Propensiteettipistemääriin perustuvat tasapainottavat painokertoimet havainnoivassa tutkimuksessa

Emmi Heinonen

Pro gradu -tutkielma
Toukokuu 2023

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Emmi Heinonen: Propensiteettipistemääriin perustuvat tasapainottavat painokertoimet havainnoivassa tutkimuksessa
Pro gradu -tutkielma, 44 s., 9 liites.
Tilastotiede
Toukokuu 2023

Havainnoivassa tutkimuksessa tarkastellaan tietyn altistuksen vaikutusta vasteeseen ilman että voidaan kontrolloida, mitkä tutkimusyksilöt kohtaavat altistuksen. Koska taustamuuttujat usein vaikuttavat altistuksen toteutumiseen, niiden jakaumat ovat todennäköisesti erilaiset eri altistusryhmissä. Tämän vuoksi havainnoivaa tutkimusta tehtäessä on tärkeää varmistaa, että vastemuuttujassa havaitut erot altistusryhmien välillä eivät johdu taustamuuttujista.

Propensiteettipistemäärä tarkoittaa todennäköisyyttä, että yksilö kohtaa tietyn altistuksen ehdolla taustamuuttujat. Yleisimmin propensiteettipistemääriä käytetään tutkimuksissa, joissa altistusryhmiä on kaksi: toisessa ryhmässä altistuneet ja toisessa altistumattomat. Ne voidaan kuitenkin yleistää myös tilanteisiin, joissa mahdollisia altistuksia on useampia. Propensiteettipistemäärät tiivistävät taustamuuttujien informaation yhteen tai muutamaan lukuun. Altistuksen vaikutusta vasteeseen voidaan tämän vuoksi tutkia ehdollisena propensiteettipistemäärille sen sijaan että ehdollistettaisiin kaikilla taustamuuttujilla.

Tutkielmassa keskitytään erityisesti yhteen propensiteettipistemääriä hyödyntävistä menetelmistä: painokertoimiin. Propensiteettipistemääriin perustuvien painokertoimien avulla voidaan muodostaa pseudoaineisto, jossa taustamuuttujien jakaumat ovat kaikissa altistusryhmissä samanlaiset kuin valitussa kohdejoukossa. Kohdejoukko voi olla koko tutkimusjoukko tai osa siitä. Jos painokertoimen nimittäjä eli toteutuneen ryhmän propensiteettipistemäärä on hyvin pieni, painokerroin voi vastaavasti olla hyvin suuri, jolloin yksi yksilö saattaa vaikuttaa tuloksiin voimakkaasti. Tästä aiheutuvien ongelmien ehkäisyä varten on kehitetty erilaisia menetelmiä.

Painokertoimien avulla voidaan estimoida altistuksen vaikutusta vasteeseen hyödyntäen joko regressiomalleja tai erilaisia parametrittomia estimaattoreita. Menetelmä soveltuu niin jatkuvien, luokallisten kuin tapahtuma-aikaa mittaavien vasteiden tutkimiseen. Tutkielman soveltavassa osassa tarkastellaan äidin raskaudenaikaisen masennuslääkkeiden käytön vaikutusta lapsen myöhempään masennus- tai ahdistuneisuusdiagnoosiin. Tutkimusaineisto perustuu laajoihin viranomaisrekistereihin, ja sitä tarkastellaan sekä kumulatiivisen ilmaantuvuuden että uhkasuhteiden avulla.

Asiasanat: havainnoiva tutkimus, taustamuuttujien jakauma, propensiteettipistemäärä, painokertoimet.

Sisällys

1 Johdanto

Monilla tieteenaloilla hyvin tavallinen tutkimuskysymys on, miten jokin käsittely, tapahtuma tai altistus vaikuttaa myöhemmin havaittavaan vasteeseen. Tutkittavana voi olla esimerkiksi lääkehoidon vaikutus influenssaoireiden häviämiseen, vanhempien koulutustason vaikutus henkilön tuloihin aikuisena tai ruokavalion vaikutus sydänkohtauksen riskiin. Jatkossa sanaa altistus käytetään tarkoittamaan yleisesti hoitoa, elintapaa, ympäristötekijää, sairautta tai muuta tekijää, jonka vaikutusta vasteeseen halutaan tutkia. Altistuksen tai altistumattomuuden perusteella muodostettavia ryhmiä kutsutaan altistusryhmiksi. Tutkimusta tehtäessä on otettava huomioon, että tutkittavan altistuksen lisäksi myös monet muut seikat vaikuttavat vasteeseen ja osa näistä tekijöistä saattaa olla yhteydessä myös altistukseen. Esimerkiksi aiemmat sairaudet saattavat vaikuttaa sekä hoitoon hakeutumiseen että hoidon tehoon.

Usein kokeellista tutkimusta pidetään luotettavimpana tapana tutkia altistuksen vaikutusta vasteeseen. Esimerkiksi lääkkeiden vaikutusta tutkitaan tyypillisesti jakamalla koehenkilöt satunnaisesti kahteen ryhmään, joista toiselle annetaan oikeaa ja toiselle lumelääkettä. Jos koehenkilöiden määrä on tarpeeksi suuri, taustatekijöiden jakauma on satunnaistamalla saaduissa ryhmissä samanlainen. Kaksi ryhmää ovat siis keskimäärin samanlaiset esimerkiksi elintapojen ja geeniperimän suhteen. Ainoa ero on, että toinen ryhmä saa lääkettä ja toinen ei. Lääkkeen vaikutusta voidaan tällöin tutkia vertaamalla näitä kahta ryhmää.

Joissain tilanteissa satunnaistaminen ei kuitenkaan onnistu. Se voisi esimerkiksi tulla liian kalliiksi tai viedä liikaa aikaa. Joskus satunnaistaminen on mahdotonta eettisten syiden takia: tutkimushenkilöitä ei voi määrätä satunnaisesti esimerkiksi tupakoiviin tai tupakoimattomiin. Tällöin havainnoiva tutkimus on ainoa mahdollisuus selvittää käsittelyn vaikutusta. Se tarkoittaa, että tutkijat tarkastelevat eroja altistuneiden ja altistumattomien välillä mutta eivät voi kontrolloida tutkittavien jakautumista näihin ryhmiin. Havainnoivissa tutkimuksissa ongelma onkin, että taustamuuttujat (esimerkiksi ikä, sukupuoli, koulutustausta ja terveydentila) usein ovat yhteydessä altistuksen toteutumiseen. Altistuneet ja altistumattomat eroavat siis toisistaan muutenkin, joten ryhmien suora vertailu ei todennäköisesti anna oikeaa kuvaa altistuksen kausaalisesta vaikutuksesta vasteeseen.

Jotta havainnoivassa tutkimuksessa voitaisiin tehdä oikeita johtopäätöksiä, on ryhmien välinen ero huomioitava jotenkin. Perinteinen tapa on lisätä taustamuuttujat selittävinä tekijöinä regressiomalliin, jolla altistuksen vaikutusta vasteeseen estimoidaan. Jos taustamuuttujia on paljon, voi mallin parametrien estimointi kuitenkin vaikeutua. Yksi vaihtoehtoisista tavoista pienentää sekoittavien tekijöiden vaikutusta on käyttää propensiteettipistemääriä hyödyntäviä menetelmiä. Propensiteettipistemäärällä tarkoitetaan todennäköisyyttä, että yksilö kuuluu tiettyyn altistusryhmään ehdolla taustamuuttujat. Menetelmiä on käytetty jo 1980-luvulla, ja niitä on tutkittu paljon. Propensiteettipistemäärien etu on, että taustamuuttujien vaikutus altistukseen tiivistetään yhteen suureeseen. Tämä helpottaa altistuksen vaikutuksen estimointia, koska esimerkiksi regressiomallissa ei tarvita yhtä paljon parametreja kuin jos kaikki taustamuuttujat olisivat mukana erillisinä selittävinä tekijöinä.

Tässä pro gradu -tutkielmassa esitellään propensiteettipistemäärien käyttöä havainnoivissa tutkimuksissa. Huomiota kiinnitetään erityisesti propensiteettipistemääriin perustuviin painokertoimiin. Aihetta pohjustetaan luvussa 2 esittelemällä kausaalipäättelyyn liittyviä käsitteitä ja propensiteettipistemääriä hyödyntävissä menetelmissä tarvittavat keskeiset oletukset. Käsitteet määritellään aluksi tilanteessa, jossa tutkittavat yksilöt ovat joko altistuneita tai altistumattomia. Tämän jälkeen laajennetaan määritelmiä niin, että altistusryhmiä voi olla useampia kuin kaksi. Kausaalipäättelyn käsitteitä hyödyntäen luvussa 3 määritellään propensiteettipistemäärä sekä kahden että useamman altistusryhmän tapauksessa. Eri tavat, joilla propensiteettipistemääriä voidaan käyttää altistuksen vaikutusta estimoitaessa esitellään lyhyesti.

Luvussa 4 keskitytään tarkemmin propensiteettipistemääriä hyödyntäviin painokertoimiin. Niiden avulla tutkimusaineistosta voidaan muodostaa pseudojoukko, jossa taustamuuttujien jakauma on sama eri altistuksen kohdanneissa ryhmissä. Painokertoimet voidaan myös laskea eri tavoin riippuen siitä, missä kohdejoukossa altistuksen vaikutusta halutaan tutkia. Kausaalivaikutuksen estimointia propensiteettipistemääräpainokertoimien avulla esitellään tässä luvussa sekä lineaarisen tai kaksiluokkaisen vasteen tapauksessa että elinaika-analyysissa. Lisäksi tarkastellaan tapoja, joilla taustamuuttujien jakauman samankaltaisuutta eri altistusryhmissä voidaan arvioida ja joilla mahdollisista poikkeavan suurista painokertoimista johtuvia ongelmia voidaan ehkäistä.

Luvussa 5 on esimerkki propensiteettipistemääriin perustuvien painojen hyödyntämisestä. Rekisteriaineistoihin perustuvassa tutkimuksessa tavoitteena oli selvittää, miten äidin masennuslääkkeiden käyttö raskauden aikana vaikuttaa lapsen riskiin sairastua masennukseen tai ahdistuneisuushäiriöön lapsuus- ja nuoruusiässä. Tutkittavat henkilöt jaettiin kolmeen altistusryhmään, ja sairastumista vertailtiin elinaika-analyysin menetelmien avulla. Painokertoimista käytettiin viittä erilaista versiota, jolloin vaikutusta voitiin tarkastella eri kohdejoukoissa.

2 Kausaalipäättelyn käsitteitä

Tarkastellaan tilannetta, jossa tavoitteena on selvittää, miten altistuminen tietylle tapahtumalle tai käsittelylle vaikuttaa tietyn vastemuuttujan arvoon. Oletetaan, että tutkimusaineisto koostuu n yksilöstä. Yksilön $i = 1, \ldots, n$, saamaa altistusta kuvaa muuttuja Z_i ja saman yksilön vastetta muuttuja Y_i. Lisäksi aineistoon kuuluu p kappaletta taustamuuttujia, joiden arvo määräytyy ennen altistusta. Taustamuuttujia voivat olla esimerkiksi yksilön ikä ja sukupuoli sekä ennen altistusta mitatut tiedot kuten verenpaine. Yksilön i taustamuuttujia kuvaa p-ulotteinen satunnaisvektori $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$. Muuttujien $\boldsymbol{X}_i$ havaituista arvoista käytetään jatkossa merkintää $\boldsymbol{x}_i$.

Usein altistus on kaksiluokkainen eli yksilö joko altistuu tai ei. Tällöin aineisto jakautuu kahteen ryhmään: altistuneet muodostavat käsittelyryhmän ja altistumattomat kontrolliryhmän. Kun altistusmuuttuja Z_i on kaksiluokkainen, merkitään $Z_i = 1$, jos yksilö i kuuluu käsittelyryhmään. Muuten $Z_i = 0$. Tavoitteena on tutkia, miten yksilön i sijoittuminen käsittely- tai kontrolliryhmään vaikuttaa vasteen Y_i saamaan arvoon. Tätä varten määritellään *potentiaaliset vasteet*

$$Y_i(0) = (Y_i | Z_i = 0) \quad \text{ja} \quad Y_i(1) = (Y_i | Z_i = 1).$$

Potentiaalinen vaste $Y_i(0)$ siis tarkoittaa vastemuuttujan Y arvoa, jos yksilö i ei altistu käsittelylle ja $Y_i(1)$ vastaavasti arvoa, jos sama yksilö kuuluu käsittelyryhmään. Käsittelyn kausaalivaikutusta tutkitaan vertailemalla potentiaalisia vasteita. [1]

Oletetaan, että

$$Y_i(1), Y_i(0) \perp\!\!\!\perp Z_j \, (i \neq j)$$

eli että yksilön j sijoittuminen käsittely- tai kontrolliryhmään ei vaikuta toisen yksilön i potentiaalisiin vasteisiin[2]. Kun tämä niin kutsuttu vakausoletus (*stable unit treatment assumption, SUTVA*) pätee, sekä käsittely- että kontrolliryhmässä jokaisella yksilöllä on vain kaksi potentiaalista vastetta [1]. Muuten potentiaaliset vasteet pitäisi määritellä myös muiden yksilöiden altistuksen mukaan. SUTVA-oletus ei päde esimerkiksi rokotetutkimuksissa, koska rokotteen saaneiden määrä vaikuttaa laumasuojan takia myös rokottamattomien sairastumistodennäköisyyteen.

