varsin seikkaperäisesti kuvattu näytteiden keruun ja käsittelyn laboratorio- ym. tekniset yksityiskohdat Joissakin tuloksia esittelevissä yksittäisissä taulukoissa (Liitteen 3 taulukot 1,2 ja 5) on lyhyt, joskaan ei yksikäsitteinen maininta mahdollisesti käytetyistä tilastollisista testeistä (jää epäselväksi käytettiinkö Fisherin tarkkaa vai Pearsonin khiin neliötestiä). Enimmäkseen "merkitsevyyksiä" kuitenkin raportoidaan ilman, että eksplisiittisesti mainitaan ao. testisuuretta saati että itse testattavan nollahypoteesin yleistä mielekkyyttä, tai että testin oletusten realistisuutta millään lailla kriittisesti arvioidaan (kuten esim. Liitteen 7 numeroimattomat kuviot ja Liitteen 12 teksti)

Tilastollinen analyysi näyttääkin paljolti painottuvan vauriokohteiden ja vertailukohteiden välisten erojen "tilastollisen merkitsevyyden" testaamiseen. Tässä suhteessa hankkeessa on toimittu noudattaen monilla tieteenaloilla vallitsevia konventionaalisia mutta tilastotieteelliseltä ja

sisällölliseltä kannalta perustelemattomia rituaaleja. Tilastollisen testaamisen ja P-arvojen raportoinnin mielekkyys ja informatiivisuus voidaan tässäkin yhteydessä asettaa yleisesti kyseenalaiseksi. Se on sitä erityisesti, kun testaamisen kohteena ovat keskimääräiset erot erilaisten oireiden esiintyvyydessä kuin myös esiintyvyyttä kuvaavissa summamuuttujissa vaurio- ja vertailukohteiden välillä (ks. Liitteen 3 taulukot 1, 2 ja 5, sekä Liitteen 7 numeroimaton kuvio), jos kerran nämä osajoukot alun perin muodostettiin sillä periaatteella, että ne olisivat mahdollisimman erilaisia oireiden esiintymisen suhteen!

Tilastollinen testaus ei ylipäätään anna hyödyllisiä ja käyttökelpoisia vastauksia kysymyksiin, jotka koskevat diagnostisen tai seulontatestin (joihin toksisuustestit ovat rinnastettavissa) erottelukykyä ja ennustearvoa. Näiltä osin raportissa ollaan enimmäkseen oikeilla linjoilla, koska "tilastolliseen merkitsevyyteen" vedotaan vain liitteessä 12, jossa verrataan kuolleiden bakteerien tiheyksien eroja vaurio- ja vertailukohteiden muodostamien ryhmien välillä. Tässä kohdassa valitun kuvaus- ja analyysitavan perusongelmana on, että siinä ei ole mitenkään otettu huomioon kohdemuuttujan jakauman merkittävää nollainflatoitumista: vertailuryhmässä suurin osa lukuarvoista oli nollia, ja nollien osuus oli huomattava myös vauriokohteiden joukossa. Mikään tavanomainen *t*-testi tai Mannin–Whitneyn testi ei ole pätevä – ei edes yksittäisten oudokkien poistamisen jälkeenkään – nollainflatoituneen aineiston analyysiin, vaan siihen tarvitaan toisenlaista menettelytapaa. Lukijalle jää myös epäselväksi, onko keskiarvon rinnalla operaattorin ± jälkeen annettu luku keskihajonta (SD) vai keskivirhe (SE), joskaan kumpikaan niistä ei olisi erityisen informatiivinen nollainflatoidussa aineistossa.

Raportin keskeiset tulokset ja niiden arviointi

Raportin keskeiset tulokset on tiivistetty taulukkoon 2. Siinä on periaatteessa asianmukaisesti raportoitu piste-estimaatit eri menetelmien näytekohtaisille sensitiivisyyksille ja spesifisyyksille nojautuen valittuihin kynnysarvoihin sekä yksinkertaistaviin oletuksiin eri näytteiden täydellisestä keskinäisestä riippumattomuudesta ja samanarvoisuudesta. Kuitenkaan siinä eikä muuallakaan ole dokumentoitu näille piste-estimaateille minkäänlaisia virhemarginaaleja. Lisäksi olisi ollut hyödyllistä raportoida esimerkiksi Youdenin indeksit ja niiden luottamusvälit antamaan vielä tiiviimpi ja havainnollisempi kuva tutkimustulosten antaman tilastollisen näytön vahvuudesta – tai lähinnä sen puuttumisesta.

