Tampereen yliopisto

Mari Ahola

ALTISTUKSEN VAIKUTUKSEN TUTKIMINEN FOUR-WAY DECOMPOSITION -MENETELMÄLLÄ

Tilastollisen data-analyysin kandidaatintutkielma
Informaatioteknologian ja viestinnän tiedekunta
Huhtikuu 2021

TIIVISTELMÄ

Mari Ahola: Altistuksen vaikutuksen tutkiminen four-way decomposition -menetelmällä
Tilastollisen data-analyysin kandidaatintutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Huhtikuu 2021

Four-way decomposition on hajotusmenetelmä, jolla tutkitaan altistuksen suoraa ja epäsuoraa vaikutusta lopputulokseen. Altistuksen epäsuora vaikutus johtuu välittäjästä, joka vaikuttaa lopputulokseen yksin ja yhdessä altistuksen kanssa. Lopputulos voi täten syntyä altistuksen, välittäjän tai niiden vuorovaikutuksen kautta. Tässä tutkielmassa käydään läpi, miten altistuksen vaikutusta voidaan tarkastella four-way decomposition -hajotusmenetelmän avulla.

Menetelmän ideana on jakaa altistuksen kokonaisvaikutus neljään komponenttiin. Ensimmäinen komponentti on altistuksen suora, kontrolloitu vaikutus lopputulokseen, kun välittäjä ei ole läsnä. Toinen komponentti on viitteellinen vuorovaikutus, missä altistus ja välittäjä ovat läsnä vain niiden välisen vuorovaikutuksen kautta. Kolmas komponentti, välitetty vuorovaikutus, vaatii altistuksen ja välittäjän sekä niiden vuorovaikutuksen läsnäolon. Neljäs komponentti on puhtaasti epäsuora vaikutus, eli välittäjän vaikutus ilman altistusta. Näiden komponenttien muodostamista käsitellään tutkielmassa laajasti.

Tutkielmassa perehdytään myös four-way decomposition -menetelmän ja tilastollisten mallien yhteyteen. Koska komponenttien tarkkoja arvoja on mahdotonta sanoa, niiden osuuksia voidaan estimoida komponenttien odotusarvojen avulla. Komponenttien odotusarvoja on mahdollista tarkastella suhdeasteikollisten muuttujien tilanteessa regressiomallien avulla. Regressiomallit ja tilastolliset ohjelmistot auttavat estimoimaan välittäjän, altistuksen, välittäjän ja altistuksen vuorovaikutuksen sekä niistä riippumattoman osuuden lopputuloksesta. Lasketuista estimaateista voidaan tarkastella muuttujien välistä kausaalisuhdetta. Komponenttien estimointia ja tulkitsemista havainnollistetaan tutkielmassa esimerkin avulla. Havaintoesimerkissä ohjeistetaan myös, kuinka four-way decomposition voidaan toteuttaa R-ohjelmiston avulla.

Avainsanat: hajotelma, hajotusmenetelmä, välittäjä, välitetty vuorovaikutus

SISÄLLYSLUETTELO

LYHENTEET JA MERKINNÄT

CDE	Altistuksen kontrolloitu suora vaikutus (engl. controlled direct effect)
INT_{med}	Välitetty vuorovaikutus (engl. mediated interaction)
INT_{ref}	Viitteellinen vuorovaikutus (engl. reference interaction)
PDE	Altistuksen puhtaasti suora vaikutus (engl. pure direct effect)
PIE	Altistuksen puhtaasti epäsuora vaikutus (engl. pure indirect effect)
TE	Altistuksen kokonaisvaikutus (total effect)
TIE	Altistuksen epäsuora kokonaisvaikutus (engl. total indirect effect)

1. JOHDANTO

Viime vuosikymmenien ajan on kehitetty useita menetelmiä altistusten vaikutuksien tutkimiseksi, sillä erilaiset altistukset vaikuttavat elämäämme jatkuvasti. Altistusten vaikutus on laaja: geneettiset ominaisuudet voivat altistaa sairauksille, lääkkeet altistavat sivuoireille ja sosiaalinen ympäristö altistaa erilaisille käyttäytymismalleille. Koska altistusten vaikutus on niin suuri, tutkijat kokevat erityisen tärkeäksi löytää tehokkaita keinoja tarkastella vaikutuksia mahdollisimman kattavasti. Altistusten ja lopputulosten välisten kausaalisuhteiden tutkiminen voi vaikuttaa yksinkertaiselta, mutta altistus ei usein vaikuta lopputulokseen yksinään.

Lopputuloksen voidaan nähdä rakentuvan useiden muuttujien vaikutuksesta. Kovariaatit ovat kontrolloitavia muuttujia, jotka selittävät lopputulosta. Kovariaatit vaikuttavat aktiivisesti altistukseen, ja altistus vaikuttaa lopputulokseen. Näin syntyy altistuksen suora vaikutus. Altistukseen voi kuitenkin vaikuttaa jokin toinen altistus, joka toimii välittäjänä. Altistuksen ja välittäjän vuorovaikutus sekä välittäjän itsenäinen vaikutus muodostavat yhdessä altistuksen epäsuoran vaikutuksen lopputulokseen. Tätä puolta kausaalisuhteissa ei usein huomioida. On kuitenkin syytä tarkastella altistuksen epäsuoraa vaikutusta, jos haluaa selkeän kuvan lopputulokseen vaikuttavien muuttujien osuuksista.

Lopputuloksen voidaan osoittaa muodostuvan altistuksen, välittäjän sekä altistuksen ja välittäjän vuorovaikutuksen pohjalta. Four-way decomposition on VanderWeelen (2014a, 2014b) kehittämä hajotusmenetelmä, jonka avulla on mahdollista tarkastella altistuksen suoraa ja epäsuoraa vaikutusta lopputulokseen. Hajotusmenetelmän perusajatuksena on se, että välittäjämuuttujan läsnä ollessa altistuksen vaikutus lopputulokseen voidaan jakaa neljään komponenttiin. Komponentit kuvaavat välittäjän, altistuksen ja niiden vuorovaikutuksen osuutta lopputulokseen. Tässä tutkielmassa käydään läpi four-way decomposition -hajotusmenetelmän perusteita ja tilastollisten mallien yhteyttä komponenttien muodostamiseen. Työn lopussa havainnollistetaan esimerkin avulla komponenttien tulkintaa ja R-ohjelmiston käyttöä komponenttien estimoimiseen.

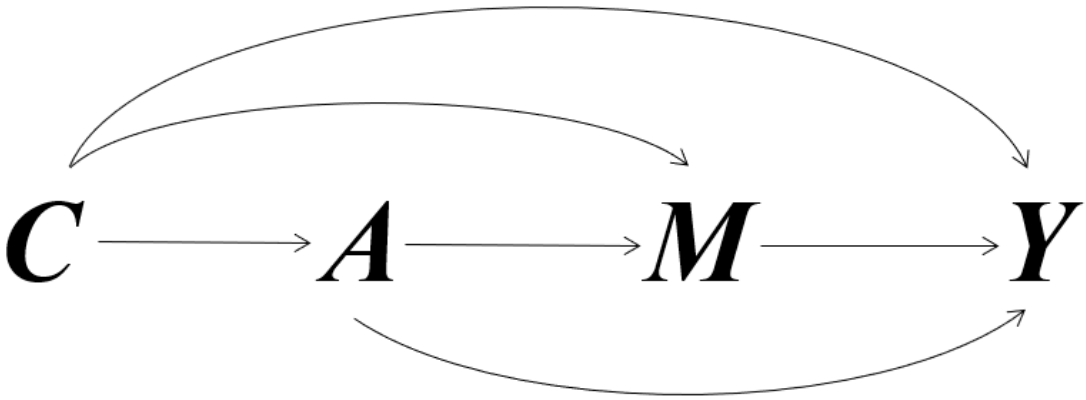
2. FOUR-WAY DECOMPOSITION -MENETELMÄN MÄÄRITELMÄ

2.1 Hajotelman muodostaminen

Four-way decomposition (VanderWeele, 2014b) on hajotusmenetelmä, jonka avulla voidaan tarkastella jonkin altistuksen suoraa ja epäsuoraa vaikutusta lopputulokseen. Menetelmän avulla voidaan estimoida, kuinka suuri osuus lopputuloksesta on välittäjän, altistuksen tai välittäjän ja altistuksen interaktion vaikutusta. Menetelmän lähtökohtana on lopputuloksen jakaminen neljään komponenttiin (VanderWeele, 2014a): (i) altistuksen vaikutus lopputulokseen ilman välittäjää, (ii) välittäjän ja altistuksen vuorovaikutus, kun altistusta ei tapahdu, (iii) välittäjän ja altistuksen vaikutus lopputulokseen, kun välittäjä ja altistus ovat läsnä, sekä (iv) välittäjän vaikutus lopputulokseen ilman altistusta.

Lopputuloksen Y perustana on joukko C perustason kovariaatteja. Kovariaatit voivat yksinään muodostaa lopputuloksen, mutta ne voivat olla myös yhteydessä altistukseen A, välittäjään M tai altistuksen ja välittäjän vuorovaikutukseen. Välitettyä vuorovaikutusta tapahtuu, kun sekä altistus että välittäjä vaikuttavat lopputulokseen Y. Kuten kuvasta (2.1) voidaan nähdä, lopputulos voi muodostua pelkkien kovariaattien vaikutuksesta, kovariaattien ja altistuksen vaikutuksesta, kovariaattien, altistuksen ja välittäjän vaikutuksesta tai kovariaattien ja välittäjän vaikutuksesta.