SUTVAn lisäksi käytössä on kaksi tärkeää oletusta: sekoittumattomuus ja positiivisuus. Sekoittumattomuusoletus

$$(Y_i(1), Y_i(0) \perp\!\!\!\perp Z_i) \, | \boldsymbol{X}_i$$

tarkoittaa, että $\boldsymbol{X}_i$ sisältää kaikki taustamuuttujat, jotka aiheuttavat eroja potentiaalisissa vasteissa altistusryhmien välillä. Tällaisia muuttujia kutsutaan sekoittaviksi tekijöiksi (*confounders*), ja niillä on kaksi ominaisuutta: ne voivat vaikuttaa vasteeseen ja niiden jakauma on erilainen eri altistusryhmissä [3]. Sekoittumattomuusoletuksen ollessa voimassa havaitsemattomia sekoittavia tekijöitä ei ole, joten kausaalivaikutusta voidaan tutkia ehdolla havaitut taustamuuttujat. Sen sijaan ei ole välttämätöntä, että kaikki taustamuuttujat $\boldsymbol{X}_i$ ovat sekoittavia tekijöitä. Kutsutaan jatkossa tätä ehtoa vahvaksi sekoittumattomuudeksi (*strong unconfoundedness*). Positiivisuusoletus

$$0 < P(Z_i = 1 | \boldsymbol{X}_i) < 1$$

kaikilla $i = 1, \ldots, n$, puolestaan määrää, että jokaisella yksilöllä on oltava nollasta poikkeava todennäköisyys päätyä sekä käsittely- että kontrolliryhmään ehdolla havaitut taustamuuttujat. Tällöin $\boldsymbol{X}_i$ ei sisällä muuttujia, joiden tietyillä arvoilla yksilöt sijoittuvat aina samaan ryhmään. Jos sekä vahva sekoittumattomuus että positiivisuus ovat voimassa, niin ryhmiin sijoittumisen sanotaan olevan vahvasti valikoitumaton (*strongly ignorable*).[2]

Yleensä yksilö voi kuulua vain jompaan kumpaan ryhmään, joten vain toinen potentiaalisista vasteista havaitaan. Yksilön i havaittu vaste on [4]

$$Y_i = Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

Tavallaan kausaalipäättelyssä onkin kyse puuttuvan aineiston käsittelemisestä, koska useimmiten joko $Y_i(0)$ tai $Y_i(1)$ puuttuu. Yksilökohtainen potentiaalisten vasteiden vertailu ei siis suoraan onnistu. Päätelmiä voidaan kuitenkin tehdä altistuksen vaikutuksesta laajemmassa joukossa.[2]

Jos halutaan tarkastella käsittelyn vaikutusta koko tutkimusjoukossa, estimoidaan *käsittelyn keskimääräistä vaikutusta* (*average treatment effect, ATE*) $E[Y(1) - Y(0)]$. Se kuvaa vasteen odotusarvojen erotusta siinä tapauksessa, että koko tutkimusjoukko kuuluisi käsittelyryhmään, verrattuna siihen, että kaikki kuuluisivat kontrolliryhmään. Toisissa tapauksissa taas halutaan verrata käsittelyryhmään kuuluvan yksilön havaittua vastetta arvoon, jonka vastemuuttuja olisi saanut, jos kyseinen yksilö olisikin kuulunut kontrolliryhmään. Tällöin tutkittavana on käsittelyn keskimääräinen vaikutus käsitellyille (*ATE for the treated, ATT*) $E[Y(1) - Y(0)|Z = 1]$. Vastaavasti voidaan tutkia myös käsittelyn vaikutusta kontrolliryhmässä eli $E[Y(1) - Y(0)|Z = 0]$. Jos käsittelyn vaikutus on sama kaikille yksilöille, niin ATE ja ATT ovat samat. Sen sijaan jos esimerkiksi tietty hoitomenetelmä vaikuttaa yhteen ihmisryhmään tehokkaammin kuin toiseen, voivat $E[Y(1) - Y(0)]$ ja $E[Y(1) - Y(0)|Z = 1]$ olla erilaiset.[4, 5]

2.1 Useamman altistusryhmän tapaus

Vertailtavia altistusryhmiä on usein kaksi, mutta joskus halutaan tutkia eroja useamman ryhmän välillä. Mukana voi olla esimerkiksi kaksi erilaista hoitomuotoa ja yksi kontrolliryhmä. Altistusmuuttuja voi olla myös jatkuva, kuten esimerkiksi lääkkeen annostus, mutta tässä keskitytään vain luokallisiin tapauksiin. Oletetaan, että altistusmuuttuja Z saa arvoja joukossa $\mathcal{Z}$. Kun altistusryhmiä on kaksi, $\mathcal{Z} = \{0, 1\}$. Yleisesti K-luokkaiselle altistusmuuttujalle voidaan merkitä $\mathcal{Z} = \{0, \ldots, K - 1\}$. Samoin kuin kaksiluokkaisen altistuksen tilanteessa, jokaisella yksilöllä $i = 1, \ldots, n$, on jokaista altistusmuuttujan arvoa $z \in \mathcal{Z}$ kohti potentiaalinen vaste $Y_i(z)$. SUTVA-oletuksen ollessa voimassa potentiaalisia vasteita on tässäkin tapauksessa yhtä monta kuin vastemuuttujan mahdollisia arvoja.[6, 7]

Toteutunutta altistusta kuvataan binäärisillä muuttujilla $D_i(z)$ $(z = 0, \ldots, K - 1)$ niin, että

$$D_i(z) = \mathrm{I}(Z_i = z) = \begin{cases} 1, & \text{kun } Z_i = z, \\ 0, & \text{muuten.} \end{cases}$$

Vahva sekoittumattomuus eli oletus, että ei-havaittuja sekoittavia tekijöitä ei ole, voidaan yleistää moniluokkaiselle altistukselle muodossa

$$Z_i \perp\!\!\!\perp \{Y_i(z)\}_{z\in\mathcal{Z}} \,|\boldsymbol{X}_i,$$

mutta voidaan käyttää myös heikkoa sekoittumattomuutta (*weak unconfoundedness*)

$$D_i(z) \perp\!\!\!\perp Y_i(z) \,|\boldsymbol{X}_i,$$

jossa riittää paikallinen riippumattomuus vastemuuttujan ja altistuksen välillä tarkasteltavassa altistusryhmässä sen sijaan että vaadittaisiin altistuksen riippumattomuutta koko potentiaalisten vasteiden joukosta. Positiivisuuden oletetaan olevan voimassa myös moniluokkaiselle altistukselle eli

$$P(Z = z|\boldsymbol{X}) > 0$$

kaikilla $z \in \mathcal{Z}$.[6, 7]

Kun ryhmiä on enemmän kuin kaksi, käsittelyn keskimääräinen vaikutus ATE voidaan estimoida pareittain ryhmien välillä. Esimerkiksi $E[Y(z') - Y(z'')]$ $(z' \neq z'')$ vertaa vastemuuttujan odotusarvoa tilanteessa, jossa kaikki tutkimusjoukon yksiköt kuuluisivat ryhmään z', siihen että ne kaikki kuuluisivat ryhmään z''. Kun ryhmiä on K kappaletta, mahdollisia ATE:n arvoja on $\binom{K}{2}$ kappaletta. Myös ATT:n kaltaisia kausaalivaikutuksia yhden altistusryhmän sisällä voidaan tutkia, mutta silloin on mietittävä tarkkaan mikä ryhmä on 'käsittelyryhmä'. Mahdollisia estimoitavia ovat sekä $E[Y(z') - Y(z'')|Z = z']$ että $E[Y(z') - Y(z'')|Z = z'']$ eli kaksi mahdollisesti paljonkin toisistaan eroavaa odotusarvoa. Erilaisia ryhmäkohtaisia vertailuja on siis $2 \cdot \binom{K}{2}$ kappaletta. Kahden altistusryhmän tapauksessa vaihtoehtoja on kaksi: käsittelyn vaikutus käsittelyryhmässä tai kontrolliryhmässä.[5]

3 Propensiteettipistemäärä

Potentiaalisten vasteiden malli on sama sekä satunnaistetuille kokeille että havainnoiville tutkimuksille. Niiden välinen ero on, että kokeellisessa tutkimuksessa yksilöiden sijoittuminen altistusryhmiin määräytyy satunnaisesti tutkimussuunnitelmassa määritellyn todennäköisyyden mukaan. Taustamuuttujat voivat vaikuttaa todennäköisyyteen, mutta tämä vaikutus tunnetaan. Potentiaalisista vasteista voidaan tällöin tehdä päätelmiä havaittujen vasteiden perusteella. Havainnoivassa tutkimuksessa sen sijaan taustamuuttujien vaikutusta altistusryhmiin sijoittumiseen ei yleensä tunneta eikä taustamuuttujien jakaumaa eri ryhmissä voida kontrolloida. Koska taustamuuttujat voivat vaikuttaa myös vastemuuttujan arvoon, havaittujen vasteiden erot ryhmien välillä voivat johtua pikemminkin taustamuuttujista kuin itse käsittelystä. Esimerkiksi ryhmässä $Z = 1$ havaitut vasteet eivät ole otos potentiaalisen vasteen $Y(1)$ jakaumasta koko joukossa vaan ehdollisesta jakaumasta $f\left(Y(1)|Z = 1\right)$. Niinpä jos vertaillaan satunnaisesti valittuja käsittely- ja kontrolliryhmän yksilöitä, erotuksen odotusarvo $E[Y(1)|Z = 1] - E[Y(0)|Z = 0]$ ei ole sama asia kuin ATE eli $E[Y(1)] - E[Y(0)]$. Sekoittumisen vaara on huomioitava jotenkin, kun aineiston perusteella tehdään päätelmiä, ja *propensiteettipistemäärä* (*propensity score*, engl. propensity = taipumus/alttius) on yksi mahdollinen ratkaisu tähän. Menetelmän esittelivät ensimmäisen kerran Rosenbaum ja Rubin vuonna 1983.[2].

Kahden altistusryhmän tapauksessa propensiteettipistemäärällä $e(\boldsymbol{x})$ tarkoitetaan yksilön todennäköisyyttä päätyä käsittelyryhmään ehdolla havaitut taustamuuttujat eli $e(\boldsymbol{x}) = P(Z = 1|\boldsymbol{X} = \boldsymbol{x})$. Propensiteettipistemäärä tasapainottaa havaittujen taustamuuttujien jakauman niin, että

$$\boldsymbol{X} \perp\!\!\!\perp Z|e(\boldsymbol{X}).$$

Tämä tarkoittaa, että yksilöillä, joilla on sama propensiteettipistemäärä, taustamuuttujien jakauma ei riipu siitä, kuuluvatko ne käsittely- vai kontrolliryhmään. Voidaan osoittaa (ks. [2]), että jos altistusryhmiin sijoittuminen on vahvasti sekoittumaton ehdolla havaitut taustamuuttujat eli $\big(Y(1), Y(0) \perp\!\!\!\perp Z\big)|\boldsymbol{X}$, se on vahvasti sekoittumaton myös ehdolla propensiteettipistemäärä. Tällöin siis $\big(Y(1), Y(0) \perp\!\!\!\perp Z\big)|e(\boldsymbol{x})$, joten

$$\begin{aligned} &E[Y(1)|e(\boldsymbol{x}), Z = 1] - E[Y(0)|e(\boldsymbol{x}), Z = 0] \\ =&E[Y(1)|e(\boldsymbol{x})] - E[Y(0)|e(\boldsymbol{x})]. \end{aligned}$$

Tästä puolestaan seuraa, että

$$\begin{aligned} &E_{e(\boldsymbol{X})}\left[E[Y(1)|e(\boldsymbol{X}), Z = 1] - E[Y(0)|e(\boldsymbol{X}), Z = 0]\right] \\ =&E_{e(\boldsymbol{X})}\left[E[Y(1)|e(\boldsymbol{X})] - E[Y(0)|e(\boldsymbol{X}),]\right] \\ =&E[Y(1) - Y(0)], \end{aligned}$$

jossa $E_{e(\boldsymbol{X})}$ tarkoittaa odotusarvoa propensiteettipistemäärien jakauman suhteen koko populaatiossa.[2]

Havaittiin, että jos vahva sekoittumattomuus on voimassa ja kahdella eri altistusryhmiin kuuluvalla yksilöllä on sama propensiteettipistemäärä, näiden yksi-

löiden havaittujen vasteiden erotuksen odotusarvo on sama kuin käsittelyn keskimääräinen vaikutus eli ATE ehdolla propensiteettipistemäärä. Itse taustamuuttujien arvot voivat olla yksilöillä erilaiset. Propensiteettipistemäärä siis tiivistää taustamuuttujien sekoittavan vaikutuksen yhteen lukuun. Sen sijaan, että vertailtaisiin vastemuuttujan arvoja ehdolla havaitut taustamuuttujat, riittää ehdollistaminen propensiteettipistemäärällä.[2]

Yleistetty propensiteettipistemäärä $e(z, \boldsymbol{x})$ ottaa huomioon myös tilanteet, joissa altistusryhmiä on enemmän kuin kaksi. Se kuvaa todennäköisyyttä

$$P(Z = z|\boldsymbol{X} = \boldsymbol{x}) = E[D(z)|\boldsymbol{X} = \boldsymbol{x}], \quad z \in \mathcal{Z},$$

ja se määriteltiin ensimmäisen kerran Imbensin artikkelissa vuonna 2000 [6]. Luonnollisesti $\sum_{z\in\mathcal{Z}} e(z, \boldsymbol{x}) = 1$ kaikilla $\boldsymbol{x}$, joten jos ryhmiä on K kappaletta ja ryhmien joukko on $\mathcal{Z} = \{0, \ldots, K-1\}$, riittää estimoida yleistetyt propensiteettipistemäärät $e(1, \boldsymbol{x}), \ldots, e(K-1, \boldsymbol{x})$. Kun $K = 2$, niin $e(1, \boldsymbol{x}) = e(\boldsymbol{x})$ ja $e(0, \boldsymbol{x}) = 1 - e(\boldsymbol{x})$.[8]

On syytä huomioida, että vaikka kahdella yksilöllä olisi sama $e(z', \boldsymbol{x})$ jollakin $z' \in \mathcal{Z}$, muut yleistetyt propensiteettipistemäärät $e(z, \boldsymbol{x})\,(z \neq z')$ eivät välttämättä ole samoja. Kun $K = 2$, sama propensiteettipistemäärä $e(1, \boldsymbol{x})$ tarkoittaa automaattisesti, että myös $e(0, \boldsymbol{x}) = 1 - e(1, \boldsymbol{x})$ on sama, mutta tämä ei päde jos $K > 2$. Niinpä jotta vahva sekoittumattomuus ehdolla havaitut taustamuuttujat $\boldsymbol{X}$ säilyisi, täytyy ehdollistaa kaikilla yleistetyillä propensiteettipistemäärillä $e(z, \boldsymbol{X})$. Sekoittumattomuusehto on siis

$$Z \perp\!\!\!\perp \big(Y(0), \ldots, Y(K-1)\big)|\big(e(1, \boldsymbol{X}), \ldots, e(K-1, \boldsymbol{X})\big).$$

Jos ryhmiä on kovin monta, propensiteettipistemäärän dimensiota pienentävä ominaisuus ei näin ollen toimi yhtä hyvin kuin jos $K = 2$. Ei myöskään tunneta skalaarifunktiota $b(\boldsymbol{x})$, jolle pätisi

$$Z \perp\!\!\!\perp \big(Y(0), \ldots, Y(K-1)\big)|b(\boldsymbol{X})$$

aina kun $Z \perp\!\!\!\perp \big(Y(0), \ldots, Y(K-1)\big)|\boldsymbol{X}$. [6, 8, 9]