On yhtäältä todettava, että kun kustakin kohteesta kerätään ja analysoidaan useampi kuin yksi näyte, niin aineisto on lähtökohtaisesti luonteeltaan hierarkkisesti ryvästynyttä siten, että saman kohteen eri näytteiden tulosten voidaan odottaa olevan jossain määrin keskenään korreloituneita mutta riippumattomia muiden kohteiden tuloksista. Tämän ryvästymisen huomioonottaminen johtaisi monimutkaisempaan analyysiin, jonka seurauksena todennäköisesti ainakin erottelukykyä kuvaavien parametrien estimaattien virhemarginaalit levenisivät verrattuna yksinkertaiseen estimointiin, jossa sisäkorrelaatiota ei oteta huomioon. Toisaalta yksinkertainen analyysi kohtelee saman kohteen eri näytteitä yhtäläisesti ja täysin symmetrisesti. Se ei ota huomioon mm. sitä mahdollisuutta, että näytteiden keruupisteet kohteen sisällä olisi valittu jotenkin harkinnanvaraisesti, eli esim. yksi niistä erityisen läheltä todennäköistä vauriokohtaa ja toinen selvästi kauempaa.

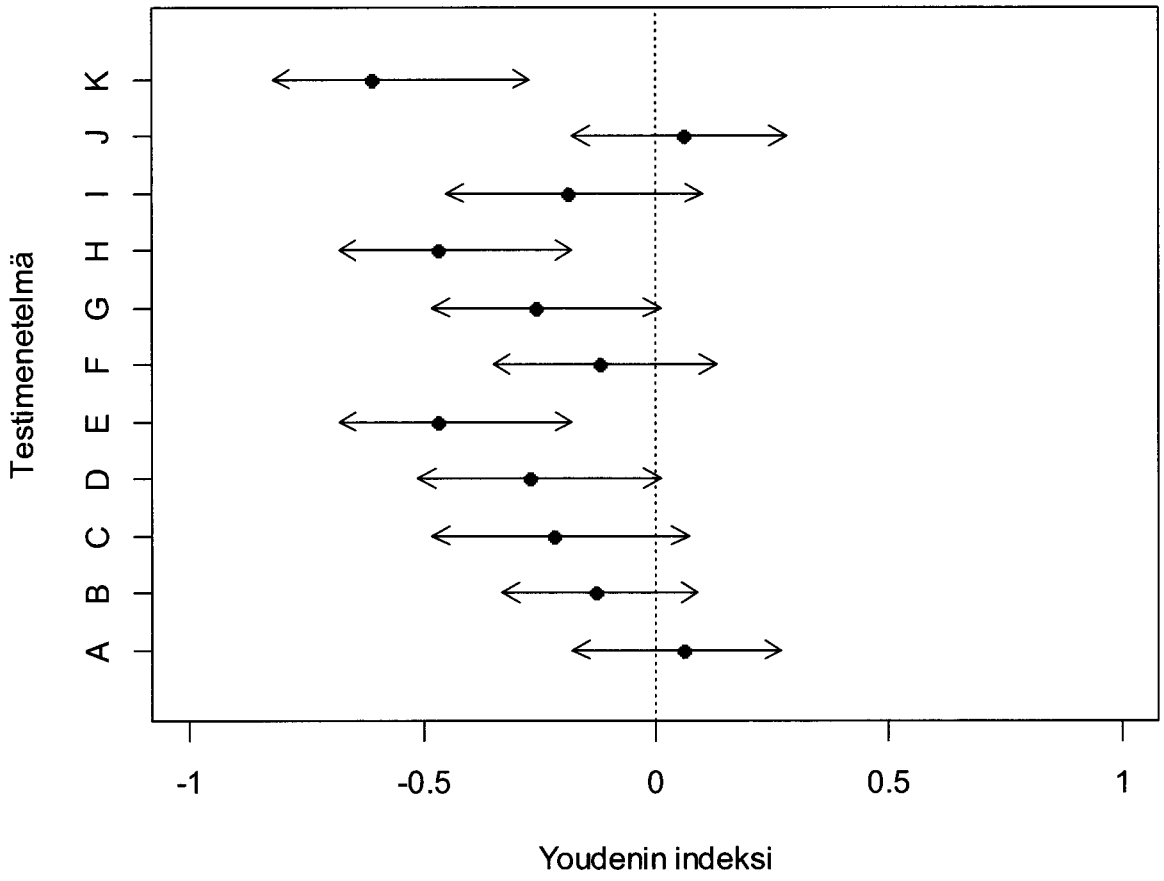
Kun tarkastellaan loppuraportin yhteenvedon taulukossa 2 esitettyjä päätuloksia täydentäen niitä eri menetelmien Youdenin indeksien piste-estimaateilla ja niiden 95% luottamusväleillä, jotka on laskettu edellä mainittuihin yksinkertaistaviin oletuksiin nojautuen (ks. Kuva 1), voidaan tehdä mm. seuraavia havaintoja:

(1) kaikkien laatikkonäytteistä tehtyjen testien estimoitu sensitiivisyys jää alle 50 % samoin kuin pyyhintänäytteiden *E.coli* -kuolleisuustestin,

(2) väärien positiivisten osuus on yli 50 % useiden menetelmien kohdalla,

(3) Youdenin indeksien piste-estimaatit ovat joko negatiiviset tai lähellä 0:aa,

(4) kaikkien em. estimaattien virhemarginaalit ovat leveähköjä.