Altistus A voi vaikuttaa lopputulokseen sellaisia reittejä pitkin, mitkä eivät vaadi lainkaan välittäjämuuttujaa. Tällöin kyseessä on kontrolloitu suora vaikutus, CDE (*controlled direct effect*). Kontrolloidussa vaikutuksessa välittäjä M tai välittäjän ja altistuksen vuorovaikutus eivät ole läsnä. Vastaavasti tilanteessa, jossa altistus A ja välittäjä M ovat läsnä vain niiden vuorovaikutuksen kautta, ei huomioida altistusta tai välittäjää itsenäisinä vaikuttajina. Tätä kutsutaan viitteelliseksi vuorovaikutukseksi, INT_{ref} (*reference interaction*). Kun altistus A ja välittäjä M ovat molemmat läsnä ja niiden välillä tapahtuu vuorovaikutusta, kyseessä on välitetty vuorovaikutus, INT_{med} (*mediated interaction*). Jos altistus on läsnä vaikuttamatta kuitenkaan lopputulokseen ja välittäjä M on aktiivisesti läsnä, tapahtuu puhtaasti epäsuoraa vaikutusta, PIE (*pure indirect effect*).



Kuva 2.1. *Muuttujien vaikutus lopputuloksen syntyyn (mukaillen VanderWeele 2014b)*

2.2 Hajotelmassa käytetyt merkinnät

Oletetaan yksinkertaisuuden vuoksi, että altistus A, välittäjä M ja lopputulos Y ovat binäärisiä. Tällöin altistuksen A tasot ovat $a = 1$ ja $a^\star = 0$. Lopputulos Y voidaan merkitä muotoon Y_a, kun se on yhteydessä altistukseen A. Kun altistus tapahtuu, toisin sanoen $a = 1$, lopputulos on muotoa $Y_a = Y_1$. Vastaavasti $Y_{a^\star} = Y_0$, kun altistusta ei tapahdu. Altistuksen kokonaisvaikutus, TE (*total effect*), on nyt $Y_1 - Y_0$. Vastaavasti merkitään, että välittäjä M tapahtuu, kun $m = 1$. Määritellään, että M_a kuvaa muuttujaa M, kun altistus A on yhteydessä välittäjään. Tällöin altistuksen A kokonaisvaikutus välittäjään on $M_1 - M_0$. Kun sekä välittäjä M että altistus A tapahtuvat, lopputulos on $Y_{am} = Y_{11}$. (VanderWeele, 2014a.)

Kun $A = a$, Y_a ja M_a vastaavat muuttujia Y ja M. Kun $A = a$ ja $M = m$, $Y_{am} = Y$. Merkitään, että $Y_a = Y_{aM_a}$. Näiden merkintöjen valossa voidaan muodostaa altistuksen kokonaisvaikutuksesta neljä komponenttia. Komponenttien tarkkoja arvoja ei ole mahdollista selvittää, mutta niiden odotusarvoja voidaan estimoida tilastollisten menetelmien avulla. (VanderWeele, 2014a, 2014b.)

2.3 Komponenttien muodostaminen

Kontrolloitu vaikutus, CDE, vertaa altistuksen kumpaakin tasoa, kun välittäjä on vakio. CDE kuvastaa näin ollen altistuksen suoraa vaikutusta lopputulokseen, kun välittäjä on asetettu vakioksi m. Määritetään kontrolloitu vaikutus muotoon $Y_{1m} - Y_{0m}$. Altistuksen kokonaisvaikutus, TE, voidaan määritellä seuraavasti:

$$\begin{aligned} Y_1 - Y_0 = (Y_{10} - Y_{00}) + (Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \\ + (Y_{11} - Y_{10} - Y_{01} - Y_{00})(M_1 - M_0) + (Y_{01} - Y_{00})(M_1 - M_0). \end{aligned} \quad (1)$$

Altistuksen kokonaisvaikutus on jaettu yhtälössä (1) neljään komponenttiin. Ensimmäinen komponentti on altistuksen kontrolloitu vaikutus. Välittäjä ei ole läsnä, joten sen arvo on $m^*= 0$. CDE voidaan tällöin kirjoittaa muotoon $Y_{10} - Y_{00}$. Toinen komponentti, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0)$, kuvaa viitteellistä vuorovaikutusta. Komponentti muodostuu kahdesta termistä. Näistä ensimmäinen, $(Y_{11} - Y_{10} - Y_{01} + Y_{00})$, kuvastaa tilannetta, jossa vain altistus tai välittäjä on läsnä, mutta molemmat eivät ole yhtä aikaa. Viitteellisen vuorovaikutuksen toinen termi merkitsee välittäjää, kun altistusta ei tapahdu. Yhtälön (1) kolmas komponentti, $(Y_{11} - Y_{10} - Y_{01} - Y_{00})(M_1 - M_0)$, sisältää altistuksen ja välittäjän vuorovaikutuksen, kun välittäjä on läsnä. Tätä komponenttia voidaan täten kutsua välitetyksi vuorovaikutukseksi. Välitetty vuorovaikutus koostuu kahdesta termistä, joista ensimmäinen on sama kuin viitteellisessä vuorovaikutuksessa. Toinen termi, $(M_1 - M_0)$, on vuorovaikutus, kun altistus vaikuttaa välittäjään. On tärkeä huomata, että $(M_1 - M_0) \neq 0$. Viimeinen komponentti, $(Y_{01} - Y_{00})(M_1 - M_0)$, sisältää välittäjän vaikutuksen, kun altistus ei ole läsnä. Sen ensimmäinen termi, $(Y_{01} - Y_{00})$, kuvastaa siis tilannetta, missä välittäjä tapahtuu mutta altistus ei. Komponentin toinen termi on sama kuin välitetyssä vuorovaikutuksessa, eli altistus on läsnä vain välittäjän kautta. Täten komponentti kuvaa altistuksen puhtaasti epäsuoraa vaikutusta. (VanderWeele, 2014a.)

Four-way decomposition -menetelmän oletus on, että altistus vaikuttaa vähintään yhteen neljästä komponentista (VanderWeele, 2014b). Altistus voi vaikuttaa lopputulokseen ilman välittäjämuuttujaa, jolloin $Y_{10} - Y_{00} \neq 0$. Vaihtoehtoisesti altistuksella on vaikutusta vain vuorovaikutuksessa välittäjän kanssa. Välittäjä voi olla myös läsnä riippumatta altistuksen läsnäolosta. Tällöin $(Y_{11} - Y_{10} - Y_{01} + Y_{00})(M_0) \neq 0$. Kun $(Y_{11} - Y_{10} - Y_{01} - Y_{00})(M_1 - M_0) \neq 0$, altistus vaikuttaa vain välittäjän läsnäollessa ja altistus tarvitaan välittäjän läsnäoloon. Toisin sanoen välittäjä johtuu altistuksesta, ja altistus ja välittäjä ovat vuorovaikutuksessa keskenään. Neljäntenä vaihtoehtona välittäjä toimii yksin ilman altistuksen aktiivista toimintaa, mutta altistus on välttämätön välittäjän läsnäololle, siis $(Y_{01} - Y_{00})(M_1 - M_0) \neq 0$. Altistuksen kokonaisvaikutus voidaan kirjoittaa näiden komponenttien muodossa myös seuraavasti:

$$TE = CDE + INT_{ref} + INT_{med} + PIE.$$

VanderWeele (2016) mainitsee, että puhtaasti epäsuora vaikutus (PIE) ja välitetty vuorovaikutus (INT_{med}) muodostavat yhdessä epäsuoran kokonaisvaikutuksen TIE (*total indirect effect*). Kontrolloitua vaikutusta (CDE) ja viitteellistä vuorovaikutusta (INT_{ref}) vastaavasti voidaan kutsua puhtaasti suoraksi vaikutukseksi, PDE (*pure direct effect*). Näistä voidaan käyttää myös termejä luonnollinen epäsuora vaikutus, NIE, ja luonnollinen suora vaikutus, NDE. Termejä hyödynnetään erityisesti muussa hajotelmiin liittyvässä kirjallisuudessa (esim. VanderWeele, 2013; Tchetgen Tchetgen ja VanderWeele, 2014).