Sen sijaan kun heikko sekoittumattomuus $D(z) \perp\!\!\!\perp Y(z)|\boldsymbol{X}$ on voimassa, se pätee myös yleistetyille propensiteettipistemäärille eli

$$D(z) \perp\!\!\!\perp Y(z)|e(z, \boldsymbol{X})$$

toteutuu kaikilla $z \in \mathcal{Z}$. Tällöin kaikilla $z \in \mathcal{Z}$

$$E[Y(z)|e(z, \boldsymbol{X})] = E[Y|Z = z, e(Z, \boldsymbol{X})]$$

ja

$$E[Y(z)] = E_{e(z,\boldsymbol{X})}\left[E[Y(z)|e(z, \boldsymbol{X})]\right],$$

jossa $E_{e(z,\boldsymbol{X})}$ tarkoittaa odotusarvoa yleistetyn propensiteettipistemäärän $e(z, \boldsymbol{X})$ suhteen.[6] Heikon sekoittumattomuuden avulla ei siis vertailla yksilöitä, joilla jokin yleistetty propensiteettipistemäärä on sama, vaan estimoidaan erikseen potentiaalisten vasteiden odotusarvoja altistusryhmien sisällä. Käsittelyjen vaikutusten eroja voidaan estimoida vertailemalla näitä odotusarvoja.[8]

Propensiteettipistemäärien tasapainottava vaikutus toteutuu vain, jos sekoittumattomuus- ja positiivisuusoletus ovat voimassa. Näiden oletusten voimassaoloa ei kuitenkaan voida testata havaitusta aineistosta. Siksi tarvitaan hyvää tietämystä tutkittavasta aiheesta, jotta oletusten uskottavuutta voidaan arvioida. Lisäksi on syytä huomioida, että oletusten toteutuminen ja sitä myöten koko menetelmän toimiminen riippuvat taustamuuttujista $\boldsymbol{X}$ ja niiden välisistä suhteista. Jos $\boldsymbol{X}$ on määritelty huonosti, sekoittumattomuus ei toteudu eivätkä $\boldsymbol{X}$:n perusteella lasketut propensiteettipistemäärät tasapainota tutkimusasetelmaa halutulla tavalla. Tämän vuoksi propensiteettipistemääriä hyödyntävien menetelmien toimivuudesta ei voida tehdä yleispäteviä johtopäätöksiä yksittäisten esimerkkien perusteella. Vaikka propensiteettipistemäärien avulla saataisiin yhdessä tilanteessa samanlainen tulos kuin satunnaistetulla aineistolla, se ei kerro mitään saman menetelmän toimimisesta kokonaan toisenlaisessa tapauksessa.[5, 10]

Havainnoivissa tutkimuksissa todennäköisyyttä $P(Z = z|\boldsymbol{X} = \boldsymbol{x})$ ei yleensä tunneta, joten propensiteettipistemäärä täytyy estimoida. Edellä esitetyt tulokset kuitenkin pätevät myös estimoiduille propensiteettipistemäärille, jos estimointi on tehty hyvin. Jatkossa propensiteettipistemäärillä $e(z, \boldsymbol{x})$ tarkoitetaan nimenomaan estimoituja todennäköisyyksiä.

3.1 Altistuksen vaikutuksen estimointi propensiteettipistemäärän perusteella

Kun estimoidaan käsittelyn vaikutusta vasteeseen, voidaan propensiteettipistemäärää käyttää sekoittavien tekijöiden vaikutuksen eliminoimiseen. Menetelmiä on neljä: kaltaistaminen, osittaminen, käyttö selittävänä muuttujana regressiomallissa ja propensiteettipistemääriin perustuvat painokertoimet [4]. Näistä kolme ensimmäistä esiteltiin Rosenbaumin ja Rubinin alkuperäisessä artikkelissa [2]. Painokertoimet Rosenbaum esitteli vuonna 1987 [11].

- Kaltaistamisessa, silloin kun altistusryhmiä on kaksi, jokaista käsittelyryhmän yksilöä kohti valitaan yksi kontrolliryhmän yksilö niin, että niiden propensiteettipistemäärät ovat samat tai lähellä toisiaan. Käsittelyn keskimääräistä vaikutusta koko populaatiossa voidaan estimoida kaikkien kaltaistettujen parien vasteiden erotuksen keskiarvon avulla. Joskus parittaisen kaltaistamisen sijaan valitaan useampi kontrolliryhmän yksilö yhtä käsittelyryhmän yksilöä kohti tai toisin päin.[2]

- Osittaminen tarkoittaa kahden altistusryhmän tilanteessa tutkimusyksilöiden jakamista propensiteettipistemäärän perusteella erillisiin osajoukkoihin niin, että jokaisessa on vähintään yksi käsittely- ja yksi kontrolliryhmän edustaja. Suosittu tapa on jakaa otos viiteen yhtä suureen osaan propensiteettipistemäärän jakauman kvintiilien mukaan. Käsittelyn vaikutusta estimoidaan vertailemalla vastemuuttujan arvoja käsittely- ja kontrolliryhmän yksilöiden välillä osajoukkojen sisällä. [2]

- Propensiteettipistemäärän käyttäminen selittävänä muuttujana tarkoittaa mallin $g\left(E[Y(z)|Z = z, e(z, \boldsymbol{x})]\right) = \alpha_z + \beta_z e(z, \boldsymbol{x})$, $z \in \{0, \ldots, K-1\}$, sovittamista, jossa linkkifunktio g valitaan vastemuuttujan Y mukaan. Malli siis

muodostetaan erikseen eri altistusryhmille, ja ryhmien välisiä eroja tutkitaan regressiokertoimia α ja β vertailemalla. Toinen vaihtoehto on muodostaa yksi malli koko joukolle ja lisätä siihen altistusta kuvaava luokallinen muuttuja. On myös mahdollista pitää mallissa propensiteettipistemäärän lisäksi taustamuuttujat tai osa niistä sellaisenaan. Tämä on neljästä menetelmästä ainoa, joka edellyttää vastemuuttujan ja jonkin selittävän muuttujan (tässä tapauksessa propensiteettipistemäärän) välistä suhdetta kuvaavan regressiomallin määrittelyä. Regressiomallit ovat kuitenkin mahdollisia myös muiden menetelmien yhteydessä.[2, 4]

- Propensiteettipistemäärien avulla muodostettuja painokertoimia käytetään muodostamaan havainnoista synteettinen aineisto, jossa havaittujen taustamuuttujien jakauma on riippumaton altistusryhmään sijoittumisesta [4]. Nämä painokertoimet toimivat samoin kuin kyselytutkimuksissa joskus käytettävät painot, joiden avulla otosta painotetaan niin, että se vastaa paremmin kohdepopulaatiota eikä esimerkiksi tietyn ihmisryhmän muita pienempi vastausprosentti vääristä tuloksia [12]. Propensiteettipistemääriin perustuvia painokertoimia käsitellään tarkemmin luvussa 4.

Kun altistusryhmiä on enemmän kuin kaksi, vahva sekoittumattomuus on voimassa vain ehdollistettuna kaikille yleistetyille propensiteettipistemäärille, mikä on otettava huomioon kaltaistamisessa ja osittamisessa. Jako pareihin ja ryhmiin on joko tehtävä kaikkien yleistettyjen propensiteettipistemäärien perusteella tai sitten on vertailtava vain kahta ryhmää kerrallaan. Jos kohdejoukoksi valitaan vain altistusryhmiin z ja z' kuuluvat, tutkitaan erotusta $E[Y(z) - Y(z')|Z \in \{z, z'\}]$, joka ei välttämättä ole sama asia kuin $E[Y(z) - Y(z')]$ eli käsittelyjen z ja z' vaikutusten ero koko populaatiossa. Niinpä käsittelyvaikutuksia $E[Y(z) - Y(z')|Z \in \{z, z'\}]$ ja $E[Y(z) - Y(z'')|Z \in \{z, z''\}]$ ei voi verrata toisiinsa, koska niissä tarkastellaan eri joukkoja. Kaltaistaminen kaikkien propensiteettipistemäärien perusteella puolestaan voi olla laskennallisesti vaikeaa, erityisesti jos altistusryhmiä on monta [8, 9, 13].

Se, mikä propensiteettipistemääriä hyödyntävistä menetelmistä sopii parhaiten, riippuu aineistosta ja tutkimuskysymyksestä. Joissain tutkimuksissa on havaittu, että kaltaistaminen pienentää taustamuuttujista johtuvaa harhaa altistuksen vaikutuksen estimaateissa enemmän kuin osittaminen tai propensiteettipistemäärän käyttö selittävänä muuttujana. Kaltaistamisen ja painokertoimien paremmuudesta harhan suhteen sen sijaan on ristiriitaisia tuloksia. Kaltaistamisen ja osittamisen etu on, että sen jälkeen kun aineisto on jaettu kaltaistettuihin ryhmiin tai osajoukkoihin, propensiteettipistemäärät eivät suoraan vaikuta kausaalivaikutuksen estimointiin. Tällöin tulosten luotettavuus ei välttämättä kärsi pahasti, vaikka estimoidut propensiteettipistemäärät eivät täysin vastaisikaan todellisia todennäköisyyksiä. Painokertoimia käytettäessä varsinkin suuret painot voivat vaikuttaa vahvasti tuloksiin, joten väärin estimoidut propensiteettipistemäärät saattavat vääristää tuloksia herkemmin. Painotetun aineiston avulla – kuten myös kaltaistamisessa ja osittamisessa – voidaan kuitenkin estimoida kausaalivaikutus suoraan aineistosta eikä regressiomallia välttämättä tarvita. Sitä vastoin jos propensiteettipistemäärää käytetään selittävänä muuttujana, täytyy päätellä millainen sen ja vasteen välinen yhteys on (esimerkiksi lineaarinen vai epälineaarinen). Jos tämä malli valitaan väärin, ovat

tuloksetkin epäluotettavia.[4]

Kaltaistamisessa ongelmaksi saattaa muodostua, että kaikille yksilöille ei löydy toisesta ryhmästä (tai ryhmistä, jos $K > 2$) paria, jonka propensiteettipistemäärä olisi tarpeeksi lähellä. Jos aineistossa on yksi käsittely- ja yksi kontrolliryhmä, olisi kontrolliryhmän hyvä olla selvästi suurempi kuin käsittelyryhmä. Lisäksi propensiteettipistemäärien jakaumien eri ryhmissä pitäisi olla riittävän päällekkäiset. Muuten useakin käsittelyryhmän yksilö saattaa jäädä ilman kaltaistettua paria, joilloin ne joudutaan jättämään pois myöhemmistä analyyseista. Useamman altistusryhmän tapauksessa kaltaistaminen voi olla vielä vaikeampaa, varsinkin jos samaan kaltaistettuun ryhmään halutaan saada yksilö jokaisesta altistusryhmästä. Lisäksi kaltaistaminen kaikkien propensiteettipistemäärien perusteella käy laskennallisesti sitä vaikeammaksi mitä enemmän altistusryhmiä on. Painokertoimien laskeminen monelle ryhmälle on useimmiten helpompaa. [13, 14]

3.2 Propensiteettipistemäärän estimointi

Kokeellisissa tutkimuksissa $P(Z = z|\boldsymbol{X})$ tunnetaan, koska se määritellään tutkimussuunnitelmassa. Havainnoivissa tutkimuksissa tämä ei kuitenkaan yleensä ole mahdollista, ja propensiteettipistemäärät $e(z, \boldsymbol{x})$ ovat vain todennäköisyyden estimaatteja. Toisaalta on havaittu, että silloinkin kun todellinen todennäköisyys $P(Z = z|\boldsymbol{X})$ on tunnettu, on taustamuuttujien tasapainottaminen onnistunut paremmin estimoitujen propensiteettipistemäärien avulla [2, 11].

Kun altistus on K-luokkainen, propensiteettipistemäärät lasketaan usein multinomisen logistisen regression avulla. Tällöin

$$e(0, \boldsymbol{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\beta_{0k} + \boldsymbol{\beta}_k' \boldsymbol{x})}$$

ja

$$e(z, \boldsymbol{x}) = \frac{\exp(\beta_{0z} + \boldsymbol{\beta}_z' \boldsymbol{x})}{1 + \sum_{k=1}^{K-1} \exp(\beta_{0k} + \boldsymbol{\beta}_k' \boldsymbol{x})}$$

kaikilla $z = 1, \ldots, K-1$, jossa $\boldsymbol{\beta}_z = (\beta_{1z}, \ldots, \beta_{pz})'$ ja β_{0z} ovat regressiokertoimet. Kun $K = 2$, malli typistyy tavalliseksi logistiseksi regressioksi. Myös erilaisten koneoppimisen menetelmien, esimerkiksi puupohjaisten menetelmien ja neuroverkkojen, käyttöä propensiteettipistemäärien estimoinnissa on tutkittu.[4, 7]

Taustamuuttujien $\boldsymbol{X}$ valinnasta propensiteettipistemäärät laskevaan malliin ei ole yleisesti hyväksyttyjä kaikenkattavia ohjeita. Selvä sääntö kuitenkin on, että altistus tai käsittely ei saa vaikuttaa taustamuuttujiin. Sen takia $\boldsymbol{X}$ saa sisältää vain muuttujia, joiden arvot havaitaan tai mitataan ennen altistusryhmiin sijoittumista. Jotta heikko sekoittumattomuusoletus $D(z) \perp\!\!\!\perp Y(z) \,|\boldsymbol{X}$ olisi voimassa kaikilla $z = 0, \ldots, K-1$, taustamuuttujien $\boldsymbol{X}$ tulisi sisältää kaikki sekoittavat tekijät. Siksi malliin usein pyritään sisällyttämään paljon muuttujia, jotta sekoittavia tekijöitä ei vahingossa jäisi ulkopuolelle. [4, 5]

Artikkelissa [4] mainitaan neljä tapaa valita muuttujat propensiteettipistemäärien malliin: kaikki havaitut taustamuuttujat, kaikki muuttujat, jotka vaikuttavat altistusryhmiin sijoittumiseen, kaikki muuttujat, jotka vaikuttavat vastemuuttujan

arvoon (mahdolliset sekoittavat tekijät) ja kaikki muuttujat, jotka vaikuttavat sekä altistukseen että vasteeseen (todelliset sekoittavat tekijät). Käytännössä on kuitenkin yleensä vaikea tietää, mitkä tekijät vaikuttavat toisiinsa, joten esimerkiksi mahdollisten tai todellisten sekoittavien tekijöiden joukkoa ei ole helppo määritellä.