Kuva 1. Raportin taulukon 2 sisältämien testimenetelmien Youdenin indeksien piste-estimaatit ja likimääräiset 95% luottamusvälit annettujen sensitiivisyys- ja spesifisyyslukujen pohjalta. Testimenetelmät on koodattu taulukon 2 mukaisessa järjestyksessä, jossa ensin ovat pyyhintänäytteisiin perustuvat A = sian siittiöiden liikkuvuus, B = *E.coli* -kuolleisuus, ja siitä eteenpäin laatikkonäytteisiin perustuvat: C = sian siittiöiden liikkuvuus, ... , K = Hiiren makrofagi, tulehdus (TNFα-tuotanto).

Kuva 1 osoittaa, että Taulukon 2 pohjalta lasketut Youdenin indeksien arvot ovat joko "väärällä kaistalla" tai "keskiviivalla", kun hyvien testimenetelmien kohdalla niiden pitäisi sijoittua tiukasti kuvion oikeaan laitaan.

Näiden tulosten nojalla raportin pääosassa esitetyt tulkinnat ja loppupäätelmät vaikuttavat erittäin perustelluilta sikäli kuin taulukon 2 luvut on laskettu asianmukaisista mittaussarjoista. Mitkään

hienojakoisemmat analyysit, jotka paremmin hyödyntäisivät käytettyjen mittausten kvantitatiivisen luonteen tai ottaisivat huomioon aineiston ryvästymisen ja siitä aiheutuvan sisäkorreloitumisen, eivät todennäköisesti johtaisi olennaisesti toisenlaisiin lopputuloksiin.

Virhemarginaalien leveys on luonnollinen seuraus aineiston pienuudesta. Yksittäisen luottamusvälin leveys on kääntäen verrannollinen aineiston koon neliöjuureen. Nyrkkisääntönä voidaan siis sanoa, että jos luottamusväli halutaan puolittaa, niin aineisto pitää nelinkertaistaa. Tällöin informatiivisempaa on lisätä toisistaan riippumattomien vaurio- ja vertailukohteiden lukumäärää kuin kasvattaa keskenään korreloivien kohdekohtaisten näytteiden määrää.

Koska Helsingin yliopiston sekä Turun yliopiston edustajat ovat ilmaisseet raportin päätulosten ja johtopäätösten osalta joiltakin osin eriäviä mielipiteitä, on syytä tarkastella yksityiskohtaisemmin heidän laatimiaan liitteitä raportissa.

Helsingin yliopiston osuus – Liite 11

Hankeorganisaatiossa Helsingin yliopistoa edustavien tutkijoiden ryhmän vastuulla oli sian siittiöiden liikkuvuuteen ja sytotoksisuuteen perustuvien testimenetelmien ominaisuuksien arviointi. Ryhmä dokumentoi Liitteessä 11 oman osuutensa toteutusta, siinä käytettyjä laboratoriotekniikoita, mittausarvojen tulkinnan periaatteita, tilastollisen analyysin menetelmiä, analyysien tuloksia ja lopuksi omia johtopäätöksiään. Päähuomio suuntautuu pyyhkäisynäytteistä mitattuihin EC50-lukuihin, jotka kuvaavat siittiöiden liikkuvuuden heikentymistä, sekä tilastolliseen yhteyteen, joka EC50-arvoilla on paitsi kohteiden vauriostatuksen niin myös niiden asukkaiden keskimääräisten oirepistemäärien kanssa.

Tämän liitteen sisällössä meitä askarruttavat erityisesti seuraavat kaksi seikkaa HY:n ryhmän omaksumassa metodisessa paradigmassa: (1) aineiston valikointi ja usean kohteen poissulkeminen analyyseistä, ja (2) EC50-lukujen ja kohteista saatujen oirepistemäärien välistä tilastollista yhteyttä kuvaavien tunnuslukujen valinta ja niiden pohjalta tehtyjen päätelmien osuvuus. Näiden lisäksi kiinnitämme huomiota liitteen eräisiin sisäisiin epäjohdonmukaisuuksiin ja esityksen huomattavaan viimeistelemättömyyteen.

Aineiston valikointi

HY:n ryhmä päätyi valikoimaan analyyseihinsä vain sellaisia kohteita, joiden kaikille kolmelle pyyhkäisypölynäytteelle onnistuttiin tekemään tuloksellinen toksisuusmittaus. Siten kaikkiaan 10 vauriotalosta, joista oli saatavissa ainakin yhden näytteen tulokset, analyysiin kelpuutettiin vain puolet eli niiden 5 talon aineisto (n:ot 14, 22, 27, 46, 48), joista kaikkien kolmen näytteen tulokset saatiin. Todettakoon, että analyysien ulkopuolelle jäi vielä 2 vauriotaloa (n:ot 19 ja 41), joiden yhdestäkään näytteestä ei kyetty saamaan onnistunutta toksisuusmittausta. Lisäksi, periaatteessa käytettävissä olleista 11 verrokkikohteesta suljettiin analyysien ulkopuolelle jostain julkilausumattomasta syystä kohteet 5 ja 35, vaikka myös niistä löytyy dokumentoituna ao. testimenetelmään perustuva EC50-arvo kolmesta eri näytteestä (ks. saamamme Excel-tiedoston välilehti "Pyyhintänäytteet, lukuarvot") .