3. KOMPONENTTIEN OSUUS ALTISTUKSEN KOKONAISVAIKUTUKSESTA

3.1 Komponenttien osuuksien estimointi

Four-way decomposition -menetelmän heikkoutena on se, ettei altistuksen kokonaisvaikutuksen komponenteille voi laskea tarkkoja arvoja. Tiettyjen oletusten (VanderWeele, 2014a) valossa on kuitenkin mahdollista tarkastella komponentteja niiden odotusarvon avulla. Jokainen komponentti voidaan ilmaista jakamalla komponentin osuus altistuksen kokonaisvaikutuksella. Määritellään, että $p_{am} = E(Y|A = a, M = m)$. Tällöin komponenttien odotusarvot ovat

$$E[CDE] = p_{10} - p_{00},$$

$$E[INT_{ref}] = (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0),$$

$$E[INT_{med}] = (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\},$$

$$E[PIE] = (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}.$$

Oletetaan, että $p_a = E(Y|A = a)$. Nyt four-way decomposition voidaan ilmaista muodossa

$$\begin{aligned} p_{a=1} - p_{a=0} = {} & (p_{10} - p_{00}) + (p_{11} - p_{10} - p_{01} + p_{00})P(M = 1|A = 0) \\ & + (p_{11} - p_{10} - p_{01} + p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\ & + (p_{01} - p_{00})\{P(M = 1|A = 1) - P(M = 1|A = 0)\}. \quad (2) \end{aligned}$$

Odotusarvojen perusteella voidaan arvioida, kuinka suuri on välittäjän vaikutus, altistuksen vaikutus, niiden välitetty vuorovaikutus tai viitteellinen vuorovaikutus. Määritellään, että $E[TE]$ kuvaa populaation kokonaisvaikutuksen odotusarvoa. Toisin sanoen

$$E[TE] = p_{a=1} - p_{a_0} = E(Y|A = 1) - E(Y|A = 0).$$

Tällöin kunkin komponentin osuus voidaan määritellä jakamalla komponentin odotusarvo kokonaisvaikutuksen odotusarvolla: $\frac{E[CDE]}{E[TE]}$, $\frac{E[INT_{ref}]}{E[TE]}$, $\frac{E[INT_{med}]}{E[TE]}$ ja $\frac{E[PIE]}{E[TE]}$. Välittäjämuuttujan tarkastelussa on huomioitava sekä puhtaasti epäsuora vaikutus että välitetty vuoro-

vaikutus. Välitettyä kokonaisvaikutusta voidaan tarkastella jakamalla nämä kaksi komponenttia altistuksen kokonaisvaikutuksella: $(E[INT_{med}] + E[PIE]) \big/ E[TE]$. Vastaavasti voidaan tarkastella vuorovaikutuksen osuutta lopputulokseen jakamalla viitteellisen ja välitetyn vuorovaikutuksen odotusarvojen summa kokonaisvaikutuksen odotusarvolla, $(E[INT_{ref}] + E[INT_{med}]) \big/ E[TE]$.

Osuuksien laskeminen on mahdollista, kun joko kaikki komponentit ovat positiivisia tai kaikki ovat negatiivisia. Jos näin ei ole, kokonaisvaikutuksen estimaatti voi lähennellä nollaa ja estimointi voi olla mahdotonta.

3.2 Teoreettiset oletukset

Yksittäisten komponenttien tarkan arvon laskeminen aineistosta on four-way decomposition -menetelmän suurin haaste (VanderWeele, 2014a). Jotta komponenttien osuuksia voidaan määritellä edellä mainituin tavoin, täytyy tiettyjen oletusten täyttyä. Oletusten käsittelyn tueksi määritellään ensiksi, että $X \perp\!\!\!\perp Y|Z$ tarkoittaa X:n olevan riippumaton Y:stä ehdolla Z. Oletetaan nyt, että altistuksen A vaikutus lopputulokseen Y on ehdollistettu kovariaattien C arvoihin. Tämä voidaan ilmaista muodossa $Y_{am} \perp\!\!\!\perp A|C$. Toisekseen, oletetaan välittäjän M vaikutuksen lopputulokseen Y olevan ehdollistettu kovariaattien C ja altistuksen A arvoihin, eli $Y_{am} \perp\!\!\!\perp M|\{A, C\}$. Seuraavaksi oletetaan, että altistuksen A vaikutus välittäjään M on ehdollistettu kovariaattien C arvoihin. Tällöin $M_a \perp\!\!\!\perp A|C$. Viimeisenä oletetaan, että altistus A ei ole yhteydessä välittäjään M ja sekoittaviin tekijöihin, jolloin $Y_{am} \perp\!\!\!\perp M_{a^*}|C$. Näiden oletusten ollessa voimassa voidaan komponenttien osuudet ilmaista esitettyjen odotusarvojen avulla. (VanderWeele, 2014a, 2014b.) Edellä mainittuja oletuksia voidaan myös tarkastella aiemmin esitetyn kuvan (2.1) parissa, sillä muuttujien yhteydet toisiinsa on siitä helposti havaittavissa.

Huomataan, että kontrolloitu suora vaikutus vaatii tuekseen vain kaksi ensimmäistä oletusta: $Y_{am} \perp\!\!\!\perp A|C$ ja $Y_{am} \perp\!\!\!\perp M|\{A, C\}$. Siksi CDE voidaan joskus myös vähentää altistuksen kokonaisvaikutuksesta kaavalla $E[PE] := E[TE] - E[CDE]$, missä $PE = INT_{ref} + INT_{med} + PIE$ (*proportion eliminated effect*) määrittää joko välitetyn, interaktiivisen tai molempien osuuden lopputuloksesta. Toisin sanoen PE kuvastaa viitteellisen vuorovaikutuksen, välitetyn vuorovaikutuksen ja puhtaasti epäsuoran vaikutuksen kokonaismäärää. Ilman kahta viimeistä oletusta näiden kolmen komponentin osuuksia ei voi erotella toisistaan, mutta niiden yhteismäärä selviää helposti kontrolloidun vaikutuksen ollessa tiedossa. (VanderWeele, 2014a.)

4. KOMPONENTTIEN ESTIMOINTI SUHDEASTEIKOLLISEN MUUTTUJAN TILANTEESSA

Four-way decomposition on mahdollista kirjoittaa suhdeasteikon mukaisesti jakamalla hajotus $p_{a=o}$:lla (VanderWeele, 2014b). Tällöin

$$\begin{aligned} RR_{a=1} - 1 = \kappa(RR_{10} - 1) + \kappa(RR_{11} - RR_{10} - RR_{01} + 1)P(M = 1|A = 0) \\ + \kappa(RR_{11} - RR_{10} - RR_{01} + 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\} \\ + \kappa(RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}, \end{aligned} \quad (3)$$

missä $RR_{a=1} = p_{a=1}/p_{a=0}$ on suhteellinen riski altistuksen tasoja verratessa, $RR_{am} = p_{am}/p_{00}$ on suhteellinen riski verrattaessa kategoriaa $A = a$ ja $M = m$ viitekategoriaan $A = 0$ ja $M = 0$, ja κ on skaalakerroin, joka lasketaan kaavalla $\kappa = p_{00}/p_{a=0}$. Termi $(RR_{11} - RR_{10} - RR_{01} + 1)$ on Rothmanin suhteellinen lisäriski vuorovaikutuksen takia (Rothman, 1986; Correia ja Williams, 2018). Termiä voidaan merkitä lyhenteellä $RERI$ (*Rothmans excess relative risk due to interaction*), ja se määrittää skaalattua vuorovaikutuksen lisäosuutta.

Yhtälön (3) avulla voidaan estimoida altistuksen kokonaisvaikutuksesta välittäjän osuus, vuorovaikutuksen osuus, välittäjän ja vuorovaikutuksen osuus ja näistä riippumaton osuus. Komponenttien osuus kokonaisvaikutuksesta voidaan estimoida jakamalla komponentti kaikkien komponenttien summalla (VanderWeele, 2014a). Skaalakerroin κ jakautuu pois, kun tarkastellaan yksittäisen komponentin osuutta. Merkitään tällöin, että

$$\begin{aligned} Q = (RR_{10} - 1) + (RR_{11} - RR_{10} - RR_{01} + 1)P(M = 1|A = 1) \\ + (RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}. \end{aligned}$$

Nyt komponentit saadaan yhtälöillä

$$PA_{CDE} = \frac{RR_{10} - 1}{Q},$$

$$PA_{INTref} = \frac{(RR_{11} - RR_{10} - RR_{01} + 1)P(M = 1|A = 0)}{Q},$$

$$PA_{INTmed} = \frac{(RR_{11} - RR_{10} - RR_{01} + 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}{Q},$$

$$PA_{PIE} = \frac{(RR_{01} - 1)\{P(M = 1|A = 1) - P(M = 1|A = 0)\}}{Q}.$$

Estimaatit voidaan laskea suhdeasteikollisten muuttujien tilanteessa regressiomallien avulla hyödyntäen esimerkiksi R-ohjelmistoa (LIITE). Vastaavan skaalatun hajotelman voisi toteuttaa myös uhkasuhteen avulla (VanderWeele, 2014a).

5. YHTEYS REGRESSIOMALLEIHIN

5.1 Hajotuksen mallinnus käytännössä

Four-way decomposition -hajotusta voi mallintaa regressionanalyysin avulla (VanderWeele, 2014a). Ennen tilastollista mallinnusta on kuitenkin oletettava, että luvussa (3.2) mainitut ehdot täyttyvät. Määritellään, että $C = c$. Oletetaan nyt, että vaste Y ja välittäjä M ovat jatkuvia muuttujia sekä seuraavat regressiomallit muuttujille Y ja M pitävät paikkaansa:

$$E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c,$$

$$E[M|a, c] = \beta_0 + \beta_1 a + \beta_2' c.$$

Asetetaan altistus tasoille a ja $a^\star$. Määritellään, että $m^\star$ on muuttujan M taso, jossa välittäjä ei ole läsnä. VanderWeele ja Vansteelandt (2009) osoittivat, että komponenttien kovariaateista riippuvat odotusarvot ovat

$$E[CDE(m^\star)|c] = (\theta_1 + \theta_{3m^\star})(a - a^\star),$$

$$E[INT_{ref}(m^\star)|c] = \theta_3(\beta_0 + \beta_1 a^\star + \beta_2' c)(a - a^\star),$$

$$E[INT_{med}|c] = \theta_3\beta_1(a - a^\star)(a - a^\star),$$

$$E[PIE|c] = (\theta_2\beta_1 + \theta_3\beta_1 a^\star)(a - a^\star).$$

VanderWeele (2014a) osoitti, että puhtaasti suora vaikutus, PDE, voidaan laskea yhtälöllä $E[PDE|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^\star + \beta_2' c)\}(a - a^\star)$. Viitteellinen vuorovaikutus on saatu juuri puhtaasti suoran vaikutuksen ja kontrolloidun suoran vaikutuksen erotuksesta: $E[INT_{ref}(m^\star)|c] = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^\star + \beta_2' c)\}(a - a^\star) - (\theta_1 + \theta_3 m^\star)(a - a^\star) = \theta_3(\beta_0 + \beta_1 a^\star + \beta_2' c - m^\star)(a - a^\star)$.