On esitetty, että olisi järkevää käyttää propensiteettipistemäärien estimoinnissa kaikkia yksilökohtaisia havaittuja taustamuuttujia, koska usein ne ovat yhteydessä sekä altistukseen että vasteeseen. Esimerkiksi altistusaikaa koskevat muuttujat sen sijaan ovat ongelmallisempia. Jos vaikka yksi hoitomuoto on ollut suosiossa aiemmin ja toinen, tehokkaampi, myöhemmin, voi näyttää siltä että altistuksen ajankohta vaikuttaa sekä käsittelyyn että hoidon tehoon, vaikka todellista kausaaliyhteyttä ei olisikaan.[4]

Tapaa sisällyttää malliin mahdollisimman paljon muuttujia varmuuden vuoksi on myös kritisoitu. Todellisuudessa se ei takaa sekoittumattomuutta, vaan muuttujat olisi valittava tarkastelemalla huolellisesti eri kausaalivaikutuksia. Sitä varten puolestaan tarvitaan hyvää taustatietämystä tutkittavasta aiheesta. Sekoittavia tekijöitä etsittäessä onkin syytä käyttää apuna kirjallisuutta ja alan asiantuntijoiden kokemuksia. Kausaaliyhteyksiä on myös hyvä tutkia laajemminkin kuin pelkästään taustamuuttujien vaikutuksena altistukseen ja vasteeseen. Taustamuuttujien väliset yhteydet saattavat aiheuttaa sen, että ehdollistaminen yhden muuttujan suhteen saa aikaan toisesta, mahdollisesti havaitsemattomasta, muuttujasta johtuvaa epätasapainoa.[10]

4 Propensiteettipistemääriin perustuvat painokertoimet

4.1 Taustamuuttujien jakauma ja kallistusfunktio

Propensiteettipistemäärien perusteella lasketut painokertoimet esiteltiin ensimmäisen kerran vuonna 1987 [11]. Myöhemmin teoriaa on kehitetty eteenpäin ja painokertoimien määritelmää laajennettu. Seuraava määrittely perustuu artikkeleihin [7] ja [15].

Oletetaan, että altistusmuuttuja Z saa arvoja joukossa $\mathcal{Z} = \{0, \ldots, K-1\}$ ja $K \geq 2$. Oletetaan myös, että havaitut taustamuuttujat $\boldsymbol{X}$ noudattavat yhteisjakaumaa, jonka tiheysfunktio on $f(\boldsymbol{x})$. Altistusryhmille $Z = z$ voidaan määritellä omat jakaumansa $f_z(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x}|Z = z)$ niin, että

$$f_z(\boldsymbol{x}) \propto f(\boldsymbol{x})e(z, \boldsymbol{x}).$$

Yleensä altistuksen vaikutusta ei haluta tutkia vain yhdessä pisteessä $\boldsymbol{x}$ kerrallaan, vaan laajemmassa taustamuuttujien joukossa. Tämä kohdejoukko ei välttämättä ole koko populaatio, vaan mahdollisesti tutkimuksen kannalta kiinnostavampi tai tilastollisesti optimaalisempi osa siitä. Tarkasteltavan kohdejoukon mukaan valitaan niin kutsuttu *kallistusfunktio* (*tilting function*) $h(\boldsymbol{x})$. Funktion h tehtävä on muuntaa taustamuuttujien jakauma $f(\boldsymbol{x})$ kohdejoukon jakaumaksi $f(\boldsymbol{x})h(\boldsymbol{x})$. Altistusryhmässä z jakauma $f_z(\boldsymbol{x})$ voidaan tällöin painottaa haluttuun kohdejoukkoon propensiteettipistemäärän ja kallistusfunktion avulla asettamalla painokertoimet

$$w_z(\boldsymbol{x}) \propto \frac{f(\boldsymbol{x})h(\boldsymbol{x})}{f(\boldsymbol{x})e(z, \boldsymbol{x})} = \frac{h(\boldsymbol{x})}{e(z, \boldsymbol{x})}, \quad z = 0, \ldots, K-1.$$

Nämä painokertoimet tasapainottavat taustamuuttujien jakaumia altistusryhmien välillä, koska

$$f_z(\boldsymbol{x})w_z(\boldsymbol{x}) = f(\boldsymbol{x})h(\boldsymbol{x})$$

kaikilla $z \in \mathcal{Z}$. Painotuksen jälkeen yksilöt muodostavat pseudojoukon, joka edustaa haluttua kohdepopulaatiota. Jokaisen yksilön painokerroin määräytyy oman ryhmän propensiteettipistemäärän mukaan. Merkitään yksilön i painokerrointa $w_i \propto h(\boldsymbol{x}_i)/e(z_i, \boldsymbol{x}_i)$, jossa z_i on yksilön toteutunut altistusryhmä.

Kallistusfunktio h voidaan valita useista eri vaihtoehdoista. Jos esimerkiksi kohdepopulaationa on koko tutkimusjoukko, niin $h(\boldsymbol{x}) = 1$ ja $f(\boldsymbol{x})h(\boldsymbol{x}) = f(\boldsymbol{x})$. Tällöin ryhmään z kuuluvan yksilön i painokerroin on $w = 1/e(z, \boldsymbol{x}_i)$ eli yksilön oman ryhmän estimoidun todennäköisyyden käänteisluku. Tällaisia painokertoimia kutsutaankin *käänteistodennäköisyypainokertoimiksi* (*inverse probability weights, IPW* tai *inverse probability of treatment weights, IPTW*). Kohdepopulaatioksi saatetaan valita myös altistuksen z' saaneet yksilöt, jolloin $h(\boldsymbol{x}) = e(z', \boldsymbol{x})$. Tässä tapauksessa kohdejakauma on $f(\boldsymbol{x})e(z', \boldsymbol{x}) \propto f_{z'}(\boldsymbol{x})$. Jokaisen altistusryhmään z' kuuluvan yksilön painokerroin on 1 ja muissa ryhmissä painokerroin $e(z', \boldsymbol{x})/e(z, \boldsymbol{x})$ $(z \neq z')$ on sitä suurempi, mitä suurempi on estimoitu todennäköisyys että kyseinen yksilö saisi altistuksen z' ehdolla havaitut taustamuuttujat. Luvussa 2 määriteltiin ATT eli käsittelyn keskimääräinen vaikutus käsitellyille. Painokertoimia $e(z', \boldsymbol{x})/e(z, \boldsymbol{x})$ voidaankin kutsua ATT-painoiksi, koska niiden avulla tutkitaan odotusarvoa $E[Y(z') - Y(z'')|Z = z']$ eli altistusten z' ja z'' vaikutusten erotusta ryhmässä z'.

4.2 Kausaalivaikutuksen parametriton estimaattori jatkuville vasteille

Kun kallistusfunktio h on valittu ja painokertoimet laskettu, altistuksen vaikutusta vastemuuttujan arvoon voidaan tutkia painotetun aineiston avulla. Tähän voidaan käyttää esimerkiksi sopivaa regressiomallia, kuten lineaarista tai logistista regressiota, jossa altistusryhmä on selittävänä muuttujana. Eroja ryhmien välillä voidaan kuitenkin tarkastella myös suoraan aineistosta käyttämällä parametritonta estimaattoria. Seuraavaksi esiteltävä estimaattori sopii käytettäväksi, jos vaste on jatkuva. Tässä tutkielmassa ei ole esimerkkiä tämän estimaattorin käytöstä, mutta siihen on hyvä tutustua eräänlaisena perustyökaluna. Kaksiluokkaiselle altistukselle parametriton estimaattori esiteltiin ensimmäisen kerran vuonna 2003 artikkelissa [16]. Myöhemmin määritelmää on muokattu sopimaan myös useammalle altistusryhmälle. Seuraava esitys perustuu artikkeliin [7].

Määritellään potentiaalisen vasteen $Y(z)$ odotusarvo m_z^h kallistusfunktion $h(\boldsymbol{x})$ mukaisessa kohdejoukossa niin, että

$$m_z^h = \frac{E_{\boldsymbol{X}}\big[E[Y(z)|\boldsymbol{X}]h(\boldsymbol{X})\big]}{E_{\boldsymbol{X}}[h(\boldsymbol{X})]}.$$

Kausaalivaikutusta kuvaa näiden odotusarvojen lineaarikombinaatio $\tau^h(\boldsymbol{a})$, joka puolestaan määritellään kerroinvektorin $\boldsymbol{a} = (a_0, \ldots, a_{K-1})'$ avulla niin, että

$$\tau^h(\boldsymbol{a}) = \sum_{z=0}^{K-1} a_z m_z^h.$$

Kun altistusmuuttuja on luokallinen, tavoitteena on yleensä vertailla kahta ryhmää toisiinsa. Tällöin $\boldsymbol{a}$ valitaan niin, että yksi alkio on -1, yksi alkio 1 ja loput ovat nollia. Esimerkiksi kun $K = 2$, valitaan $\boldsymbol{a} = (-1, 1)'$, jolloin

$$\tau^h(\boldsymbol{a}) = \frac{E_{\boldsymbol{X}}\left[h(\boldsymbol{X})E[Y(1) - Y(0)|\boldsymbol{X}]\right]}{E_{\boldsymbol{X}}\left[h(\boldsymbol{X})\right]}.$$

Jos käsittely olisi järjestysasteikollinen tai jatkuva, saattaisi olla järkevää valita muun tyyppinen vektori $\boldsymbol{a}$.

Heikon sekoittumattomuusehdon $D(z) \perp\!\!\!\perp Y(z)|\boldsymbol{X}$, jossa $D(z) = \mathrm{I}(Z = z)$, ollessa voimassa $E[Y(z)|\boldsymbol{X}]E[D(z)|\boldsymbol{X}] = E[Y(z)D(z)|\boldsymbol{X}]$. Lisäksi jos yleistetty propensiteettipistemäärä on estimoitu oikein, $E[D(z)/e(z, \boldsymbol{X})|\boldsymbol{X}] = 1$ kaikilla $z = 0, \ldots, K-1$. Näin ollen odotusarvo m_z^h voidaan kirjoittaa muodossa

$$\begin{aligned} m_z^h &= \frac{E_{\boldsymbol{X}}\big[E[Y(z)|\boldsymbol{X}]h(\boldsymbol{X})\big]}{E_{\boldsymbol{X}}[h(\boldsymbol{X})]} \\ &= \frac{E_{\boldsymbol{X}}\big[E[Y(z)D(z)h(\boldsymbol{X})/e(z, \boldsymbol{X})|\boldsymbol{X}]\big]}{E_{\boldsymbol{X}}\big[E[D(z)h(\boldsymbol{X})/e(z, \boldsymbol{X})|\boldsymbol{X}]\big]} \\ &= \frac{E_{\boldsymbol{X}}\big[E[Y(z)D(z)w_z(\boldsymbol{X})|\boldsymbol{X}]\big]}{E_{\boldsymbol{X}}\big[E[D(z)w_z(\boldsymbol{X})|\boldsymbol{X}]\big]}. \end{aligned}$$

Tämän perusteella $\frac{1}{n}\sum_{i=1}^{n} D_i(z)Y_i w_z(\boldsymbol{x}_i)$ on odotusarvon m_z^h osoittajan tarkentuva estimaattori ja $\frac{1}{n}\sum_{i=1}^{n} D_i(z)w_z(\boldsymbol{x}_i)$ vastaavasti nimittäjän. Potentiaalisten vasteiden odotusarvot kohdejoukossa voidaan siis estimoida painokertoimien avulla:

$$\hat{m}_z^h = \frac{\sum_{i=1}^{n} D_i(z)Y_i w_i}{\sum_{i=1}^{n} D_i(z)w_i}.$$

Tällöin kausaalivaikutukselle $\tau^h(\boldsymbol{a}) = \sum_{z=0}^{K-1} a_z m_z^h$ puolestaan saadaan tarkentuva estimaattori $\hat{\tau}^h(\boldsymbol{a}) = \sum_{z=0}^{K-1} a_z \hat{m}_z^h$. Luokallisen altistusmuuttujan tapauksessa verrataan kahta ryhmää, z' ja z'', jolloin $\hat{\tau}^h(\boldsymbol{a}) = \hat{m}_{z'}^h - \hat{m}_{z''}^h$.

Merkitään tutkimusyksilöiden altistusryhmiä $\underline{\boldsymbol{Z}} = \{Z_1, \ldots, Z_n\}$ ja taustamuuttujia $\underline{\boldsymbol{X}} = \{\boldsymbol{X}_i, \ldots, \boldsymbol{X}_n\}$. Kausaalivaikutuksen estimaattorin varianssi voidaan hajottaa

$$\mathrm{Var}[\hat{\tau}^h(\boldsymbol{a})] = E_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\big[\mathrm{Var}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}]\big] + \mathrm{Var}_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\big[E[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}]\big].$$

Näistä kahdesta termistä ensimmäinen, ehdollisen varianssin odotusarvo, edustaa estimaattorin vaihtelua ehdolla havaitut taustamuuttujat ja altistusluokat. Ehdollisen odotusarvon varianssi puolestaan kuvaa sitä, kuinka paljon otos vaikuttaa estimaattorin odotusarvoon. Tämän termin estimointi vaatii potentiaalisen vasteen $Y(z)$ ja taustamuuttujien välisen yhteyden tarkastelua. Ensimmäisen termin edustama yksilöllinen varianssi on kuitenkin usein suurempi kuin ehdollisen odotusarvon varianssi, joten voidaan keskittyä vain ensimmäiseen termiin. Voidaan osoittaa (ks. artikkelin [7] liite), että kun $n \to \infty$, niin ehdollisen varianssin odotusarvo konvergoituu

$$n \cdot E_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\big[\mathrm{Var}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}]\big] \to E_{\boldsymbol{X}}\left[\Big(\sum_{z=0}^{T-1} a_z^2 \mathrm{Var}[Y(z)|\boldsymbol{X}]/e(z,\boldsymbol{X})\Big)h^2(\boldsymbol{X})\right]/E[h(\boldsymbol{X})]^2.$$

Jos homoskedastisuus on voimassa eli $\mathrm{Var}[Y(z)|\boldsymbol{X}] = v$ kaikilla $z = 0, \ldots, K-1$, niin raja-arvo yksinkertaistuu vielä lisää.