Tekemälleen valikoinnille ei HY:n ryhmä anna liitteen tekstissä nähdäksemme mitään pätevää perustelua. Koko raportissa kauttaaltaan (ml. liitteet 11 ja 12) on eri testimenetelmien erottelukykyä kuvaavana parametrina ensisijaisesti käytetty yksittäisen näytteen toksisuumittauksen sensitiivisyyttä ja spesifisyyttä. Tässä hengessä myös HY:n ryhmä esittää omia tuloksiaan sian

siittiöiden EC50-lukuihin perustuvan testin osalta Taulukossa 2-11. Näytekohtaisen sensitiivisyyden ja spesifisyyden estimoinnissa periaatteessa yhtä informatiivisia ovat nimittäin kaikki pätevät mittaustulokset myös niistä kohteista, joissa toksisuustestin tuloksia on saatavilla vain 1 tai 2 näytteestä.

Aineiston valikoinnilla ja sinänsä kelpoisten havaintojen poissulkemisella on ainakin seuraavat ei-toivotut seuraukset: (a) estimaattien virhemarginaalien leveneminen, ja (b) mahdollinen harha estimoinnissa, jos mittaustulosten puuttuminen ja poissulkeminen korreloivat kohteen statuksen ja/tai potentiaalisen mutta saamatta jääneen mittaustuloksen kanssa. Jälkimmäisen huolen herääminen on tässä yhteydessä luonnollista, ottaen huomioon näytteitä kohdanneessa kadossa havaitun suuren kontrastin vaurio- ja vertailukohteiden välillä. Toisaalta, jos toksisuustestillä on sellainen ominaisuus, että terveissä kohteissa se enimmäkseen onnistuu, mutta yli puolella vauriokohteista (tässä aineistossa 7/12) näytteiden kerääminen ei tuota hyödynnettävissä olevia mittausarvoja, niin menetelmän yleinen kenttäkelpoisuus ei vaikuta kovin hyvältä.

Liitteen 11 taulukossa 2-11 raportoidaan HY:n tutkijoiden laskemat piste-estimaatit (mutta ei virhemarginaaleja) valikoimastaan 14 kohteen aineistosta sekä siittiötestin että TY:n *E.coli*-lux-testin yksittäisen näytteen sensitiivisyydelle (Se) ja väärien positiivisten osuudelle (FPR) kuin myös niiden suhteelle eli positiivisen tuloksen uskottavuusosamäärälle (LR+, ks. edellä), kun positiivisen tuloksen eli toksisuuden kriteerinä on EC50 < 40 µg/ml. Kun tarkistimme laskelmat nojautuen EC50-lukuarvoihin, jotka on dokumentoitu em. Excel-taulukon välilehden "Pyyhintänäytteet, lukuarvot" ao. sarakkeissa, niin saimme siittiötestin osalta hieman erilaiset toksisen pölyn esiintyvyysluvut vaurio- ja vertailukohteissa sekä HY:n ryhmän valikoimasta aineistosta että kaikkien kelvollisten näytteiden aineistosta [suluissa 95% luottamusväli]:

parametri	valikoitu aineisto	kaikki kelvolliset näytteet
Se	10/15 = 0.67 [0.41, 0.85]	14/24 = 0.58 [0.39, 0.75]
FPR	6/27 = 0.22 [0.10, 0.41]	8/33 = 0.24 [0.13, 0.41]
J	0.44 [0.13, 0.68]	0.34 [0.08, 0.56]
LR+	3.0 [1.5, 6.6]	2.4 [1.3, 4.8]