Regressiomallin hyödyntäminen komponenttien estimoinnissa on yksinkertaista tilastollisten ohjelmistojen avulla. Mutlu et al. (2017) ovat kehittäneet helppokäyttöisen koodin R-ympäristöön (LIITE), kun taas VanderWeele (2014a, 2014b) on suosinut artikkeleissaan SAS-koodia komponenttien estimointiin. Molemmat mainituista koodeista laskevat komponenttien suhteellisen riskin estimaatit sekä niiden osuudet altistuksen kokonaisvai-

kutuksesta. Käytössä olevat koodit eivät vaadi käyttäjältään erityisen paljoa tietoa muuttujien jatkuvuuden tai binäärisyyden vaikutuksesta malleihin. Tarkastellaan kuitenkin yksityiskohtaisemmin muuttujien ominaisuuksien vaikutusta regressioanalyysiin.

5.2 Jatkuva vaste ja jatkuva välittäjä

Oletetaan, että vaste Y ja välittäjämuuttuja M ovat jatkuvia ja altistus on binäärinen, eli $a = 1$ ja $a^{\star} = 0$. Määritellään, että $m^{\star} = 0$, kun välittäjä ei ole läsnä. Tällöin kontrolloitu vaikutus on muotoa $E[CDE|c] = \theta_1$, sillä altistuksen tasojen erotus $(a - a^{\star}) = 1$. Altistuksen puhtaasti suora vaikutus on $E[PDE|c] = \theta_1 + \theta_3(\beta_0 + \beta_2' c)$. Suoran vaikutuksen ja kontrolloidun vaikutuksen erotus antaa tällöin viitteellisen vuorovaikutuksen odotusarvolle yhtälön $E[INT_{ref}|c] = \theta_3(\beta_0 + \beta_1 + \beta_2' c)$. Välittäjän tason $m^{\star}$ ollessa nolla, puhtaasti epäsuora vaikutus voidaan laskea yhtälöllä $E[PIE] = \theta_2\beta_1$. Välitetty vuorovaikutus on nyt muotoa $E[INT_{med}|c] = \theta_3\beta_1$. (VanderWeele, 2014a.)

VanderWeele ja Vansteelandt (2009) osoittivat, että varianssiapproksimaation avulla on mahdollista tarkastella komponenttien odotusarvojen keskivirhettä. Myös bootstrap-menetelmän käyttäminen on mahdollista keskivirheen tarkasteluun. Tilastolliset ohjelmistot kykenevät laskemaan jatkuvan vasteen ja välittäjän tilanteessa komponenttien suhteelliset riskit yksinkertaisten käskyjen avulla.

5.3 Jatkuva vaste, binäärinen välittäjä

Oletetaan, että vaste Y on jatkuva ja välittäjä M binäärinen. Tällöin regressiomallit muuttujille Y ja M ovat

$$E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c,$$

$$logit\{P(M = 1|a, c)\} = \beta_0 + \beta_1 a + \beta_2' c.$$

Valerin ja VanderWeelen (2013) todistuksen mukaan kontrolloidun suoran vaikutuksen kovariaatista riippuva odotusarvo on muotoa $E[CDE(m^{\star})|c] = (\theta_1 + \theta_3 m^{\star})(a - a^{\star})$. Kontrolloitu vaikutus ei siis muutu aiemmasta, sillä välittäjämuuttuja ei ole läsnä altistuksen suorassa vaikutuksessa. Altistuksen puhtaasti epäsuoran vaikutuksen odotusarvo binäärisen välittäjämuuttujan tapauksessa on

$$E[PIE|c] = (\theta_2 + \theta_3 a^{\star})\frac{exp[\beta_0 + \beta_1 a + \beta_2' c]}{1 + exp[\beta_0 + \beta_1 a + \beta_2' c]} - \frac{exp[\beta_0 + \beta_1 a^{\star} + \beta_2' c]}{1 + exp[\beta_0 + \beta_1 a^{\star} + \beta_2' c]}.$$

Viitteellisen vuorovaikutuksen odotusarvo saadaan epäsuoran vaikutuksen ja kontrolloidun vaikutuksen erotuksesta:

$$E[INT_{ref}(m^\star)|c] = \{\theta_1(a-a^\star)\} + \{\theta_3(a-a^\star)\}\frac{exp[\beta_0+\beta_1 a^\star+\beta_2' c]}{1+exp[\beta_0+\beta_1 a^\star+\beta_2' c]}$$
$$-(\theta_1+\theta_3 m^\star)(a-a^\star) = \theta_3(a-a^\star)\left(\frac{exp[\beta_0+\beta_1 a^\star+\beta_2' c]}{1+exp[\beta_0+\beta_1 a^\star+\beta_2' c]} - m^\star\right).$$

Vastaavasti välitetyn vuorovaikutuksen odotusarvo voidaan estimoida epäsuoran kokonaisvaikutuksen ja puhtaasti epäsuoran vaikutuksen erotuksesta (Valeri ja VanderVeele, 2013):

$$E[INT_{med}|c] = (\theta_2+\theta_3 a)\left\{\frac{exp[\beta_0+\beta_1 a+\beta_2' c]}{1+exp[\beta_0+\beta_1 a+\beta_2' c]} - \frac{exp[\beta_0+\beta_1 a^\star+\beta_2' c]}{1+exp[\beta_0+\beta_1 a^\star+\beta_2' c]}\right\}$$
$$-(\theta_2+\theta_3 a^\star)\left\{\frac{exp[\beta_0+\beta_1 a+\beta_2' c]}{1+exp[\beta_0+\beta_1 a+\beta_2' c]} - \frac{exp[\beta_0+\beta_1 a^\star+\beta_2' c]}{1+exp[\beta_0+\beta_1 a^\star+\beta_2' c]}\right\}$$
$$= \theta_3(a-a^\star)\left\{\frac{exp[\beta_0+\beta_1 a+\beta_2' c]}{1+exp[\beta_0+\beta_1 a+\beta_2' c]} - \frac{exp[\beta_0+\beta_1 a^\star+\beta_2' c]}{1+exp[\beta_0+\beta_1 a^\star+\beta_2' c]}\right\}.$$

Binäärinen välittäjämuuttuja ei estä varianssiapproksimaation ja bootstrap-menetelmän käyttöä estimaattien keskivirheen tarkastelussa (VanderWeele ja Vansteelandt, 2009; VanderWeele, 2014a). Tilastolliset ohjelmistot (LIITE; VanderWeele, 2014a) mahdollistavat välittäjän asettamisen binääriseksi estimoinnin yhteydessä.

5.4 Binäärinen vaste ja binäärinen välittäjä

Kun muuttujat Y ja M ovat binäärisiä sekä lopputuloksen voi olettaa olevan harvinainen, muuttujien Y ja M regressiomallit ovat muotoa

$$logit\{P(Y=1|a,m,c)\} = \theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c,$$

$$logit\{P(M=1|a,c)\} = \beta_0+\beta_1 a+\beta_2' c.$$

Altistuksen kokonaisvaikutuksen, kontrolloidun suoran vaikutuksen ja puhtaasti suoran vaikutuksen kovariaateista $C = c$ riippuvat riskisuhteet voidaan estimoida ristisuhteen avulla, sillä lopputulos on harvinainen (Valeri ja VanderWeele, 2013). Estimoidaan riskit tällöin seuraavasti:

$$RR_c^{TE} \approx \frac{\exp(\theta_1 a)\{1+\exp(\beta_0+\beta_1 a^\star+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a)\}}{\exp(\theta_1 a^\star)\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star)\}},$$

$$RR_c^{CDE}(m^\star) \approx \exp\{(\theta_1+\theta_3 m)(a-a^\star)\},$$

$$RR_c^{PIE} \approx \frac{\{1+\exp(\beta_0+\beta_1 a^\star+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^\star)\}}{\{1+\exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+\exp(\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star)\}}.$$

Oletetaan yhä, että lopputulos on harvinainen. Merkitään, että

$$\kappa = \frac{E(Y_{a^\star m^\star}|c)}{E(Y_{a^\star}|c)} = \frac{E[Y|a^\star, m^\star, c]}{\int E[Y|a^\star, m, c]dP(m|a^\star, c)}$$
$$\approx \frac{\exp(\theta_0 + \theta_1 a^\star + \theta_2 m^\star + \theta_3 a^\star m^\star + \theta_4' c)}{\exp\{\theta_0 + \theta_1 a^\star + \theta_4' c\} \int \exp\{(\theta_2 + \theta_3 a^\star)m\}dP(m|a^\star, c)}$$
$$= \frac{\exp(\theta_2 m^\star + \theta_3 a^\star m^\star)}{\frac{1+\exp(\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star)}{1+\exp(\beta_0+\beta_1 a^\star+\beta_2' c)}} = \frac{exp[\theta_2 m^\star + \theta_3 a^\star m^\star]\{1 + exp[\beta_0 + \beta_1 a^\star + \beta_2' c]\}}{1 + exp[\beta_0 + \beta_1 a^\star + \beta_2' a + \theta_2 + \theta_3 a^\star]}.$$