4.3 Painokertoimien käyttö elinaika-analyysissa

Edellä esitellyssä kausaalivaikutuksen estimaattorissa pyritään vertailemaan jatkuvan vastemuuttujan odotusarvojen erotusta eri altistusten välillä. Kaikkiin tilanteisiin tämä menetelmä ei kuitenkaan sovi. Elinaika-analyysissa vasteena on tapahtuma-aika eli aika joka kuluu seurannan alusta siihen, että yksilö kohtaa päätetetapahtuman. Päätetapahtuma voi olla esimerkiksi kuolema, sairastuminen tai parantuminen. Päätetapahtuman kohtaamisaikaa merkitään muuttujalla T. Potentiaalisten vasteiden voidaan ajatella olevan yksilön tapahtuma-aikoja eri altistuksilla eli $T(z)$, $z = 0, \ldots, K-1$. Sekoittumattomuusoletuksen on oltava voimassa myös tässä asetelmassa eli potentiaalisten tapahtuma-aikojen ja altistusryhmiin sijoittumisen oletetaan olevan riippumattomia ehdolla havaitut taustamuuttujat.[14, 17]

Yleensä ollaan kuitenkin kiinnostuneempia eroista päätetapahtuman todennäköisyydessä tietyllä aikavälillä. Tarkastellaan siis todennäköisyyttä, että yksilö niin sanotusti selviytyy eli ei kohtaa päätetapahtumaa tiettyyn hetkeen t mennessä. Estimoitavana on välttöfunktio $S(t) = P(T > t)$, jota usein tarkastellaan kuvaajan eli

välttökäyrän kautta. Sen sijaan että vertailtaisiin suoraan tapahtuma-aikojen odotusarvoja, voi olla järkevämpää ajatella potentiaalisia vasteita potentiaalisina välttökäyrinä. Ne kuvaavat välttöfunktiota joukoissa, jotka edustavat eri altistusryhmiä, mutta ovat taustamuuttujien suhteen samanlaisia. Käyrien estimoinnissa voidaan hyödyntää propensiteettipistemääriin perustuvia painokertoimia. Kohdejoukko, jossa selviytymisen todennäköisyyttä vertaillaan, määrää mitä kallistusfunktiota h painokertoimissa käytetään.[14]

Elinaika-analyysissa on otettava huomioon myös, että yleensä kaikkien tutkittavien kohdalla ei tunneta päätetapahtuman aikaa, koska heistä ei ole tietoja tietyn ajan jälkeen. Yksilö on saattanut jäädä pois tutkimuksesta omasta halustaan tai muuttaa toisaalle, jolloin tietoja ei ole enää saatavilla. Tietojen keruu tutkimusta varten on myös saattanut päättyä ennen päätetapahtumaa. Jos päätetapahtuma on esimerkiksi sairastuminen, kaikki yksilöt eivät välttämättä edes kohtaa sitä koko elämänsä aikana, mutta tätä ei voida tutkimusaineiston perusteella tietää varmasti. Tiedetään vain, mihin aikaan mennessä yksilö ei ollut kohdannut päätetapahtumaa. Näitä havaintoja kutsutaan sensuroituneiksi, ja niiden havaittu aika T tarkoittaa sensuroitumisaikaa.

4.3.1 Painotettu Kaplan-Meierin estimaattori ja logrank-testi

Yleinen tapa estimoida välttöfunktiota on Kaplan-Meierin estimaattori. Painotetulla aineistolla käytetään muokattua Kaplan-Meierin estimaattoria, joka esitellään artikkelissa [18] seuraavasti: Määritellään δ_i niin, että $\delta_i = 1$, jos yksilö i on kohdannut päätetapahtuman, ja $\delta_i = 0$, jos yksilö i on sensuroitunut. Oletetaan että tapahtumat ilmaantuvat H eri ajanhetkillä $t_1 < t_2 < \cdots < t_H$. Nyt voidaan muodostaa painotettu tapahtumien määrä altistusryhmässä z hetkellä t_j

$$d_{jz}^h = \sum_{i:T_i=t_j} w_i \delta_i D_i(z).$$

Tarvitaan myös riskijoukon koko, eli niiden ryhmään z kuuluvien yksilöiden määrä, jotka eivät ole kohdanneet päätetapahtumaa tai sensuroituneet ennen hetkeä t_j. Painotettu riskijoukon koko on

$$Y_{jz}^h = \sum_{i:T_i \geq t_j} w_i D_i(z).$$

Altistusryhmäkohtaista välttöfunktiota $S(z,t) = P(T > t|Z = z)$ voidaan nyt estimoida painotetun Kaplan-Meierin estimaattorin

$$\hat{S}^h(z,t) = \begin{cases} 1, & \text{jos } t < t_1, \\ \prod_{t_j \leq t}(1 - \frac{d_{jz}^h}{Y_{jz}^h}), & \text{jos } t_1 \leq t \end{cases}$$

avulla.

Joskus tarkastellaan päätetapahtuman välttämisen todennäköisyyden sijaan mieluummin kumulatiivista ilmaantuvuutta eli todennäköisyyttä, että päätetapahtuma kohdataan viimeistään hetkellä t. Ryhmäkohtainen kumulatiivinen ilmaantuvuus on siis $F(z,t) = P(T \leq t|Z = z) = 1 - S(z,t)$, ja sen estimaattori on $\hat{F}^h(z,t) = 1 - \hat{S}^h(z,t)$.

Välttöfunktion ja kumulatiivisen ilmaantuvuuden perusteella voidaan tarkastella ryhmien välisiä eroja eri hetkillä. Estimoiduista funktioista voidaan piirtää myös kuvaajia, joista nähdään päätetapahtuman välttämistodennäköisyyden tai kumulatiivisen ilmaantuvuuden kulku koko seuranta-ajalta. Pelkistä kuvaajista ei kuitenkaan voida päätellä, kuinka paljon tilastollista epävarmuutta ryhmien välillä havaittuihin eroihin liittyy. Kaplan-Meierin kuvaajien visuaalisen tarkastelun lisäksi yleensä halutaan myös testata ryhmien välisiä eroja. Nollahypoteesina voi olla, että kaikkien ryhmien välttöfunktioiden arvot ovat samat kaikkina seurannan hetkinä eli

$$H_0: \quad S(0,t) = S(1,t) = \cdots = S(K-1,t) \quad \text{kaikilla } t \leq \tau,$$

jossa τ on suurin aika, jolloin kaikissa ryhmissä on vähintään yksi yksilö riskijoukossa. Vastahypoteesi on tällöin

$$H_1: \quad S(z',t) \neq (z'',t) \quad \text{jollakin } t \leq \tau \text{ ja } z' \neq z''.$$

Kaplan-Meierin estimaattien samankaltaisuuden testaamiseen käytetään usein logrank-testiä. Painotetulla aineistolla käytetään siitä muokattua versiota. Olkoon $d_j^h = \sum_{z=0}^{K-1} d_{jz}^h$ painotettu päätetapahtumien määrä hetkellä t_j kaikissa ryhmissä ja $Y_j^h = \sum_{z=0}^{K-1} Y_{jz}^h$ painotettu riskijoukon koko hetkellä t_j koko joukossa. Vastaavasti $d_j = \sum_{z=0}^{K-1} \sum_{i:T_i=t_j} \delta_i D_i(z)$ ja $Y_j = \sum_{z=0}^{K-1} \sum_{i:T_i \geq t_j} D_i(z)$ ovat päätetapahtumien ja riskijoukossa olevien painottamattomat määrät koko joukossa hetkellä t_j. Kallistusfunktion h mukaisesti painotettua logrank-testiä varten lasketaan jokaiselle altistusryhmälle z

$$G^h(z) = \sum_{j=1}^{H} \left(d_{jz}^h - Y_{jz}^h \frac{d_j^h}{Y_j^h} \right), \quad z = 0, \ldots, K-1.$$

Jos välttöfunktion arvo $P(T > t|Z = z)$ on sama kaikilla $z = 0, \ldots, K-1$, päätetapahtumien määrän odotusarvo kussakin altistusryhmässä hetkellä t riippuu päätetapahtumien kokonaismäärästä kyseisellä hetkellä ja ryhmäkohtaisen riskijoukon koosta Y_{jz}^h suhteessa koko riskijoukkoon. Kukin summan $G^h(z)$ termi siis vertaa ryhmässä z toteutunutta painotettua päätetapahtumien määrää hetkellä t_j nollahypoteesin vallitessa laskettuun odotusarvoon. Altistusryhmäkohtaiset testisuureet kootaan vektoriin $\boldsymbol{G}^h = [G^h(0), \ldots, G^h(K-1)]'$. Vektorin $\boldsymbol{G}^h$ kovarianssimatriisi on $\mathrm{Var}(\boldsymbol{G}^h)$, joka sisältää elementit

$$\begin{aligned} \mathrm{Var}(G^h(z)) = \sum_{j=1}^{H} \Bigg\{ & \frac{d_j(Y_j - d_j)}{Y_j(Y_j - 1)} \\ & \times \sum_{i:T_i \geq t_j} \left[\left(\frac{Y_{jz}^h}{Y_j^h} \right)^2 w_i^2 I(Z_i \neq z) + \left(\frac{Y_j^h - Y_{jz}^h}{Y_j^h} \right)^2 w_i^2 I(Z_i = z) \right] \Bigg\} \end{aligned}$$

ja

$$\begin{aligned}\mathrm{Cov}(G^h(z'), G^h(z'')) = \sum_{j=1}^{H} \Bigg\{ & \frac{d_j(Y_j - d_j)}{Y_j(Y_j-1)} \\ & \times \sum_{i:T_i \geq t_j} \left[\left(1 - \frac{Y^h_{jz'}}{Y^h_j} \right) w_i I(Z_i = z') - \frac{Y^h_{jz'}}{Y^h_j} w_i I(Z_i \neq z') \right] \\ & \times \left[\left(1 - \frac{Y^h_{jz''}}{Y^h_j} \right) w_i I(Z_i = z'') - \frac{Y^h_{jz''}}{Y^h_j} w_i I(Z_i \neq z'') \right] \Bigg\}.\end{aligned}$$

Varsinainen testisuure on $(\boldsymbol{G}^h)'(\mathrm{Var}(\boldsymbol{G}^h))^- \boldsymbol{G}^h$, jossa $(\mathrm{Var}(\boldsymbol{G}^h))^-$ on kovarianssimatriisin yleistetty käänteismatriisi eli matriisi, jolle pätee $\mathrm{Var}(\boldsymbol{G}^h)(\mathrm{Var}(\boldsymbol{G}^h))^- \mathrm{Var}(\boldsymbol{G}^h) = \mathrm{Var}(\boldsymbol{G}^h)$. Nollahypoteesin ollessa voimassa testisuure noudataa χ^2-jakaumaa vapausastein $K-1$. [18, 19]

Kun altistusryhmiä on enemmän kuin kaksi, logrank-testi vertaa elossaolofunktion samankaltaisuutta kaikissa ryhmissä. Vain kahta ryhmää voidaan vertailla kontrastivektorin $\boldsymbol{c} = (c_0, c_1, \ldots, c_{K-1})'$ avulla. Vektori $\boldsymbol{c}$ määritellään niin, että $\sum_{z=0}^{K-1} c_z = 0$. Jos esimerkiksi $K = 4$ ja verrataan ryhmää 2 ryhmään 0, valitaan $\boldsymbol{c} = (-1, 0, 1, 0)$. Testisuure on tällöin

$$(\boldsymbol{c}'\boldsymbol{G}^h)'(\boldsymbol{c}'\mathrm{Var}(\boldsymbol{G}^h)\boldsymbol{c})^-(\boldsymbol{c}'\boldsymbol{G}^h) = \frac{(\boldsymbol{c}'\boldsymbol{G}^h)^2}{\boldsymbol{c}'\mathrm{Var}(\boldsymbol{G}^h)\boldsymbol{c}},$$

ja se noudattaa χ^2-jakaumaa vapausastein 1.[19]

Painotettu tapahtumien määrä ryhmässä z' voidaan ilmaista muiden ryhmien tapahtumien määrän avulla $d^h_{jz'} = d^h_j - \sum_{z \neq z'} d^h_{jz}$ ja vastaavasti $Y^h_{jz'} = Y^h_j - \sum_{z \neq z'} Y^h_{jz}$. Niinpä toteutuneiden ja odotettujen tapahtumien määrän ero $G^h(z')$ on

$$\begin{aligned} G^h(z') &= \sum_{j=1}^{H} \left((d^h_j - \sum_{z \neq z'} d^h_{jz}) - (Y^h_j - \sum_{z \neq z'} Y^h_{jz}) \frac{d^h_j}{Y^h_j} \right) \\ &= \sum_{j=1}^{H} \left(-\sum_{z \neq z'} d^h_{jz} + \sum_{z \neq z'} Y^h_{jz} \frac{d^h_j}{Y^h_j} \right) \\ &= -\sum_{z \neq z'} \sum_{j=1}^{H} \left(d^h_{jz} - Y^h_{jz} \frac{d^h_j}{Y^h_j} \right) \\ &= -\sum_{z \neq z'} G^h(z). \end{aligned}$$

Varianssin ja kovarianssin laskusääntöjen perusteella tästä seuraa, että

$$\mathrm{Var}(G^h(z')) = \sum_{z \neq z'} \mathrm{Var}(G^h(z)) + \sum_{z^k \neq z^l; z^k, z^l \neq z'} \mathrm{Cov}(G^h(z^k), G^h(z^l))$$

ja

$$\mathrm{Cov}(G^h(z'), G^h(z^k)) = -\left(\mathrm{Var}(G^h(z^k)) + \sum_{z \notin \{z', z^k\}} \mathrm{Cov}(G^h(z^k), G^h(z)) \right),\ z^k \neq z'.$$

Erityisesti tilanteessa, jossa altistusryhmiä on kaksi, $G^h(0) = -G^h(1)$, $\mathrm{Var}(G^h(0)) = \mathrm{Var}(G^h(1))$ ja $\mathrm{Cov}(G^h(0), G^h(1)) = -\mathrm{Var}(G^h(1))$. Nyt yleistetyksi käänteismatriisiksi voidaan valita $(\mathrm{Var}(\boldsymbol{G}^h))^- = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1/\mathrm{Var}(G^h(1)) \end{smallmatrix}\right]$, jolloin testisuureeksi saadaan

$$\begin{bmatrix} G^h(0) & G^h(1) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1/\mathrm{Var}(G^h(1)) \end{bmatrix} \begin{bmatrix} G^h(0) \\ G^h(1) \end{bmatrix} = \frac{(G^h(1))^2}{\mathrm{Var}(G^h(1))}.$$

Testisuure noudattaa χ^2-jakaumaa vapausastein 1, joten voidaan yhtä hyvin sanoa, että $G^h(1)/\sqrt{\mathrm{Var}(G^h(1))}$ noudattaa standardinormaalijakaumaa nollahypoteesin ollessa voimassa. Tämä tulos toki pätee logrank-testille myös painottamattoman aineiston kanssa.