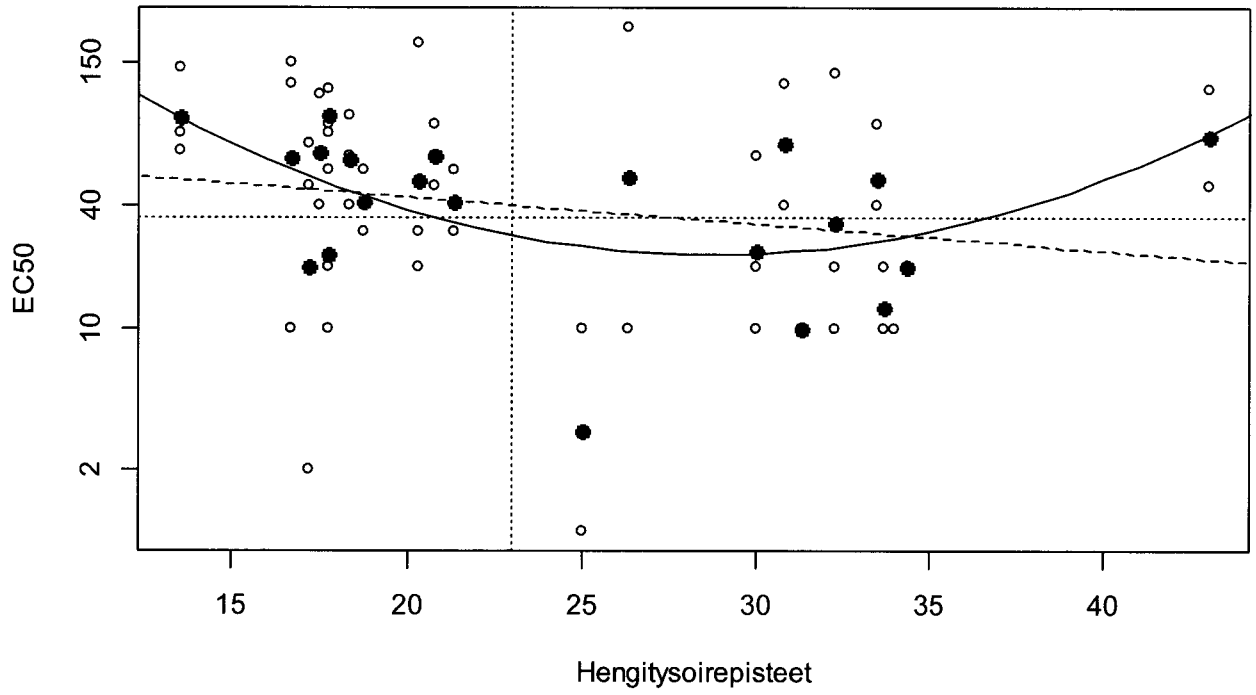
Valikoidusta aineistosta saadut piste-estimaatit näyttävät siis antavan hieman paremman kuvan siittiötestin erottelukyvystä kuin valikoimattomasta aineistosta lasketut, joskin havaittavat erot aineistojen välillä mahtuvat hyvin tavanomaisen virhemarginaalin sisään. Joka tapauksessa, vaikka valikointi ei sisältäisikään mitään systemaattista virhettä, niin valikoidunkaan aineiston antama näyttö siittiötestin erottelukyvystä ei näiden lukujen valossa ole lähimainkaan niin hyvä, kuin haluttuun tarkoitukseen käyttökelpoiselta menetelmältä voidaan odottaa.

Todettakoon tässä yhteydessä havaitsemamme epäjohdonmukaisuus koskien tulososassa (s. 4) viitattua Taulukkoa 1/11. Sen voisi asiayhteydestä päätellen odottaa sisältävän valikoitujen 5 vauriokohteen ja 9 vertailukohteen toksisuusmittausten tulokset. Kuitenkin seuraavalla sivulla oleva Taulukko 11-1 (ei siis nimeltään Taulukko 1/11!) kattaa vain osan näistä 14 kohteesta. Siitä erityisesti puuttuu vauriokohde 22 eikä siinä myöskään ole vertailukohteita 5, 9, 17, 21, 29, 34 ja 35. Toisaalta tässä taulukossa on mittaustuloksia myös kohteista 7, 15, 19, 32, 37, 41, 43, jotka eivät sisältyneet em. 14 kohteen valikoituun osajoukkoon. Lukijalle jää epäselväksi, mitä tarkoitusta tämä taulukko palvelee.

Toksisuustestin tulosten ja oirepistemäärien tilastollisen yhteyden tarkastelu

Taulukossa 2-11 on edellä mainittujen tulosten lisäksi raportoitu oirekyselyistä saatujen kohdekohtaisten keskimääräisten oiresummapistemäärien (per asukas) ryhmittäiset keskiarvot erikseen vauriokohteissa ja vertailukohteissa sekä näiden ryhmäkeskiarvojen osamäärät vaurio- ja vertailuryhmien välillä. Näiden keskiarvojen suhteellisella vertailulla voi parhaimmillaan olla jonkin verran lisäarvoa aineiston kuvailevassa analyysissä. HY:n tutkijat kuitenkin käyttävät tätä suhdelukua ja sen samanlaisuutta estimoidun LR+:n kanssa vahvana argumenttina oman testimenetelmänsä hyvyyden puolesta. Tällaista menetelmää emme kuitenkaan ole tähän mennessä nähneet esiteltävän missään tuntemassamme diagnostisten testien arvioinnin tilastollisia menetelmiä käsittelevässä kirjallisessa lähteessä. Oiresummien keskiarvojen ryhmien välisten osamäärien ja positiivisten uskottavuusosamäärien keskinäinen vertailu ei nimittäin anna minkäänlaista relevanttia ja käyttökelpoista kvantitatiivista arviota tutkittavien toksisuustestien erottelukyvystä. Tämä johtuu siitä, että sekä vaurio- että vertailukohteiden osapopulaatioissa on merkittävässä määrin ryhmien sisäistä vaihtelua niin oirepistemäärissä kuin toksisuustestin tuloksissa, jota ryhmäkohtaiset suhdeluvut eivät millään tavoin ota huomioon.