Merkitään myös, että

$$\int \frac{E[Y|a, m, c]}{E[Y|a^\star, m^\star, c]} dP(m|a^\dagger, c)$$
$$\approx \int \exp(\theta_1 a + \theta_2 m + \theta_3 am - \theta_1 a^\star - \theta_2 m^\star - \theta_3 a^\star m^\star)dP(m|a^\dagger, c)$$
$$= \exp\{\theta_1(a - a^\star) - \theta_2 m^\star - \theta_3 a^\star m^\star\} \int \exp\{(\theta_2 + \theta_3 a)m\}dP(m|a^\dagger, c)$$
$$= \frac{exp[\theta_1(a - a^\star) - \theta_2 m^\star - \theta_3 a^\star m^\star}{1 + exp[\beta_0 + \beta_1 a^\dagger + \beta_2' c]}\{1 + exp[\beta_0 + \beta_1 a^\dagger + \beta_2' c + \theta_2 + \theta_3 a]\}.$$

Tällöin viitteellisen vuorovaikutuksen riski voidaan laskea seuraavasti:

$$RR_c^{INT_{ref}}(m^\star) = \int \left\{ \frac{E[Y|a, m, c]}{E[Y|a^\star, m^\star, c]} - \frac{E[Y|a^\star, m, c]}{E[Y|a^\star, m^\star, c]} - \frac{E[Y|a, m^\star, c]}{E[Y|a^\star, m^\star, c]} + 1 \right\} dP(m|a^\star, c)$$
$$= \frac{e^{\theta_1(a-a^\star)-\theta_2 m^\star-\theta_3 a^\star m^\star}(1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_{2a}})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c}}$$
$$- \frac{e^{-\theta_2 m^\star-\theta_3 a^\star m^\star}(1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c}} - e^{(\theta_1+\theta_3 m^\star)(a-a^\star)} + 1.$$

Viitteellisen vuorovaikutuksen komponentti saadaan yhtälöllä

$$\kappa RR_c^{INT_{ref}}(m^\star) = \frac{e^{\theta_1(a-a^\star)}(1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star}} - 1$$
$$- \frac{e^{\theta_1(a-a^\star)+\theta_2 m^\star+\theta_3 a m^\star}(1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star}} + \frac{e^{\theta_2 m^\star+\theta_3 a^\star m^\star}(1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star}}.$$

Välitetyn vuorovaikutuksen riski voidaan laskea yhtälöllä

$$RR_c^{INT_{med}} = \int \left\{ \frac{E[Y|a, m, c]}{E[Y|a^\star, m^\star, c]} - \frac{E[Y|a^\star, m, c]}{E[Y|a^\star, m^\star, c]} \right\} \{dP(m|a, c) - dP(m|a^\star, c)\}$$
$$= \frac{e^{\theta_1(a-a^\star)-\theta_2 m^\star-\theta_3 a^\star m^\star}(1 + e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c}} - \frac{e^{-\theta_2 m^\star-\theta_3 a^\star m^\star}(1 + e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^\star})}{1 + e^{\beta_0+\beta_1 a+\beta_2' c}}$$
$$- \frac{e^{\theta_1(a-a^\star)-\theta_2 m^\star-\theta_3 a^\star m^\star}(1 + e^{\beta_0+\beta_1 a^\star+\beta_2' a+\theta_2+\theta_3 a})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c}} + \frac{e^{-\theta_2 m^\star-\theta_3 a^\star m^\star}(1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star})}{1 + e^{\beta_0+\beta_1 a^\star+\beta_2' c}},$$

ja välitetyn vuorovaikutuksen komponentti yhtälöllä

$$\kappa RR_c^{INT_{med}} = \frac{e^{\theta_1(a-a^\star)}(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a})(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c})}{(1+e^{\beta_0+\beta_2 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star})(1+e^{\beta_0+\beta_1 a+\beta_2' c})} -$$
$$\frac{(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\beta_3 a^\star})(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c})}{(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star})(1+e^{\beta_0+\beta_1 a+\beta_2' c})} - \frac{e^{(a-a^\star)}(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a})}{(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star})} + 1.$$

Kontrolloidun suoran vaikutuksen komponentti lasketaan seuraavasti:

$$\kappa[RR_c^{CDE}(m^\star) - 1] = \kappa[e^{(\theta_1+\theta_3 m^\star)(a-a^\star)} - 1]$$
$$= \frac{e^{\theta_1(a-a^\star)+\theta_2 m^\star+\theta_3 a m^\star}(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c})}{1+e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star}} - \frac{e^{\theta_2 m^\star+\theta_3 a^\star m^\star}(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c})}{1+e^{\beta_0+\beta_2 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star}}.$$

Viimeinen komponentti, altistuksen puhtaasti epäsuora vaikutus, määritetään yhtälöllä

$$\kappa \int_m \frac{E(Y_{a^\star m}|c)}{E(Y_{a^\star m^\star}|c)}\{dP(m|a,c) - dP(m|a^\star,c)\}$$
$$= \kappa\left(\frac{e^{-\theta_2 m^\star-\theta_3 a^\star m^\star}(1+e^{\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^\star})}{1+e^{\beta_0+\beta_1 a+\beta_2' c}} - \frac{e^{-\theta_2 m^\star-\theta_3 a^\star m^\star}(1+e^{\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star})}{1+e^{\beta_0+\beta_1 a^\star+\beta_2' c}}\right)$$
$$= \frac{\{1+exp(\beta_0+\beta_1 a^\star+\beta_2' c)\}\{1+exp(\beta_0+\beta_1 a+\beta_2' c+\theta_2+\theta_3 a^\star)\}}{\{1+exp(\beta_0+\beta_1 a+\beta_2' c)\}\{1+exp(\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2+\theta_3 a^\star)\}} - 1.$$

Kun sekä vaste että välittäjä ovat binäärisiä, on niiden estimointi selvästi haasteellisempaa verrattuna jatkuvien muuttujien tapaukseen. Erityisesti tällaisessa tilanteessa tilastollisten ohjelmistojen hyödyntäminen on kannattavaa. Luvussa (6) käytetään komponenttien estimoinnissa juurikin binäärisiä muuttujia esimerkkinä siitä, kuinka R-ohjelmistossa (LIITE) voidaan toteuttaa four-way decomposition binäärisen vasteen, välittäjän ja altistuksen tapauksessa. Valeri ja VanderWeele (2013) nostavat esiin mahdollisuuksen tutkia binääristenkin muuttujien tilanteessa komponenttien keskivirheitä varianssiapproksimaation avulla.

5.5 Binäärinen vaste, jatkuva välittäjä

Oletetaan, että vaste Y on binäärinen ja välittäjä M on jatkuva. Oletetaan myös, että lopputulos on harvinainen ja seuraavat logistiset regressiomallit pitävät paikkaansa:

$$logit\{P(Y=1|a,m,c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c,$$

$$E[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c.$$

Välittäjä M on normaalisti jakautunut ja sen varianssi on σ^2. Välittäjä on riippuva muuttujista A ja C. Koska lopputuloksen oletetaan olevan harvinainen, voidaan käyttää ristisuhdetta riskisuhteen arvioimiseen. VanderWeele ja Vansteelandt (2010) osoittivat, että

altistuksen kokonaisvaikutuksen, kontrolloidun vaikutuksen ja epäsuoran vaikutuksen riskisuhteet voidaan arvioida yhtälöillä

$$RR_c^{TE} \approx exp\big[\theta_1 + \theta_2\beta_1 + \theta_3(\beta_0 + \beta_1 a^\star + \beta_1 a + \beta_2' c + \theta_2\sigma^2)(a - a^\star) + \frac{1}{2}\theta_3^2\sigma^2(a^2 - a^{\star 2})\big],$$

$$RR_c^{CDE}(m^\star) \approx \exp\big[(\theta_1 + \theta_3 m^\star)(a - a^\star)\big],$$

$$RR_c^{PIE} \approx exp\big[(\theta_2\beta_1 + \theta_3\beta_1 a^\star)(a - a^\star)\big].$$

Määritellään, että

$$\begin{aligned}
\kappa = \frac{E(Y_{a^\star m^\star}|c)}{E(Y_{a^\star}|c)} &= \frac{E[Y|a^\star, m^\star, c]}{\int E[Y|a^\star, m, c]dP(m|a^\star, c)} \\
&\approx \frac{\exp(\theta_0 + \theta_1 a^\star + \theta_2 m^\star + \theta_3 a^\star m^\star + \theta_4' c)}{\exp\{\theta_0 + \theta_1 a^\star + \theta_4' c\}\int \exp\{(\theta_2 + \theta_3 a^\star)m\}dP(m|a^\star, c)} \\
&= \frac{exp(\theta_2 m^\star + \theta_3 a^\star m^\star)}{exp\{(\theta_2 + \theta_3 a^\star)(\beta_0 + \beta_0 + \beta_1 a^\star + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a^\star)^2\sigma^2\}} \\
&= exp[\theta_2 m^\star + \theta_3 a^\star m^\star - (\theta_2 + \theta_3 a^\star)(\beta_0\beta_1 a^\star + \beta_2' c) - \frac{1}{2}(\theta_2 + \theta_3 a^\star)^2\sigma^2].
\end{aligned}$$