4.3.2 Verrannollisten uhkien malli painotetulla aineistolla

Päätetapahtuman välttämisen todennäköisyyden lisäksi voidaan tarkastella myös uhkafunktiota eli hasardia. Uhkafunktio

$$\lambda(t) = \lim_{h \to 0^+} \frac{1}{h} P(t \leq T < t + h | T \geq t)$$

kuvaa nopeutta, jolla päätetapahtuman todennäköisyys kasvaa hetkellä t niiden yksilöiden joukossa, jotka eivät ole kohdanneet päätetapahtumaa ennen sitä. Uhkafunktion arvo voi olla suurempi kuin yksi, joten se ei tarkoita päätetapahtuman todennäköisyyttä. Uhkafunktio voidaan ilmaista myös välttöfunktion $S(t)$ avulla niin, että $\lambda(t) = -S'(t)/S(t)$, jossa $S'(t)$ on välttöfunktion derivaatta.[20]

Tavallisesti ei olla kiinnostuneita itse uhkafunktion arvoista vaan niiden suhteellisista eroista eri ryhmien välillä. Altistuksen vaikutusta tutkittaessa voidaan määritellä ryhmien z' ja z'' välinen uhkasuhde (*hazard ratio, HR*) $\lambda_{z'}(t)/\lambda_{z''}(t)$. Verrannollisten uhkien mallissa oletetaan, että kahden ryhmän välinen uhkasuhde on vakio eli $\lambda_{z'}(t)/\lambda_{z''}(t) = \theta$ kaikilla $t \geq 0$. Ryhmän $Z = 0$ uhkafunktiota merkitään $\lambda_0(t)$. Verrannollisten uhkien mallin mukaan muiden ryhmien uhkafunktiot voidaan nyt määritellä $\lambda_z(t) = \lambda_0(t)\theta_z$. Ryhmien $Z = z$ ja $Z = 0$ välinen uhkasuhde on siis θ_z. Mallin yhtälöissä on välillä miellyttävämpää käyttää logaritmista uhkasuhdetta $\beta_z = \log(\theta_z)$. Verrannollisten uhkien mallissa estimoidaan vain uhkasuhteita eikä uhkafunktion perustasosta $\lambda_0(t)$ tai sen jakaumasta tehdä oletuksia, mikä onkin yksi syy mallin suosioon. Sitä kutsutaan myös Coxin malliksi kehittäjänsä mukaan.[20, 21]

Yksilölle i määritellään $\boldsymbol{D}_i = (D_i(1), \ldots, D_i(K-1))'$, jossa $D_i(z) = 1$, jos yksilö i kuuluu altistusryhmään z, ja muuten $D_i(z) = 0$. Logaritmisista uhkasuhteista puolestaan muodostetaan vektori $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K-1})'$. Painotetussa verrannollisten uhkien mallissa osittaisuskottavuusfunktio on

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \left(\frac{e^{\boldsymbol{\beta}' \boldsymbol{D}_i}}{\sum_{j:T_j \geq T_i} w_j e^{\boldsymbol{\beta}' \boldsymbol{D}_j}} \right)^{w_i}.$$

Sitä kutsutaan osittaisuskottavuusfunktioksi, koska tulo otetaan vain päätetapahtuman kohdanneiden joukossa ($\delta_i = 1$). Sensuroituneet yksilöt ovat mukana vain

riskijoukon kautta. Logaritminen osittaisuskottavuusfunktio on

$$l(\boldsymbol{\beta}) = \sum_{i:\delta_i=1} w_i \left(\boldsymbol{\beta}' \boldsymbol{D}_i - \log \sum_{j:T_j \geq T_i} w_j e^{\boldsymbol{\beta}' \boldsymbol{D}_j} \right).$$

Derivoimalla se parametrin β_z suhteen saadaan

$$l'(\beta_z) = \sum_{i:\delta_i=1} w_i \left(D_i(z) - \frac{\log \sum_{j:T_j \geq T_i} w_j D_j(z) e^{\boldsymbol{\beta}' \boldsymbol{D}_j}}{\log \sum_{j:T_j \geq T_i} w_j e^{\boldsymbol{\beta}' \boldsymbol{D}_j}} \right).$$

Ratkaisemalla derivaatan nollakohta saadaan kyseisen parametrin estimaatti $\hat{\beta}_z$. Uhkasuhteen estimaatti puolestaan on $\hat{\theta}_z = \exp(\hat{\beta}_z)$.[17, 22]

Parametrien estimointiin painotetulla aineistolla liittyy enemmän epävarmuutta kuin ilman painokertoimia. Jotta tämä epävarmuus tulisi huomioiduksi, suositellaan usein vakaan ("robustin") sandwich-estimaattorin käyttöä verrannollisten uhkien mallin parametrien varianssin estimoinnissa. Vakaata sandwich-estimaattoria käytetään myös silloin, kun tutkittavien yksilöiden välillä esiintyy korrelaatiota. Painotetulle aineistolle estimaattori esitellään artikkelissa [21]. Toisaalta on esitetty, että vakaa sandwich-estimaattori yliarvioi varianssia herkästi. Estimaattorista onkin muodostettu mm. bootstrap-menetelmän avulla erilaisia muokattuja versioita, joissa tätä virhettä on pyritty korjaamaan. Tässä työssä ei kuitenkaan syvennytä näihin muunnoksiin.[14, 17, 22]

Verrannollisten uhkien mallin painotettu versio estimoi marginaalista eli populaatiotason vaikutusta. Uhkasuhteen estimaatti kuvaa siis keskimääräistä eroa koko kohdejoukossa kahden tilanteen välillä: toisessa kaikki kohdejoukon yksilöt kuuluvat yhteen altistusryhmään ja toisessa toiseen. Tässäkin tapauksessa valittu kohdejoukko määrää, mitä kallistusfunktiota käytetään painokertoimia laskettaessa.[14]

4.4 Painokertoimet, kun positiivisuusoletus ei päde

Riippumatta altistusryhmien määrästä K oletetaan positiivisuusoletuksen $e(z, \boldsymbol{x}) > 0$, $z = 0, \ldots, K-1$, olevan voimassa kaikilla $\boldsymbol{x}$. Joskus tämä oletus ei kuitenkaan toteudu täysin, vaan $e(z, \boldsymbol{x}) \approx 0$ jollekin altistusryhmälle z joillakin taustamuuttujien arvoilla $\boldsymbol{x}$. Tämä voi johtua esimerkiksi pienestä otoskoosta tai propensiteettipistemäärien mallin epäsopivasta määrittelystä. Kyse voi olla myös siitä, että jotkin taustamuuttujien arvot määräävät yksilön lähes deterministisesti tiettyyn altistusryhmään. Esimerkiksi lääketieteellisissä sovelluksissa tietyt potilaan ominaisuudet, kuten ikä tai terveydentilaan liittyvät tekijät, voivat vaikuttaa erittäin voimakkaasti päätökseen toimenpiteen suorittamisesta tai lääkkeen tarjoamisesta. Tällaisissa tapauksissa käsittelyn vaikutuksen estimoinnille on vähemmän tarvetta kuin jos yksiköllä olisi todellinen mahdollisuus päätyä mihin tahansa ryhmään.[23]

Aiemmin todettiin, että jatkuvan vasteen tapauksessa kausaalivaikutuksen parametrittoman estimaattorin varianssiin vaikuttavat yleistettyjen propensiteettipistemäärien käänteisluvut. Varianssi siis on suuri, jos jokin $e(z, \boldsymbol{x})$ on hyvin pieni. Kallistusfunktion valinnalla voidaan kuitenkin vähentää pienien propensiteettipistemäärien vaikutusta. Yksi vaihtoehto on propensiteettipistemäärien katkaisu (*trimming*) niin että kohdejoukkoon valikoituvat vain ne yksilöt, joilla positiivisuusoletus

toteutuu tarpeeksi hyvin. Kun altistusluokkia on kaksi, voidaan valita poistettavaksi esimerkiksi ne yksilöt i, joilla $e(\boldsymbol{x}_i) < 0,1$ tai $e(\boldsymbol{x}_i) > 0,9$. Moniluokkaisessa tapauksessa voidaan määritellä $h(\boldsymbol{x}) = \mathrm{I}(\boldsymbol{x} \in \mathcal{C})$, jossa I on indikaattorifunktio. Joukko $\mathcal{C}$ kannattaa valita siten, että luvun 4.2 parametrittoman estimaattorin varianssi on mahdollisimman pieni. Jos homoskedastisuus $\mathrm{Var}[Y(z)|\boldsymbol{X}] = v$ kaikilla $z = 0, \ldots, K-1$, on voimassa, tämä toteutuu, kun $\mathcal{C} = \{\boldsymbol{x} | \sum_{z=0}^{K-1} 1/e(z, \boldsymbol{x}) \leq \alpha\}$ [8]. Ehdossa α on suurin arvo, jolle pätee

$$\alpha \leq \frac{2 \cdot E[\sum_{z=0}^{K-1} 1/e(z, \boldsymbol{X}) | \sum_{z=0}^{K-1} 1/e(z, \boldsymbol{X}) \leq \alpha]}{P\left(\sum_{z=0}^{K-1} 1/e(z, \boldsymbol{X}) \leq \alpha\right)}.$$

Kun tätä optimaalista katkaisumenetelmää sovelletaan käytännössä, korvataan odotusarvo keskiarvolla. Kallistusfunktio $h(\boldsymbol{x}) = \mathrm{I}(\boldsymbol{x} \in \mathcal{C})$ muuttaa painokertoimet nolliksi niillä yksilöillä, joilla yksi tai useampi propensiteettipistemäärä on pieni. Tämä johtaa sekä suurimpien että pienimpien painokertoimien poistumiseen.[7, 8]

Propensiteettipistemäärien katkaisussa ongelma on, että yksilöitä ja niiden myötä informaatiota voidaan menettää paljonkin. Huonosti toteutuvan positiivisuusehdon aiheuttamia ongelmia voidaan helpottaa myös poistamatta yksilöitä kohdejoukosta kokonaan. Voidaan osoittaa (ks. artikkelin [7] liite), että homoskedastisuuden ollessa voimassa luvun 4.2 estimaattorin $\hat{\tau}^h(\boldsymbol{a})$ asymptoottinen varianssi on pienin, kun

$$h(\boldsymbol{x}) \propto \frac{1}{\sum_{z=0}^{K-1} a_z^2 / e(z, \boldsymbol{x})}.$$

Altistuksen ollessa luokallinen tavoitteena on yleensä vertailla altistuksia parittain. Tällöin on hyödyllistä valita kallistusfunktio h siten että kaikkien parittaisten vertailujen estimaattoreiden kokonaisvarianssi on mahdollisimman pieni. Sitä varten valitaan $\boldsymbol{a} = \mathbf{1}_K = (1, \ldots, 1)'$, jolloin edellisestä tuloksesta saadaan kallistusfunktioksi $h(\boldsymbol{x}) = \left(\sum_{z=0}^{K-1} 1/e(z, \boldsymbol{x})\right)^{-1}$ eli yleistettyjen propensiteettipistemäärien harmoninen keskiarvo. Se saa suurimman arvonsa, kun $e(z, \boldsymbol{x}) = 1/K$ kaikilla $z = 0, \ldots, K-1$, eli kun kaikki altistusryhmät ovat taustamuuttujien perusteella yhtä todennäköisiä. Näin saadaan niin kutsutut *päällekkäisyyspainokertoimet* (*overlap weights, OW*)

$$w_z(\boldsymbol{x}) \propto \frac{1/e(z, \boldsymbol{x})}{\sum_{z=0}^{K-1} 1/e(z, \boldsymbol{x})}.$$

Kun $K = 2$, kallistusfunktio pelkistyy

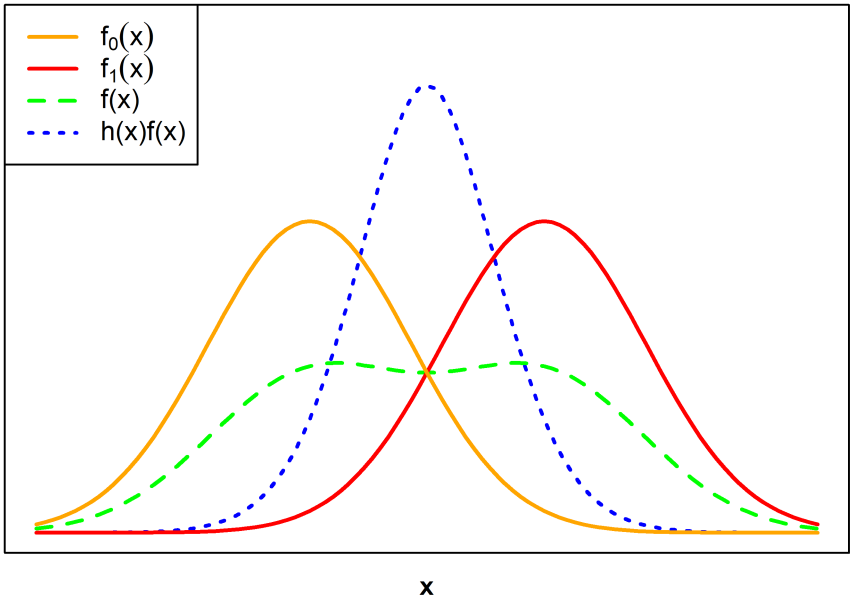
$$h(\boldsymbol{x}) = \frac{1}{\frac{1}{e(\boldsymbol{x})} + \frac{1}{1-e(\boldsymbol{x})}} = e(\boldsymbol{x})(1 - e(\boldsymbol{x})),$$

jolloin päällekkäisyyspainokertoimet ovat $w_1 \propto 1 - e(\boldsymbol{x}) = e(0, \boldsymbol{x})$ ja $w_0 \propto e(\boldsymbol{x}) = e(1, \boldsymbol{x})$ eli "väärän" ryhmän propensiteettipistemäärä.[7]

Päällekkäisyyspainokertoimet ovat aina välillä $(0, 1)$, joten yhden yksilön vaikutus painotetussa otoksessa ei kasva suhteettoman suureksi vaikka toteutuneen altistusryhmän propensiteettipistemäärä olisikin hyvin pieni. Lisäksi päällekkäisyyspainokertoimia käytettäessä yksilöitä ei poisteta otoksesta toisin kuin katkaisussa. Jakaumaa $f(\boldsymbol{x})$ kuitenkin painotetaan niin, että suurin merkitys on niillä yksilöillä,

joilla on kohtalainen todennäköisyys päätyä mihin tahansa altistusryhmään. Esimerkiksi lääketieteellisessä tutkimuksessa, jossa vertaillaan eri hoitomuotoja, huomio kiinnittyy erityisesti niihin potilaisiin, joille on vaikeinta valita sopivaa hoitoa. Tämä joukko saattaakin olla tutkimuskysymyksen kannalta kaikkein kiinnostavin.[7, 15]

Päällekkäisyyspainokertoimien vaikutusta taustamuuttujan jakaumaan kohdejoukossa, kun taustamuuttujia on yksi ja altistusryhmiä kaksi, havainnollistetaan kuvassa 1. Kuvasta nähdään, että päällekkäisyyspainokertoimilla painotetussa joukossa korostuvat erityisesti ne muuttujan X arvot, joiden kohdalla ryhmäkohtaiset jakaumat ovat eniten päällekkäin.



Kuva 1: Epävarmojen yksilöiden painottuminen kohdejoukossa, kun käytetään päällekkäisyyspainokertoimia. Kuvassa havainnollistetaan tilannetta, jossa on yksi taustamuuttuja X, joka noudattaa normaalijakaumaa kahdessa altistusryhmässä ($Z = 0$ ja $Z = 1$). Muuttujan X jakaumat altistusryhmissä ($f_0(x)$ ja $f_1(x)$) on merkitty oranssilla ja punaisella yhtenäisellä viivalla, ja niistä nähdään että eri altistusryhmissä painottuvat erilaiset muuttujan X arvot. Vihreä katkoviiva kuvaa muuttujan X jakaumaa $f(x)$ koko joukossa eli kohdepopulaatiossa kun $h(x) = 1$. Sininen pisteviiva puolestaan näyttää muuttujan jakauman $h(x)f(x)$ päällekkäisyyspainokertoimilla painotetussa joukossa ($h(x) = e(x)(1 - e(x))$). Tässä joukossa muuttujan X jakauma on tiheimmillään ryhmäkohtaisten jakaumien puolivälissä.