Osuvampi analyysi on sellainen, jossa kohdekohtaisten oirekeskiarvojen tilastollista yhteyttä tarkastellaan kohde- ja jopa näytekohtaisten testitulosten kanssa. Kuvassa 2 on tämän ajatuksen mukainen sirontakuvio, jonka x-akseli edustaa hengitysoireiden summapistemäärien keskiarvoja kohteittain ja y-akseli näytekohtaisia EC50-arvoja. Kustakin kohteesta on siis 1-3 EC50-havaintoa, jotka on merkitty pienellä avopallolla, ja kohdekohtaisten lukujen geometriset keskiarvot isommalla mustalla pallolla.



Kuva 2. Siittiötestien EC50-arvot kohdekohtaisten keskimääräisten hengitysoirepistemäärien kanssa. Kustakin kohteesta on 1-3 EC50-arvoa. Esimerkiksi kohteen 27 havaintopisteiden x-koordinaatti on 25, kohteen 22 x-koordinaatti on 33.7 ja kohteen 43 x-koordinaatti on 43. Avopallot kuvaavat näytekohtaisia tuloksia ja mustat pallot ovat kohdekohtaisten EC50-arvojen geometristen keskiarvojen kohdalla.

Tämän sirontakuvio antaa monenlaista kiinnostavaa informaatiota. Ensinnäkin, EC50-mittauksen konsistenssista saa alustavan kuvan tarkastelemalla arvojen vaihtelua y-akselin suunnassa niiden x-akselilla erotettavien arvojen kohdalla, jotka kukin liittyvät vain yhteen kohteeseen ja joista on yhtä useampi mittaustulos. Todettakoon, että menetelmän konsistenssin arviointi vaatisi oman erillisen analyysinsä, mutta mitään dokumentaatiota tällaisesta analyysistä ei löydy liitteestä 11. Toiseksi, oirepistemäärien ja EC50-lukujen välisen tilastollinen riippuvuus ei näytä kovin vahvalta, mitä kuvastaa EC50-lukujen suurehko hajonta (pieni konsistenssi?) kullakin oirepistemäärän tasolla. Kolmanneksi, tilastollisen yhteyden systemaattinen muoto vaikuttaa myös epäjohdonmukaiselta. Lineaarista monotonista yhteyttä (katkoviivan mukainen regressiosuora) paremmin aineiston kanssa sopusoinnussa on 2. asteen regressiokäyrä (yhtenäinen viiva). Kiinnitämme erityisesti huomiota siihen, että (a) kohteella 27, josta saatiin kaikkein pienimmät EC50-arvot, asukkaiden keskimääräinen hengitysoirepistemäärä oli 25, joka oli vauriokohteista kaikkein pienin, ja (b) kohteen 43, jonka asukkaiden oirepistemäärä 43 oli kaikkein korkein, onnistuneet kaksi toksisuusmittausta tuottivat EC50-arvot, jotka eivät mitenkään eronneet vertailutalojen tyypillisistä EC50-arvoista. Pelkästään tämä sirontakuvio osoittaa, kuinka heikosti siittiötestin tulos auttaa luokittelemaan kohteita sen mukaan, millaisia terveyshaittoja asukkaat kokevat.

Näytekohtaisten tulosten yhdistely ja kvantitatiivisuuden huomioon ottaminen

HY:n ryhmä on mielestämme sikäli oikeilla jäljillä, että jos yksittäisen näytteen toksisuudella on loogisen suuntainen tilastollinen yhteys siihen, onko kohde vaurioitunut vai ei, niin kolme toksisuuden kriteerin täyttävää näytettä samasta kohteesta antaa vahvemman näytön vauriosta kuin kaksi tai yksi. Tämän ajatuksen voi pukea kvantitatiivisempaan muotoon esimerkiksi seuraavasti. Merkitään symbolilla T positiivisten näytteiden lukumäärää yksittäisessä kohteessa, jolloin T voi saada arvot 0, 1, 2 tai 3, edellyttäen että kaikkien kolmen näytteen toksisuus on määritettävissä. Näytekohtaisten Se- ja FPR-estimaattien pohjalta voidaan nyt laskea binomijakaumaan (indeksiparametrilla $N = 3$) nojautuen muuttujan T mahdollisten arvojen $t = 0, ..., 3$ estimoidut todennäköisyydet $p(t \mid D)$ ja $p(t \mid H)$ erikseen vauriotalojen ja vertailutalojen osapopulaatioissa D ja H. Näistä saadaan estimoiduksi uskottavuusosamäärä $\mathrm{LR}(t) = p(t \mid D) / p(t \mid H)$ kullekin mahdolliselle positiivisten näytteiden lukumäärälle t. Seuraavassa taulukossa on näiden parametrien piste-estimaatit ja LR:ien likimääräiset 95% luottamusvälit (CI), jotka on johdettu edelleenkin erittäin yksinkertaistaviin riippumattomuusoletuksiin nojautuvalla mallilla.