Voidaan määrittää myös, että

$$\begin{aligned}
&\int \frac{E[Y|a, m, c]}{E[Y|a^\star, m^\star, c]}dP(m|a^\dagger, c) \\
&\quad\approx \int \exp(\theta_1 a + \theta_2 m + \theta_3 am - \theta_1 a^\star - \theta_2 m^\star - \theta_3 a^\star m^\star)dP(m|a^\dagger, c) \\
&\quad= \exp\{\theta_1(a - a^\star) - \theta_2 m^\star - \theta_3 a^\star m^\star\}\int \exp\{(\theta_2 + \theta_3 a)m\}dP(m|a^\dagger, c) \\
&= exp\big\{\theta_1(a - a^\star) - \theta_2 m^\star - \theta_2 a^\star m^\star\big\}exp\big\{(\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^\dagger + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2\big\} \\
&= exp\{\theta_1(a - a^\star) - \theta_2 m^\star - \theta_3 a^\star m^\star + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^\dagger + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2\}.
\end{aligned}$$

Tällöin viitteellisen vuorovaikutuksen riski saadaan yhtälöllä

$$\begin{aligned}
RR_c^{INT_{ref}}(m^\star) &= \int \Big\{\frac{E[Y|a, m, c]}{E[Y|a^\star, m^\star, c]} - \frac{E[Y|a^\star, m, c]}{E[Y|a^\star, m^\star, c]} - \frac{E[Y|a, m^\star, c]}{E[Y|a^\star, m^\star, c]} + 1\Big\}dP(m|a^\star, c) \\
&= e^{\theta_1(a - a^\star) - \theta_2 m^\star - \theta_3 a^\star m^\star + (\theta_2 + \theta_3 a)(\beta_0 + \beta_1 a^\star + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a)^2\sigma^2} \\
&\quad - e^{-\theta_2 m^\star - \theta_3 a^\star m^\star + (\theta_2 + \theta_3 a^\star)(\beta_0 + \beta_1 a^\star + \beta_2' c) + \frac{1}{2}(\theta_2 + \theta_3 a^\star)^2\sigma^2} - e^{(\theta_1 + \theta_3 m^\star)(a - a^\star)} + 1,
\end{aligned}$$

ja komponentti yhtälöllä

$$\begin{aligned}\kappa RR_c^{INT_{ref}}(m^\star) = e^{\{\theta_1+\theta_3(\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2\sigma^2)\}(a-a^\star)+\frac{1}{2}\theta_3^2\sigma^2(a^2-a^{\star 2})} - 1 \\ -e^{\theta_1(a-a^\star)+\theta_2 m^\star+\theta_3 a m^\star-(\theta_2+\theta_3 a^\star)(\beta_0+\beta_1 a^\star+\beta_2' c)-\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2} \\ +e^{\theta_2 m^\star+\theta_3 a^\star m^\star-(\theta_2+\theta_3 a^\star)(\beta_0+\beta_1 a^\star+\beta_2' c)-\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2}.\end{aligned}$$

Välitetyn vuorovaikutuksen riski voidaan laskea seuraavasti:

$$\begin{aligned}RR_c^{INT_{med}} &= \int \Big\{\frac{E[Y|a,m,c]}{E[Y|a^\star,m^\star,c]} - \frac{E[Y|a^\star,m,c]}{E[Y|a^\star,m^\star,c]}\Big\}\{dP(m|a,c) - dP(m|a^\star,c)\} \\ &\approx e^{\theta_1(a-a^\star)-\theta_2 m^\star-\theta_3 a^\star m^\star+(\theta_2+\theta_3 a)(\beta_+\beta_1 a+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2} \\ &\quad - e^{-\theta_2 m^\star-\theta_3 a^\star m^\star+(\theta_2+\theta_2 a^\star)(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2} \\ &\quad - e^{\theta_1(a-a^\star)-\theta_2 m^\star-\theta_2 a^\star m^\star+(\theta_2+\theta_3 a)(\beta_0+\beta_1 a^\star+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a)^2\sigma^2} \\ &\quad + e^{-\theta_2 m^\star-\theta_3 a^\star m^\star+(\theta_2+\theta_3 a^\star)(\beta_0+\beta_1 a^\star+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2}.\end{aligned}$$

Tällöin välitetyn vuorovaikutuksen komponentti on

$$\begin{aligned}\kappa RR_c^{INT_{med}} &= e^{\{\theta_1+\theta_2\beta_1+\theta_3(\beta_0+\beta_1 a+\beta_2' c+\theta_2\sigma^2)\}(a-a^\star)+\frac{1}{2}\theta_3^2\sigma^2(a^2-a^{\star 2})} \\ &\quad - e^{(\theta_2\beta_1+\theta_3\beta_1 a^\star)(a-a^\star)} - e^{\{\theta_1+\theta_3(\beta_0+\beta_1 a^\star+\beta_2' c+\theta_2\sigma^2)\}(a-a^\star)+\frac{1}{2}\theta_3^2\sigma^2(a^2-a^{\star 2})} + 1.\end{aligned}$$

Viimeiseksi määritellään, että altistuksen kontrolloidun vaikutuksen komponentti on

$$\begin{aligned}\kappa[RR_c^{CDE}(m^\star) - 1] &= \kappa[e^{(\theta_1+\theta_3 m^\star)(a-a^\star)} - 1] \\ &= e^{\theta_1(a-a^\star)+\theta_2 m^\star+\theta_3 a m^\star-(\theta_2+\theta_3 a^\star)(\beta_0+\beta_1 a^\star+\beta_2' c)-\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2} \\ &\quad - e^{\theta_2 m^\star+\theta_3 a^\star m^\star-(\theta_2+\theta_3 a^\star)(\beta_0+\beta_1 a^\star+\beta_2' c)-\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2}\end{aligned}$$

ja puhtaasti epäsuoran vaikutuksen komponentti on

$$\begin{aligned}(RR_c^{PIE} - 1) &= \kappa \int \frac{E(Y_{a^\star m}|c)}{E(Y_{a^\star m^\star}|c)}\{dP(m|a,c) - dP(m|a^\star,c)\} \\ &= \kappa\Big\{e^{-\theta_2 m^\star-\theta_3 a^\star m^\star+(\theta_2+\theta_3 a^\star)(\beta_0+\beta_1 a+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2} \\ &\quad - e^{-\theta_2 m^\star-\theta_3 a^\star m^\star+(\theta_2+\theta_3 a^\star)(\beta_0+\beta_1 a^\star+\beta_2' c)+\frac{1}{2}(\theta_2+\theta_3 a^\star)^2\sigma^2}\Big\} \\ &= e^{(\theta_2\beta_1+\theta_3\beta_1 a^\star)(a-a^\star)} - 1.\end{aligned}$$

Kuten aiemmissakin tilanteissa, myös binäärisen vasteen ja jatkuvan välittäjän kohdalla voidaan hyödyntää varianssiapproksimaatiota ja bootstrap-menetelmää keskivirheiden estimointiin (VanderWeele ja Vansteelandt, 2010).

6. SOVELLUSESIMERKKI: OPPIMISVAIKEUKSIEN VAIKUTUS MASENNUKSEEN

On todettu, että oppimisvaikeuksien ja masennuksen välillä on jonkinlainen yhteys (Maag ja Reid, 2006). Vaikuttaisi siltä, että oppimisvaikeuksista, kuten luki- ja tarkkaavaisuushäiriöstä, kärsivät voivat olla normaalia alttiimpia sairastumaan masennukseen. On myös mahdollista, että oppimishäiriöt altistavat erityisesti vakavalle masennukselle. Vaikka Maag ja Reid (2006) toteavat, etteivät masennusta mittaavat menetelmät ole välttämättä hyviä mittareita tutkimaan masennuksen kliinisiä eroja oppimishäiriöisten ja ei-oppimishäiriöisten välillä, kokevat he oppimisvaikeuksien liittyvän vahvasti masennukseen. Artikkelissaan he kommentoivat, kuinka aihe kaipaisi enemmän tutkimusta tarkempien yhteyksien tarkasteluun. Tutkitaankin nyt four-way decomposition -hajotusmenetelmän avulla oppimisvaikeuksien yhteyttä masennukseen.

Tiedossa on myös, että alkoholi ja masennus ovat yhteydessä toisiinsa. Flensborg-Madsen (2011) nostaa esiin, kuinka runsas alkoholinkäyttö voi johtaa masennukseen tai johtua jo olemassaolevasta masennuksesta. Tarkastellaan alkoholin yhteyttä masennukseen käyttämällä yksilön alkoholinkäytön herättämiä tunteita välittäjänä esimerkissä.