Funktion $h(\boldsymbol{x}) = \left(\sum_{z=0}^{T-1} 1/e(z, \boldsymbol{x})\right)^{-1}$ lisäksi on myös muita kallistusfunktioita, jotka saavat suurimman arvonsa, kun kaikkien altistusryhmien propensiteettipistemäärät ovat yhtä suuret. Jos valitaan $h(\boldsymbol{x}) = \min_z\{e(z, \boldsymbol{x})\}$, saadaan niin kutsutut *kaltaistuspainokertoimet* (*matching weights, MW*). Nimensä mukaisesti kaltaistuspainokertoimet muodostavat samankaltaisen asetelman kuin jos aineisto jaettaisiin kaltaistettuihin ryhmiin propensiteettipistemäärien perusteella niin, että ryhmässä olisi aina yksi yksilö jokaisesta altistusryhmästä. Tämä on helpointa havaita tilan-

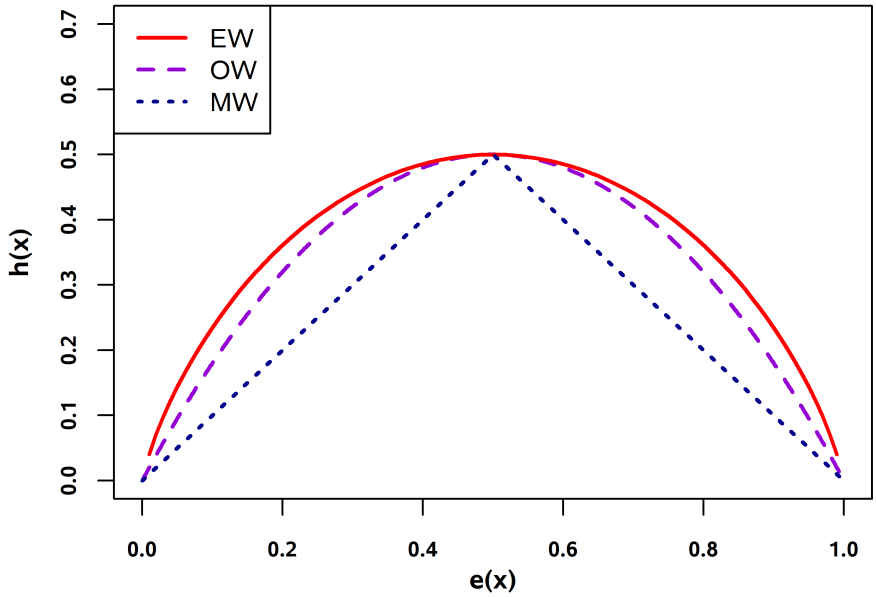
teessa, jossa on yksi altistuneiden tai käsiteltyjen ryhmä ja yksi kontrolliryhmä. Tällöin $h(\boldsymbol{x}) = \min\{e(\boldsymbol{x}), 1-e(\boldsymbol{x})\}$. Oletetaan, että on m yksilöä, joiden propensiteettipistemäärät ovat lähellä lukua $\tilde{e}$. Jos propensiteettipistemäärät estimoivat hyvin todellista todennäköisyyttä kuulua altistuneiden ryhmään, voidaan olettaa tässä joukossa olevan $m\tilde{e}$ altistunutta. Vastaavasti kontrolliryhmän yksilöitä on $m(1-\tilde{e})$. Jos $\tilde{e} \leq 0,5$, altistuneiden ryhmän yksilöitä on tässä m yksilön joukossa vähemmän kuin kontrolliryhmään kuuluvia. Parittaisessa kaltaistuksessa kaikille $m\tilde{e}$ altistuneelle siis saataisiin kontrolliryhmästä pari, jolla on lähes samansuuruinen propensiteettipistemäärä. Näiden altistuneiden todennäköisyys olla mukana lopullisessa kaltaistetussa joukossa on siis 1. Kontrolliryhmän yksilöistä taas osa jäisi ilman altistunutta paria. Kun $m(1-\tilde{e})$ yksilöstä valitaan $m\tilde{e}$ paria altistuneille, niin todennäköisyys olla mukana lopullisessa aineistossa on $\tilde{e}/(1-\tilde{e})$. Jos taas $\tilde{e} > 0,5$, kontrolliryhmän yksilöitä voidaan olettaa olevan vähemmän kuin altistuneita. Tällöin todennäköisyys olla mukana lopullisessa aineistossa on kontrolleille 1 ja altistuneille $(1-\tilde{e})/\tilde{e}$. Kaltaistuspainokerrointa $\min\{e(\boldsymbol{x}), 1-e(\boldsymbol{x})\}/e(\boldsymbol{x})$ voidaan siis pitää todennäköisyytenä tulla valituksi kaltaistettuun aineistoon. [24]

Kaltaistuspainokertoimet toimivat samaan tapaan myös, kun $K > 2$. Esimerkiksi kolmen altistusryhmän aineistossa kaltaistuspainot vastaavat kaltaistamista 1:1:1. Jos yksilön toteutuneen ryhmän propensiteettipistemäärä on pienin, saman ryhmän yksilöitä oletettavasti on vähiten kyseisen propensiteettipistemäärien yhteisjakauman pisteen ympäristössä. Tällöin yksilölle saataisiin haluttaessa poimittua kaltaistetut naapurit muista ryhmistä, joten todennäköisyys olla mukana kaltaistetussa aineistossa on 1. Muuten painokerroin pienenee, kun yksilön oman ryhmän propensiteettipistemäärä eli painokertoimen nimittäjä suurenee. Voidaan osoittaa (ks. [13, 24]), että kaltaistuspainokertoimilla saatu kausaalivaikutuksen estimaatti on asymptoottisesti sama kuin propensiteettipistemäärien perusteella kaltaistetusta aineistosta saatu estimaatti.

Kolmas tapa painottaa yksilöitä, joilla on kohtalainen todennäköisyys kuulua mihin ryhmään tahansa, on käyttää *entropiapainokertoimia* (*entropy weights, EW*). Tällöin valitaan kallistusfunktioksi $h(\boldsymbol{x}) = -\sum_{z=0}^{K-1} e(z, \boldsymbol{x})\ln(e(z, \boldsymbol{x}))$. Entropiafunktiolla arvioidaan tapahtuman epävarmuutta, ja sitä käytetään yleisesti esimerkiksi koneoppimisen menetelmissä ja informaatioteoriassa. Tässä yhteydessä tapahtuma on yksilön sijoittuminen altistusryhmään. Entropiafunktio saa suurimman arvonsa, kun kaikki vaihtoehdot ovat yhtä todennäköisiä. Kun vaihtoehtoina on K eri altistusryhmää, entropia eli epävarmuus on suurinta jokaisen ryhmän todennäköisyyden ollessa $1/K$. Entropiapainokertoimet, kuten päällekkäisyys- ja kaltaistuspainotkin, ovat aina nollan ja yhden välillä, joten erittäin suuret painokertoimet eivät muodosta ongelmaa. [23]

Päällekkäisyys-, kaltaistus- ja entropiapainokertoimet toimivat siis samaan tapaan. Kaikki antavat suurimman painon niille yksilöille, joilla kaikkien altistusryhmien propensiteettipistemäärät ovat lähellä toisiaan. Ne yksilöt, joiden estimoitu todennäköisyys kuulua johonkin tiettyyn ryhmään on hyvin suuri, saavat pienen painon. Kuvassa 2 on esitetty näihin kolmeen painokertoimeen liittyvät kallistusfunktiot, kun altistusryhmiä on kaksi. Kaikki saavat suurimman arvonsa, kun $e(\boldsymbol{x}) = 0,5$ ja lähestyvät nollaa, kun $e(\boldsymbol{x})$ lähestyy nollaa tai yhtä. Päällekkäisyys- ja entropiapainokertoimien kallistusfunktiot ovat melko samanmuotoiset, kun taas kaltaistus-

painokertoimien kallistusfunktio on terävämpi. Eri kallistusfunktiot johtavat hiukan erilaisiin painokertoimiin, mutta kaikki kolme ovat aina välillä [0,1].



Kuva 2: Kallistusfunktiot kaltaistus- (MW), päällekkäisyys- (OW) ja entropiapainokertoimille (EW) kahden altistusryhmän tapauksessa. Kaikki kolme saavat suurimman arvonsa, kun kummankin altistusryhmän estimoidut todennäköisyydet ovat yhtä suuret. Vertailun helpottamiseksi funktiot on kerrottu sopivalla vakiolla niin, että kaikki saavat pisteessä $e(\boldsymbol{x}) = 0,5$ arvon 0,5.

Taulukkoon 1 on koottu kaikki esitellyt painokertoimet ja niiden kallistusfunktiot. Voidaan sanoa, että eri painokertoimet toimivat hiukan erilaisella logiikalla. Käänteistodennäköisyyspainokertoimet ovat aina suurempia kuin 1. Niiden avulla muodostettuihin pseudojoukkoihin siis tavallaan lisätään kopioita alkuperäisen aineiston yksilöistä. Esimerkiksi jos altistusryhmään z' sijoittuneella yksilöllä propensiteettipistemäärä $e(z', \boldsymbol{x}_i)$ on hyvin pieni ja toisen ryhmän pistemäärä $e(z'', \boldsymbol{x}_i)$ suuri, sen voidaan ajatella antavan tietoa tilanteesta, jossa ryhmän z'' yksilö olisikin ryhmässä z'. Tämän yksilön vastemuuttujan arvo siis edustaa potentiaalista vastetta, jota ryhmään z'' kuuluvilta yksilöiltä ei havaita. Siksi siitä lisätään pseudojoukkoon monta kopiota. Sitä vastoin päällekkäisyys-, kaltaistus- ja entropiapainokertoimet ovat aina pienempiä tai yhtä suuria kuin 1. Sen sijaan että aliedustetuista yksilöistä lisättäisiin kopioita, yksilön painoa pienennetään, jos sen kaltaiset ovat joukossa yliedustettuina verrattuna harvinaisempiin tapauksiin. Jos painokerroin on esimerkiksi 0,6, yksilöstä otetaan pseudojoukkoon vain 60%. ATT-painoissa yhdistyvät nämä lähestymistavat: Jos toisen ryhmän yksilö muistuttaa taustamuuttujien perusteella paljon kohdejoukkona olevan ryhmän yksilöitä, sen painoarvoa kasvatetaan. Sen sijaan jos yksilön todennäköisyys kuulua kohdejoukkona olevaan ryhmään on pieni, sitä painotetaan alaspäin.[24]

Taulukko 1: Esimerkkejä mahdollisista kallistusfunktioista $h(\boldsymbol{x})$. Painokertoimet ovat aina $w_i = h(\boldsymbol{x}_i)/e(z_i, \boldsymbol{x}_i)$, $z_i = 0, \ldots, K-1$.

Nimi	**Lyhenne**	**Kohdejoukko**	**Kallistusfunktio**
Käänteistn	IPW	Koko joukko	1
Katkaistu käänteistn	TIPW	Koko joukko ilman suurimpia ja pienimpiä painoja	$\mathrm{I}(\boldsymbol{x} \in \mathcal{C})$
Ryhmän sisäinen	ATT	Altistusryhmä z'	$e(z', \boldsymbol{x}_i)$
Päällekkäisyyspaino	OW	Painotetaan	$\left(\sum_{z=0}^{K-1} 1/e(z, \boldsymbol{x})\right)^{-1}$
Kaltaistuspaino	MW	epävarmimpia	$\min_z\{e(z, \boldsymbol{x})\}$
Entropiapaino	EW	yksilöitä	$-\sum_{z=0}^{K-1} e(z, \boldsymbol{x})\ln(e(z, \boldsymbol{x}))$

4.5 Taustamuuttujien tasapainoisuuden arviointi painotetussa aineistossa

Propensiteettipistemäärän avulla on tarkoitus tasapainottaa eroavaisuuksia havaittujen kovariaattien jakaumassa. Jos estimoitu propensiteettipistemäärä $e(z, \boldsymbol{x}_i)$ vastaa todellista todennäköisyyttä $P(Z_i = z|\boldsymbol{X} = \boldsymbol{x}_i)$ kaikilla $i = 1, \ldots, n$, saman pistemäärän saaneiden joukossa havaittujen taustamuuttujien jakauma ei riipu altistusryhmästä. Painokertoimia käytettäessä tästä seuraa, että taustamuuttujien painotettu jakauma jokaisessa K altistusryhmässä on sama kuin niiden jakauma koko kallistusfunktion h määräämässä kohdejoukossa. Toisin sanoen $f_z(\boldsymbol{X})w_z(\boldsymbol{X}) = f(\boldsymbol{X})h(\boldsymbol{X})$ kaikilla $z = 0, \ldots, K-1$. Jotta menetelmät toimisivat oikein, on tärkeää varmistaa, että tämä oletus toteutuu.

Painotetun jakauman tasapainoisuuden tutkiminen, kun altistusmuuttuja on kaksiluokkainen, on esitelty artikkelissa [12]. Seuraavaksi esiteltävä yleistys tilanteeseen, jossa $K > 2$, perustuu artikkeleihin [5] ja [7]. Tasapainoisuustarkasteluja varten jokaiselle p taustamuuttujalle määritellään ryhmässä z painotettu keskiarvo

$$\bar{x}_{jz} = \frac{\sum_{i=1}^{n} D_i(z) x_{ij} w_i}{\sum_{i=1}^{n} D_i(z) w_i}, \quad j = 1, \ldots, p,\ z = 0, \ldots, K-1,$$

ja koko kohdejoukossa keskiarvo

$$\bar{x}_{jh} = \frac{\sum_{i=1}^{n} x_{ij} h(\boldsymbol{x}_i)}{\sum_{i=1}^{n} h(\boldsymbol{x}_i)}.$$

Luokallinen taustamuuttuja voidaan muuttaa yhdeksi tai useammaksi indikaattorimuuttujaksi, jotkat saavat arvon 1 tai 0. Tällöin keskiarvot vastaavat eri luokkien osuuksia. Jos esimerkiksi taustamuuttuja X_j on kolmiluokkainen ($X_j \in \{1, 2, 3\}$), voidaan tasapainoisuustarkasteluja varten muodostaa muuttujat $X_{j_2} = \mathrm{I}(X_j = 2)$ ja $X_{j_3} = \mathrm{I}(X_j = 3)$. Keskiarvot $\bar{x}_{j_2 z}$ ja $\bar{x}_{j_3 z}$ ovat tässä tapauksessa luokkien 2 ja 3 painotetut osuudet altistusryhmässä z ja $\bar{x}_{j_2 h}$ ja $\bar{x}_{j_3 h}$ vastaavat osuudet koko kohdejoukosta.