t	0	1	2	3
p(t|D)	0.075	0.31	0.42	0.195
p(t|H)	0.42	0.42	0.144	0.016
LR(t)	0.18	0.75	2.8	11.8
95% CI	[0.03,0.68]	[0.33,1.2]	[1.3,9.5]	[1.8,105]

Vaikka kolmen positiivisen näytteen uskottavuusosamäärä LR(3) = 11.8 näyttääkin vaikuttavalta verrattuna yksittäisen näytteen positiivisesta uskottavuusosamäärästä laskemaamme estimaattiin LR+ = 2.4, niin sitä ei voi sellaisenaankaan pitää erityisen suurena LR-arvona. Se ei myöskään ole tämän menetelmän yleisen erottelukyvyn kannalta ratkaiseva tilanteessa, jossa vauriokohteistakin vain 20%:lla kaikki kolme näytettä olisivat valitun kynnysarvon perusteella positiivisia, ja 40%:lla vähintään kaksi näytettä kolmesta olisi negatiivisia, kun taas vertailuryhmän taloista lähes 60%:lla ainakin yksi näyte olisi positiivinen. Huomattakoon lisäksi LR(3):n kuin kaikkien muidenkin LR-lukujen leveät luottamusvälit, joten myös niitä koskeva tilastollinen epävarmuus on varsin suuri.

Edellä tehtyyn analyysiin ja siitä saatuihin estimaatteihin on syytä suhtautua muutenkin varauksella. Siinähän lähtökohtaisena oletuksena on, että kustakin kohteesta hankitaan kolme rinnakkaista näytettä, joista kaikista saadaan onnistunut mittaustulos. Jos kuitenkin kato on niin suurta kuin tässä

aineistossa näytti olevan, eli EC50-arvo saadaan mitatuksi kaikista kolmesta näytteestä vain alle puolella vauriokohteista, niin em. laskelmien relevanssi on varsin viitteellinen. On tosin periaatteessa mahdollista rikastaa käytettyä tilastollista mallikehikkoa siten, että estimoidaan positiivisten testitulosten lukumäärän *T* eri arvojen *t* todennäköisyydet ja LR:t myös sellaisissa skenaarioissa, joissa mitatuksi saadaan vain yksi tai kaksi näytettä.

Siittiötestin kokonaisvaltaista erottelukykyä voidaan yksittäiseen kynnysarvoon sitoutumisen asemesta arvioida ROC-käyrän ja sen pinta-alan AUC avulla hyödyntäen saatuja kvantitatiivisia EC50-arvoja. Yksittäisen testin EC50-luvuista johdetun ROC-käyrän estimoiduksi pinta-alaksi saimme 0.67 [95% CI 0.51, 0.81]. Kun yhdistelimme kohdekohtaiset testitulokset laskemalla niiden geometriset keskiarvot, saimme tämän menetelmän ROC-käyrän pinta-alan estimaatiksi myös 0.67 [0.43, 0.89]. Olisi voinut odottaa geometristen keskiarvojen käytön tuottavan yksittäiseen näytteeseen verrattuna jonkin verran paremman AUC-arvon. Se, ettei näin nyt käynyt, voi osaltaan johtua siitä, että kadon takia vauriokohteiden joukossa iso osa keskiarvoista saattoi perustua vain yhteen tai kahteen mittaukseen.

Mitkään näihin siittiötestin kvantitatiivisempiin analyyseihin perustuvat tulokset eivät näkemyksemme mukaan ole vieläkään niin hyviä ja lupaavia, kuin tarkoitukseen soveltuvan menetelmän erottelukyvyltä on perusteltua odottaa, ottaen huomioon sen lähtökohdan, että ne ovat peräisin "sairaista sairaimpien" vauriokohteiden ja "terveistä terveimpien" vertailukohteiden muodostamasta aineistosta.

Liitteen sivulla 6 olevan tekstin viimeisen kappaleen pohdinta ei ole mielestämme kaikilta osin kovin vakuuttavaa. On totta, että jos näytteitä kerättäisiin kolmen asemesta 10-20 kpl per kohde, niin asianmukaisella näytekohtaisten tulosten yhdistelmällä voitaisiin tietenkin teoriassa saavuttaa huomattavasti parempi kvantitatiivinen *mittaustarkkuus* kuin yhdellä mittauksella, minkä seurauksena yhdistelmästrategian erottelukyvyn voisi odottaa olevan jonkin verran paremman kuin 1 tai 3 näytteeseen perustuvan menetelmän. Jos taas "tilastollisella varmuudella" tässä yhteydessä tarkoitetaan testimenetelmän erottelukyvyn populaatiotasoisen arvioinnin *tilastollista tarkkuutta*, niin tätä tarkkuuden lajia ei voi kovin paljoa parantaa näytemäärän moninkertaistamisella, josta suuri osa olisi itse asiassa luonteeltaan pseudoreplikaatiota. Tilastollisen tarkkuuden kannalta informatiivisempi vaihtoehtoinen tutkimusstrategia olisi kasvattaa arviointiaineistoa lisäämällä mieluummin toisistaan riippumattomien vaurio- ja vertailukohteiden lukumäärää mutta pitämällä kohdekohtaiset näytemäärät pieninä. – Herää myös kysymys, kuinka käyttökelpoinen ja kustannustehokas sellainen testimenetelmä ylipäätään voi olla, joka vaatisi yli 10 näytteen keräämiseen kustakin kohteesta.