6.1 Aineisto ja menetelmät

Aineistona käytetään korkeakouluopiskelijoiden terveyskyselyä vuodelta 2016 (Kunttu, Pesonen ja Saari, 2016). Aineiston koko on 3110. Poimitaan kovariaateiksi tutkimukseen osallistuneiden ikä, sukupuoli ja perhemuoto, kuten puolison tai perheen kanssa asuminen. Kyselyssä on selvitetty, kuinka paljon syyllisyyttä vastannut opiskelija kokee alkoholinkäytöstään asteikolla 0-4. Asetetaan välittäjämuuttuja niin, että voidaan verrata yksilöitä, jotka eivät tunne alkoholinkäytöstään lainkaan syyllisyyttä, niihin, jotka tuntevat edes vähän. Altistukseksi valitaan todettu oppimisvaikeus ja vasteeksi masennus, jonka lääkäri tai psykiatri on diagnosoinut. Käytetään R-ohjelmistoa laskemaan komponenttien arvot ja niiden suhteelliset osuudet.

Osuuksien estimoimista varten muutetaan välittäjä ja vaste binäärisiksi muuttujiksi. Asetetaan välittäjä M muotoon $m = 1$, kun yksilö kokee jossain määrin syyllisyyttä alkoholinkäytöstä. Vastaavasti $m^\star = 0$, kun yksilöllä ei ole lainkaan syyllisyyden kokemusta

alkoholin nauttimisesta. Vaste Y on 1, kun kyselyyn vastaajalla on diagnosoitu masennus, ja 0, kun masennusta ei ole diagnosoitu. Altistuksen tasot ovat $a = 2$ ja $a^\star = 1$, missä A on todettu oppimisvaikeus.

Komponenttien estimointiin käytetään valmista R-koodia (LIITE; Mutlu et al., 2017), johon tarvitsee itse lisätä vain tutkittava aineisto ja muuttujien tiedot. Koodi vaatii rinnalleen myös lähdekoodin, joka ladataan samaan tiedostosijaintiin käytettävän koodin kanssa (Mutlu et al., 2017). Aineistona tässä tutkimuksessa on daF3224-tiedosto, joka on csv-tiedostomuodossa. R-koodi mahdollistaa myös muiden tiedostomuotojen lukemisen ja tallentamisen. Ensimmäiseksi on määriteltävä tietopolku tiedostoon, joka sisältää aineiston. Sen lisäksi valitaan polku, johon tallennetaan taulukkona lasketut estimaatin tiedostoon, jonka nimen voi käyttäjä itse määrittää.

```
data_path <- 'daF3224.csv'
output <- 'Results.csv'
```

Seuraavaksi määritellään muuttujat. Altistus A vastaa binääristä muuttujaa $k7$, eli yksilön oppimisvaikeutta. Muuttuja $rek80$ on binääriseksi rekonstruoitu välittäjä M. Myös vaste $Y = rek6\_26$ on rekonstruoitu binääriseksi.

```
A <- ‘k7’
M <- ‘rek80’
Y <- ‘rek6_26’
```

Määritellään kovariaatit $k1 =$ vastaajan ikä, $k2 =$ vastaajan sukupuoli ja $k106 =$ vastaajan perhemuoto.

```
COVAR <- c(‘k1’, ‘k2’, ‘k106’)
```

Vaste ja välittäjä ovat tässä tutkimuksessa binäärisiä, mutta myös jatkuvia muuttujia voi käyttää. Altistuksen A tasot määritellään halutuiksi, ja määritetään taso $m^\star$, jolla välittäjä ei ole läsnä.

```
# 1 = binäärinen ja 0 = jatkuva
outcome <- 1
mediator <- 1

a <- 2
astar <- 1
mstar <- 0
```

Koodi käytää bootstrap-menetelmää komponenttien ja luottamusvälien estimointiin. Bootstrap-menetelmää (Zoubir ja Iskander, 2007) käytetään yleisesti tunnuslukujen ja parametrien otantajakauman estimoimiseen. Menetelmä auttaa arvioimaan estimaatteja, joiden tarkka arvo on muuten hankala saada. Bootstrap perustuu otoksiin aineistosta, joiden avulla

arvioidaan haluttuja tunnuslukuja. Tarkkaa määrää toistoille ei ole määritelty. Mitä isompi määrä toistoja on, sitä tarkempi estimaatti voidaan saada. Usein toistojen määrä on vähintään tuhat, sillä pienempi määrä vähentää tuloksen uskottavuutta. Kun käytetty algoritmi tekee tuhansia toistoja, sen nopeus ja tehokkuus kärsii. Liian iso määrä toistoja voi kestää hyvin kauan, eikä välttämättä kaikilla tietokoneilla ole tarpeeksi tehoa algoritmin ajamiseen. Tähän esimerkkiin valitaan tuhat iteraatiota, mutta määrä on koodin (LIITE) käyttäjän vapaasti päätettävissä.

```
# Bootstrapi-toistojen määrä.
N_r <- 1000
```

Tämän osion jälkeen koodissa on kutsut funktioihin, jotka sijaitsevat lähdekoodissa (Mutlu et al., 2017). Käyttäjän ei tule muokata kutsuja, mutta aineiston sisältävän tiedoston lukemiseen on määritettävä, jos tiedosto on jotain muuta muotoa kuin SPSS-tiedosto. Kun käyttäjä ajaa loput koodista, tapahtuu haluttujen komponenttien estimointi. Tulokset löytyvät määritetystä tiedostosijainnista annetun nimen mukaisesta tiedostosta.

Saaduista estimaateista ensimmäinen kuvaa komponenttien suhteellista riskiä. Tässä tapauksessa CDE on kovariaattien ja oppimisvaikeuksien vaikutus masennukseen. INT_{ref} on oppimisvaikeuksien ja alkoholinkäytön aiheuttaman syyllisyydentunteen vuorovaikutus. Vastaavasti INT_{med} on oppimisvaikeuksien ja alkoholinkäytön aiheuttamien syyllisyydentunteiden vuorovaikutuksen sekä oppimisvaikeuksien ja alkoholinkäytön herättämien tunteiden itsenäisten vaikutusten osuus. PIE kuvaa pelkästään alkoholinkäytön aiheuttaman syyllisyydentunteen vaikutusta masennukseen. Altistuksen kokonaisvaikutuksen, suhteellinen riski on ilmaistu alimpana taulukossa.

6.2 Tulokset

Tuloksia pääsee tarkastelemaan R-koodin kirjoittamasta tiedostosta. Taulukossa (6.1) näkyy tämän esimerkin olennaisimmat tulokset. Koodi antaa automaattisesti 95 % luottamusvälin parametreille ja altistuksen kokonaisvaikutusta selittäville osuuksille. Esimerkin tiedoilla lasketut luottamusvälit ovat isoja, mikä ei anna hyvää kuvaa tulosten tarkkuudesta.

Taulukko 6.1. *Komponenttien estimaatit luottamusarvoineen*

Komponentti	Suhteellinen lisäriski	95 % luottamusväli	Selittävä osuus	95 % Luottamusväli
CDE	0.607	(-0.570, 1.456)	66.38 %	(-45.98 %, 183.40 %)
INT_{ref}	0.270	(-0.470, 1.044)	29.55 %	(-72.89 %, 131.31 %)
INT_{med}	0.023	(-0.096, 0.102)	2.48 %	(-12.07 %, 13.59 %)
PIE	0.015	(-0.022, 0.040)	1.61 %	(-8.91 %, 4.97 %)
TE	0.914	(-0.015, 1.686)		

Kontrolloidun suoran vaikutuksen, CDE:n osuus on selkeästi suurin. Altistuksen kokonaisvaikutuksesta kontrolloitu vaikutus selittää noin 66 %. Oppimishäiriöillä on selvästi yksinään jotain yhteyttä masennusdiagnoosiin. Alkoholinkäytön kokemuksen välittämä vaikutus, toisin sanoen puhtaasti epäsuora vaikutus, on 1.6 % altistuksen kokonaisvaikutuksesta. Välitetty vaikutus, $(PIE + INT_{med})/TE$, selittää noin 4 % kokonaisvaikutuksesta. Interaktiivisen vaikutuksen osuus sen sijaan oli jopa yli 30 %. Luvussa 3 mainittiin muuttuja PE (proportion eliminated), joka lasketaan vähentämällä kontrolloitu vaikutus kokonaisvaikutuksesta. Nyt sen osuus on lähes 34 %, mikä kuvaa vuorovaikutuksen ja välitetyn vaikutuksen osuutta altistuksen kokonaisvaikutuksesta.

Oppimishäiriön kokonaisvaikutuksen suhteellinen riski masennuksen syntyyn on noin 1.9, eli altistuksen tuoma lisäriski on 0.9 (95 % luottamusväli: 0.986-2.686). Se vaikuttaisi olevan siis melko merkittävä masennuksen kannalta. Välitetty suhteellinen riski (0.038) on hyvin pieni, mutta viitteellisen vuorovaikutuksen ja kontrolloidun vaikutuksen suhteellinen riski, 0.877, on selkeästi merkittävämpi.

Näiden lukemien perusteella vaikuttaisi, että alkoholinkäytön aiheuttamalla syyllisyydellä on jonkinlainen yhteys masennuksen syntyyn. Epäsuora vaikutus yksinään oli kuitenkin melko pieni tekijä altistuksen kokonaisvaikutuksen kannalta, mutta altistuksen ja välittäjän viitteellisen vuorovaikutuksen suhteellinen riski, 0.270 (95 % luottamusväli: -0.470 – 1.044), on selkeästi merkittävä kokonaisvaikutuksen kannalta. Oppimisvaikeudet vaikuttaisivat lisäävän masennuksen riskiä sekä suorasti että epäsuorasti. Oppimisvaikeuden kontrolloitu suora vaikutus (suhteellinen riski: 0.607; 95 % luottamusväli: -0.570 – 1.456) oli kuitenkin merkittävin komponentti tuloksien perusteella.

7. POHDINTA

Four-way decomposition on hajotusmenetelmä, joka jakaa altistuksen vaikutuksen neljään komponenttiin. Komponenttien avulla on mahdollista tutkia välittäjän, altistuksen sekä välittäjän ja altistuksen vuorovaikutuksen osuutta lopputuloksesta. Komponentit määrittävät siten altistuksen suoran ja epäsuoran vaikutuksen lopputulokseen. Hajotusmenetelmä auttaa kausaalisuhteiden tarkastelussa merkittävästi, sillä altistuksen vaikutus ei välttämättä ole niin yksiselitteinen kuin on helppo ymmärtää perinteisten tilastollisten mallinnusten avulla. Usein altistukseen vaikuttaa jokin ulkopuolinen välittäjä, jonka vaikutus lopputulokseen yhdessä altistuksen kanssa voi olla hyvinkin merkittävä. Tämä ei kuitenkaan selviä ilman komponenttien yksityiskohtaista tarkastelua.

Hajotusmenetelmän heikkoutena on sen kykenemättömyys antaa täsmällistä arvoa komponenteille. Jokainen komponentti estimoidaan sen odotusarvon avulla, jolloin estimoinnin tarkkuuteen vaikuttaa useat aineiston ominaisuudet. Aineiston koko, mitta-asteikot, muuttujien ominaisuudet ja estimoinnin ongelmat, kuten poikkeamat ja harhat, vaikuttavat estimaattorin tehokkuuteen. Siitä huolimatta four-way decomposition antaa selkeämmän arvion altistuksen ja välittäjän sekä niiden vuorovaikutuksen osuudesta kuin pelkän altistuksen vaikutuksen estimointi, jossa välittäjää ei oteta huomioon. Välittäjän vaikutus on merkittävä monissa tutkimuksissa, eikä sitä voi jättää täysin huomiotta. Välittäjä ei itsessään välttämättä ole oleellinen tekijä lopputuloksen kannalta, mutta joissakin aineistossa välitetty vuorovaikutus on hyvin merkittävä osa lopputuloksen syntyä. Toisaalta altistuksen suoran ja epäsuoran vaikutuksen tarkastelu voi joskus osoittaa, ettei välittäjää tarvita lopputuloksen syntymiseen. Tällaisissa tilanteissa välittäjämuuttuja voidaan hylätä, jolloin muodostetut mallinnukset kuvaavat tarkemmin jäljellä olevien muuttujien välisiä suhteita.

Tämän tutkielman esimerkissä kaikki valitut muuttujat eivät vaikuttaneet kovinkaan suuresti lopputulokseen. Välittäjällä, toisin sanoen alkoholin aiheuttamalla syyllisyydentunteella, näyttäisi kuitenkin olevan jonkinlainen yhteys masennukseen, mutta käytetyn tutkimusmateriaalin valossa sitä ei voi määrittää tarkemmin. Altistuksella, eli oppimisvaikeuksilla, vaikuttaisi olevan selvästi yhteyttä masennusdiagnoosin kanssa. Oppimisvaikeuksien vaikutus masennukseen niin suorasti kuin epäsuorastikin on aihe, jota on syytä tutkia enemmän. Tämän tutkielman kannalta ei ole kuitenkaan olennaista tarkastella niiden välistä kausaalisuhdetta tarkemmin. Esimerkin avulla komponenttien muodostamista ja analysointia voitiin kuitenkin tarkastella tehokkaasti. Esimerkki antoi myös selkeän lähtö-

kohdan R-koodin käyttämiseen.

Four-way decomposition on selvästi hyvin tehokas menetelmä altistuksen suoran ja epäsuoran vaikutuksen tutkimiseen. Menetelmää on hyödynnetty erityisesti lääketieteen ja epidemiologian piirissä viime vuosien aikana, mutta jatkossa sen suosio toivottavasti leviää laajemmalle. Hajotusmenetelmä toimii hyvänä työkaluna lääketieteen ja psykologian parissa, mutta se sopii varsin hyvin muidenkin alojen tutkimuksiin. Menetelmän yleistymistä tukisi vahvasti valmiit koodi-paketit useille ohjelmistoille, sillä tällä hetkellä valmiita koodeja on vain muutamaan tilastolliseen ohjelmistoon.

LÄHTEET

Correia, K., Williams, P.L. (2018). Estimating the Relative Excess Risk Due to Interaction in Clustered-Data Settings. *American Journal of Epidemiology*, 187(11), 2470-2480, DOI: 10.1093/aje/kwy154.

Flensborg-Madsen, T. (2011). Alcohol use disorders and depression - the chicken or the egg. *Addiction*, England, 106(5), 916-918, DOI: 10.1111/j.1360-0443.2011.03406.x.

Kunttu, K., Pesonen, T. ja Saari, J. (2016). Korkeakouluopiskelijoiden terveystutkimus 2016. Yhteiskuntatieteellinen tietoarkisto. http://urn.fi/urn:nbn:fi:fsd:T-FSD3224.

Maag, J.W. ja Reid, R. (2006). Depression Among Students with Learning Disabilities: Assessing the Risk. *Journal of Learning Disabilities*, 39(1), 3-10, DOI:10.1177/00222194060390010201.

Mutlu, U., Roshchupkin, G.V., Sajeev, G., Ikram, M.K. ja Ikram, A.K. (2017). A 4-way decomposition, https://github.com/Unalmut/4way-decomposition.

Rothman, K.J. (1986). *Modern Epidemiology* 1. Little, Brown and Company, Boston, MA.

Tchetgen Tchetgen, E.J. ja VanderWeele, T.J. (2014). On identification of natural direct effects when a confounder of the mediator is direct affected by exposure. *Epidemiology*, 25(2), 282-291, DOI: 10.1097/EDE.0000000000000054.

Valeri, L., VanderWeele, T.J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18, 137-150, DOI: 10.1037/a0031034.

VanderWeele, T.J., Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 4(2), 457-468, DOI: https://dx.doi.org/10.4310/SII.2009.v2.n4.a7.

VanderWeele TJ, Vansteelandt S. (2010). Odds ratios for mediation analysis with a dichotomous outcome. *American Journal of Epidemiology*, 172, 1339-1348, https://doi.org/10.1093/aje/kwq332.

VanderWeele, T.J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2), 224-232, DOI: 10.1097/EDE.0b013e318281a64e.

VanderWeele, T.J. (2014a). A unification of mediation and interaction. *Harvard University Biostatistics Working Paper Series*, 164.

VanderWeele, T.J. (2014b). A unification of mediation and interaction: a four-way decomposition. *Epidemology*, 25 (5), 749-761, DOI: 10.1097/EDE.0000000000000121.

VanderWeele, T.J. (2016). Sufficient Cause Representation of the Four-Way Decomposition for Mediation and Interaction. *Epidemology*, 27 (5), e32-e33.

Zoubirm A.M. ja Iskander, D. R. (2007). Bootstrap Methods and Applications. *IEEE Signal Processind Magazine*, 24 (4), 10-19, DOI:10.1109/MSP.2007.4286560.

LIITE: R-KOODI

Four-way decomposition -hajotelman komponenttien estimoimiseen voidaan hyödyntää alla olevaa R-koodia. Koodi on esitetty sen alkuperäisessä muodossa, mutta ohjeet koodin käyttöön on annettu luvussa (6) suomeksi. Koodin pitää sijaita samassa kansiossa kuin siihen liittyvä lähdekoodi (Mutlu et al., 2017).

```
#------------------------------------------------------------------
# All libraries here
library(boot)
library(survival)
library(data.table)
library(foreign)
library(dummies)
library(GenABEL)
library(dummies)
#------------------------------------------------------------------
# Sources import here
# this script should be run from the same folder where src.R is
source('src.R')
#------------------------------------------------------------------
# Define your parameters here!!!

#Data pathway
data_path<-"//Test.sav" #spss/csv/txt
#Path to save results
output<-'//Test_results.csv'

#Define variables
A<<-'A2'
M<<-'M1'
Y<<-'Y1'
COVAR<<-c('C1','C2','C3')

#1=binary 0=continuous
outcome=0
mediator=0
```

```
# Assign levels for the exposure that are being compared; for mstar
# it is the level at which to compute the CDE and the remainder of
# the decomposition
a<<-1
astar<<-0
mstar<<-0

#Boostrap number of iterations
N_r=5

#------------------------------------------------------------------------
####### DONT TOUCH FROM HERE #######
#------------------------------------------------------------------------

# Reading data file
data<-read.spss(data_path, to.data.frame=T) #TODO spss/csv/txt (?)

if (! prod(c(A,Y,M,COVAR) %in% names(data) ) )
   {stop('Some of defined variable names are not in data file!')}
if (mediator==1 & outcome==1)
   {save_results(output=output, boot_function=boot.bMbO, N=N_r)}
if (mediator==0 & outcome==1)
   {save_results(output=output, boot_function=boot.cMbO, N=N_r)}
if (mediator==1 & outcome==0)
   {save_results(output=output, boot_function=boot.bMcO, N=N_r)}
if (mediator==0 & outcome==0)
   {save_results(output=output, boot_function=boot.cMcO, N=N_r)}
```