Turun yliopiston osuus – Liite 12

Hankeorganisaatiossa Turun yliopistoa edustavien tutkijoiden ryhmä esittää Liitteessä 12 tuloksiaan analyyseistään, jotka kohdistuvat pyyhkäisy- ja laatikkonäytteille tehtyjen *E.coli* -lux-testien korjattuihin mittaustuloksiin, jotka esitetään liitteen Taulukoissa 1 ja 2 samoin kuin saamamme Excel-taulukon ao. välilehdillä. TY:n ryhmä ei ole toteuttanut samanlaista aineiston valikointia kuin HY:n ryhmä vaan on sisällyttänyt analyyseihinsä kaikki kelvolliset mittaustulokset riippumatta siitä, kuinka monta sellaista onnistuttiin saamaan kustakin kohteesta.

Tällä aineistolla saatiin kummallekin menetelmälle jonkin verran parempaa erottelukykyä kuvaavat tulokset verrattuna raportin yhteenveto-osan taulukossa 2 esitettyihin lukuihin. Pyyhintänäytteistä estimoitu *E.coli* -menetelmän sensitiivisyys oli Se = 0.69 [95% CI 50, 84], väärien positiivisten

osuus FPR = 0.27 [0.15, 0.44] ja Youdenin indeksi $J = 0.42$ [0.16, 0.62]. Laatikkonäytteistä vastaavat estimaatit olivat Se = 0.58 [0.39, 0.75], FPR = 0.35 [0.18, 0.57], ja $J = 0.23$ [−0.06, 0.49]. Tekstissä viitataan myös tehtyyn ROC-analyysiin, jonka numeerisia tuloksia ei kuitenkaan dokumentoida. Trapetsimenetelmää käyttäen saimme yksittäisen pyyhintänäytteen kvantitatiiviseen tulokseen perustuvan ROC-käyrän alle jäävän pinta-alan estimaatiksi 0.70 [0.52, 0.88].

Myöskään nämä tulokset eivät näkemyksemme mukaan ole edes pyyhintänäytteiden osalta vieläkään niin hyviä ja lupaavia, kuin tarkoitukseen soveltuvan menetelmän erottelukyvyltä on perusteltua odottaa, ottaen huomioon arviointiaineiston luonteen: "sairaista sairaimmat" vauriokohteet ja "terveistä terveimmät" vertailukohteet. *E.coli* -pyyhintänäytteiden kvantitatiivista analyysiä voitaisiin periaatteessa laajentaa samalla tavoin kuin edellä tehtiin siittiötestille. Koska yksinkertaisemmista analyyseistä lasketut tunnusluvut eivät näytä olevan *E.coli* -menetelmälle olennaisesti suotuisempia kuin siittiötestille, niin tällaisista analyyseistä saatavat tulokset tuskin muuttaisivat lopullista arviota *E.coli* -menetelmän käyttökelpoisuudesta.

Arvioitsijoiden johtopäätökset

Toxtest-hankkeen tavoitteena oli "kehittää sisäympäristönäytteille soveltuva toksisuuden arviointimenetelmä, jota voidaan käyttää terveysvalvonnan tukena homevaurion vakavuuden arvioinnissa ja kosteusvauriokohteiden korjauksen priorisoinnissa". Loppuraportin tiivistelmän johtopäätös on yksiselitteinen: **"huonepölyuutoksen toksisuutta ei voida käyttää kosteusvauriokohteiden luokittelussa tai terveyshaitan arvioinnissa"**. Toisaalla raportissa (liitteet 11 ja 12) päädytään paljon positiivisempiin päätelmiin. Arviointitehtävämme mukaan olemme perehtyneet hankkeen tavoitteisiin, menetelmiin, tulosten käsittelyyn ja tulkintaan ja myös liitteissä esitettyihin osin ristiriitaisiin johtopäätöksiin. Esitämme tässä arviointiraportissa myös joitakin perustavanlaatuisia näkökohtia hankkeen lähtökohdista, menetelmistä ja aivan erityisesti tulosten tilastolliseen käsittelyyn liittyvistä tavoista. **Arviointimme mukaan loppuraportin tiivistelmän johtopäätös on oikea, eikä ole mitään tieteellisiä perusteita lieventää tai muuttaa tätä johtopäätöstä.**

Oulussa, 31.5.2013

Esa Läärä, VTL
Tilastotieteen, erit. biometrian professori
Matemaattisten tieteiden laitos
Oulun yliopisto

Olavi Pelkonen, LKT, dosentti
Farmakologian professori (emer)
Farmakologian ja toksikologian yksikkö
Oulun yliopisto