Tasapainoisuutta voidaan tarkastella kahdella tavalla. Ensimmäinen tapa on vertailla taustamuuttujien jakaumia eri ryhmissä koko kohdejoukon jakaumaan. Jos

painotetut jakaumat todella vastaavat kohdejoukon jakaumaa, ryhmäkohtaisen keskiarvon $\bar{x}_{jz}$ tulisi olla lähellä kohdejoukon keskiarvoa $\bar{x}_{jh}$ kaikilla $z = 0, \ldots, K-1$. Muuttujan X_j painottamatonta otosvarianssia ryhmässä z merkitään $S^2_{x_j,z}$, jolloin varianssien painottamaton keskiarvo on $S^2_{x_j} = \frac{1}{K}\sum_{z=0}^{K-1} S^2_{x_j,z}$. Nyt voidaan määritellä standardoitu erotus keskiarvosta (SEK) jokaiselle taustamuuttujalle ja jokaiselle altistusryhmälle z niin, että

$$SEK_{j,z} = \frac{|\bar{x}_{jz} - \bar{x}_{hz}|}{S_{x_j}}.$$

Jos suurin erotus $\max_z SEK_{j,z}$ on liian suuri, voidaan päätellä taustamuuttujan X_j jakaumassa olevan epätasapainoa ryhmien välillä.

Toisaalta taustamuuttujien jakaumien tasapainoisuus eri ryhmien välillä voidaan ilmaista myös muodossa $f_{z'}(\boldsymbol{X})w_{z'}(\boldsymbol{X}) = f_{z''}(\boldsymbol{X})w_{z''}(\boldsymbol{X})$. Niinpä voidaan myös tarkastella taustamuuttujien jakaumien samankaltaisuutta parittain eri ryhmien välillä. Tällöin tutkitaan parittaisia standardoituja erotuksia (PSE)

$$PSE_{j,z'z''} = \frac{|\bar{x}_{jz'} - \bar{x}_{jz''}|}{S_{x_j}}.$$

Myös parittaisten erotusten kohdalla tarkastellaan suurinta arvoa $\max_{z'<z''} PSE_{j,z'z''}$. Nytkin suuri suurin arvo viittaa muuttujan X_j epätasapainoisuuteen.

Sille, kuinka suuri SEK tai PSE merkitsee epätasapainoisuutta taustamuuttujan arvoissa ryhmien välillä, ei ole määritelty ehdottomia rajoja. Eri tutkijat käyttävät hiukan eri arvoja. Joidenkin mielestä muuttuja on epätasapainossa jos SEK tai PSE on suurempi kuin 0,2. Toiset taas pitävät ongelmallisena, jos jompi kumpi luku on suurempi kuin 0,25.[5]

Erotukset keskiarvosta ja parittaiset erotukset eivät riipu itse taustamuuttujien vaihteluvälistä. Niinpä niiden perusteella voidaan hyvin asettaa taustamuuttujat järjestykseen sen mukaan, kuinka hyvin ne ovat tasapainossa eri ryhmissä. Jos jokin taustamuuttuja vaikuttaa olevan epätasapainoinen, on mahdollista yrittää parantaa tilannetta lisäämällä propensiteettipistemäärät estimoivaan malliin kyseisen muuttujan muunnoksia. Voidaan kokeilla esimerkiksi muuttujan korottamista toiseen potenssiin tai kertomista jonkin toisen muuttujan arvolla eli yhteisvaikutusta. Tämän jälkeen propensiteettipistemäärät estimoidaan, painokertoimet lasketaan ja tasapainoisuus arvioidaan uudestaan. Tarvittaessa samat vaiheet voidaan toistaa useaan kertaan, kunnes taustamuuttujat ovat tasapainossa. Lisättävien interaktioiden ja muiden muunnosten tulisi kuitenkin olla myös tulkinnallisesti järkeviä. Kun propensiteettipistemäärien estimointiin käytetään koneoppimisen menetelmiä, erotuksia keskiarvosta tai parittaisia erotuksia voidaan käyttää lopetuskriteereinä. Tällöin propensiteettipistemäärät estimoivaan malliin lisätään taustamuuttujien muunnoksia sen perusteella, mitkä saavat tasapainoisuuden paranemaan eniten. Iterointi päättyy, kun valittu tasapainoisuutta kuvaava luku on tarpeeksi pieni. Näissä menetelmissä lisättävien termien tulkinnallista mielekkyyttä on tosin vaikea varmistaa.

Artikkelissa [15] esitellään mielenkiintoinen tulos tilanteessa, jossa altistusryhmiä on kaksi ($K = 2$), propensiteettipistemäärät estimoidaan logistisen regression avulla ja käytetään päällekkäisyyspainoja. Tässä tapauksessa $Z_i = 1$, kun yksilö i kuuluu

käsittelyryhmään ja $Z_i = 0$ muuten. Lisäksi merkitään $e(\boldsymbol{x}_i) = e(1, \boldsymbol{x}_i)$, jolloin $e(0, \boldsymbol{x}_i) = 1 - e(\boldsymbol{x}_i)$. Logistisessa regressiossa logaritminen uskottavuusfunktio on

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} z_i \log p(\boldsymbol{x}_i) + (1 - z_i)\log(1 - p(\boldsymbol{x}_i)),$$

jossa $p(\boldsymbol{x}_i) = P(Z = 1|X = \boldsymbol{x}_i)$. Merkitään $\boldsymbol{x}_i^* = (1, \boldsymbol{x}_i')'$ ja $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$. Kun logaritmiseen uskottavuusfunktioon sijoitetaan propensiteettipistemäärä $e(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i^*)/\big(1 + \exp(\boldsymbol{\beta}'\boldsymbol{x}_i^*)\big)$, saadaan

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log(1 - e(\boldsymbol{x}_i)) + z_i \log\left(\frac{e(\boldsymbol{x}_i)}{1 - e(\boldsymbol{x}_i)}\right) \\ &= \sum_{i=1}^{n} \log\left(1 - \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_i^*}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}_i^*}}\right) + z_i \boldsymbol{\beta}'\boldsymbol{x}_i^* \\ &= \sum_{i=1}^{n} -\log(1 + e^{\boldsymbol{\beta}'\boldsymbol{x}_i^*}) + z_i \boldsymbol{\beta}'\boldsymbol{x}_i^*. \end{aligned}$$

Derivoituna parametrin β_j, $j = 0, \ldots, p$ suhteen tästä tulee

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} -\frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_i^*}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}_i^*}} x_{ij} + z_i x_{ij} = \sum_{i=1}^{n} (z_i - e(\boldsymbol{x}_i)) x_{ij},$$

jossa $x_{i0} = 1$ kaikilla $i = 1, \ldots, n$. Kun derivaatta asetetaan nollaksi, saadaan

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} e(\boldsymbol{x}_i) \quad \text{ja} \quad \sum_{i=1}^{n} x_{ij} z_i = \sum_{i=1}^{n} x_{ij} e(\boldsymbol{x}_i),$$

mistä seuraa, että

$$\sum_{i=1}^{n} z_i (1 - e(\boldsymbol{x}_i)) = \sum_{i=1}^{n} e(\boldsymbol{x}_i) - \sum_{i=1}^{n} z_i e(\boldsymbol{x}_i) = \sum_{i=1}^{n} e(\boldsymbol{x}_i)(1 - z_i)$$

ja

$$\sum_{i=1}^{n} x_{ij} z_i (1 - e(\boldsymbol{x}_i)) = \sum_{i=1}^{n} x_{ij} e(\boldsymbol{x}_i) - \sum_{i=1}^{n} x_{ij} z_i e(\boldsymbol{x}_i) = \sum_{i=1}^{n} x_{ij} e(\boldsymbol{x}_i)(1 - z_i),$$

kaikilla $j = 1, \ldots, p$.

Kun $K = 2$, $D_{i1} = Z_i$ ja $D_{i0} = 1 - Z_i$. Päällekkäisyyspainokertoimia käytettäessä $h(\boldsymbol{x}) = e(\boldsymbol{x}_i)(1 - e(\boldsymbol{x}_i))$. Näin ollen

$$\begin{aligned} \bar{x}_{j1} &= \sum_{i=1}^{n} z_i x_{ij} (1 - e(\boldsymbol{x}_i)) / \sum_{i=1}^{n} z_i (1 - e(\boldsymbol{x}_i)) \\ &= \sum_{i=1}^{n} (1 - z_i) x_{ij} e(\boldsymbol{x}_i) / \sum_{i=1}^{n} (1 - z_i) e(\boldsymbol{x}_i) \\ &= \bar{x}_{j0}. \end{aligned}$$

Muuttujan X_j painotettu keskiarvo on siis sama kummassakin altistusryhmässä, joten $PSE_{j,01} = 0$ kaikilla $j = 1, \ldots, p$. Tämä automaattinen tasapainoisuus ei kuitenkaan päde, kun $K > 2$ tai käytetään jotain muuta kallistusfunktiota.

4.6 Täydennetyt estimaattorit

Useimmiten propensiteettipistemääräpainokertoimia hyödynnettäessä taustamuuttujia $\boldsymbol{X}$ käytetään vain propensiteettipistemäärien estimoimiseen. Altistuksen yhteyttä vasteeseen tutkitaan painotetun aineiston avulla ilman että taustamuuttujat ovat mukana muuten kuin painokertoimien kautta. Taustamuuttujat kuitenkin vaikuttavat altistuksen lisäksi myös vasteeseen. Luvussa 4.2 esiteltyä parametritonta estimaattoria voidaankin täydentää yhdistämällä siihen perinteinen regressiomalli, jossa taustamuuttujat ja altistus ovat selittäviä muuttujia. Robinsin, Rotnitzkyn ja Zhaon alkuperäisessä artikkelissa vuodelta 1994 täydennetyt estimaattorit liittyvät tilanteeseen, jossa joidenkin selittävien muuttujien arvoa ei havaita osalla yksilöistä [25]. Artikkelin kuvaamissa tilanteissa painokertoimessa esiintyvä estimoitu todennäköisyys koskee sitä, havaitaanko yksilöltä kaikkien selittävien muuttujien arvot vai ei. Täydennettyjä estimaattoreita käytetään kuitenkin yleisesti myös propensiteettipistemääräpainokertoimien kanssa. Tässä tutkielmassa ei kuitenkaan ole mukana käytännön esimerkkiä niistä.

Täydennetty estimaattori vasteen odotusarvosta ryhmässä z on

$$\hat{m}_z^{h,aug} = \hat{m}_z^h - \frac{\sum_{i=1}^n \big(D_i(z) - e(z, \boldsymbol{x}_i)\big) w_i \hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i)}{\sum_{i=1}^n h(\boldsymbol{x}_i)},$$

jossa $\hat{m}_z^h$ on luvun 4.2 parametriton estimaattori ja $\hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i)$ on regressiomallin tulos, jos yksilö i kuuluisi ryhmään z. Kahden ryhmän välistä eroa voidaan tarkastella erotuksen $\hat{\tau}^{h,aug}(z', z'') = \hat{m}_{z'}^h - \hat{m}_{z''}^h$ avulla. Ei ole välttämätöntä, että taustamuuttujien joukko $\boldsymbol{X}$ on sama sekä propensiteettipistemäärien estimoinnissa että vastemuuttujan regressiomallissa, vaan muuttujat voidaan valita kumpaankin malliin erikseen.[7, 26]

Täydennetty estimaattori myös tasapainottaa hyvin suurien painojen vaikutusta. Jos yksilö i kuuluu ryhmään z, niin $D_i(z) = 1$ ja täydennetyn estimaattorin jälkimmäisessä termissä osoittajan summaan tulee $\big(1 - e(z, \boldsymbol{x}_i)\big) w_i \hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i)$. Painokerroin w_i on suuri, kun $e(z, \boldsymbol{x}_i)$ on lähellä nollaa. Tällöin edellinen termi on lähellä lukua $w_i \hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i)$. Muistetaan, että

$$\hat{m}_z^h = \frac{\sum_{i=1}^n D_i(z) Y_i w_i}{\sum_{i=1}^n D_i(z) w_i}.$$

Jos siis Y_i on mukana parametrittomassa estimaattorissa suurella painokertoimella, niin täydennetyssä estimaattorissa $\hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i)$ on mukana miinusmerkkisenä lähes yhtä suurella painokertoimella, mikä tasapainottaa estimaattoria. Jos taas $e(z, \boldsymbol{x}_i)$ on lähellä yhtä ja yksilö i kuuluu ryhmään z, niin osoittajan summassa $\big(1 - e(z, \boldsymbol{x}_i)\big) w_i \hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i)$ on lähellä nollaa ja $\hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i)$ vaikuttaa estimaattoriin vain vähän.[26]

Jos yksilö i ei kuulu ryhmään z, niin $D_i(z) = 0$ ja osoittajan summaan jää

$$\begin{aligned} -e(z, \boldsymbol{x}_i) w_z(\boldsymbol{x}_i) \hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i) &= -e(z, \boldsymbol{x}_i) \frac{h(\boldsymbol{x})}{e(z, \boldsymbol{x}_i)} \hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i) \\ &= -h(\boldsymbol{x}) \hat{E}(Y_i | Z_i = z, \boldsymbol{x}_i). \end{aligned}$$

Näin ollen täydennettyyn estimaattoriin $\hat{\tau}^{h,aug}(z', z'')$ vaikuttavat myös ne yksilöt, jotka eivät kuulu ryhmään z' tai z'', toisin kuin pelkässä parametrittomassa estimaattorissa.

Kun käytetään käänteistodennäköisyyspainokertoimia eli $h(\boldsymbol{x}) = 1$, niin $\hat{m}_z^{h,aug}$ on odotusarvon $E[Y(z)]$ tarkentuva estimaattori, jos joko propensiteettipistemäärät estimoiva malli tai vasteen regressiomalli on määritelty oikein. Tällöin saatetaankin puhua kaksinkertaisesti vakaasta estimaattorista. Kuitenkin silloin kun on valittu jokin toinen kallistusfunktio, tarkentuvuus edellyttää että propensiteettipisteiden malli on oikea eli estimoidut propensiteettipistemäärät vastaavat todellisia todennäköisyyksiä.[7]

5 Äidin raskaudenaikaisen masennuslääkkeiden käytön vaikutus lapsen masennukseen ja ahdistuneisuuteen

Viime vuosikymmenten aikana sekä masennusdiagnoosit että masennuslääkkeiden käyttö ovat lisääntyneet merkittävästi. Naisilla masennus on miehiä yleisempää, joten ei ole yllättävää, että masennuslääkkeiden käyttö myös raskauden aikana on yleistynyt. Tämä herättää kysymyksiä lääkkeiden vaikutuksesta lapsen kehitykseen ja myöhemmin ilmeneviin oireisiin. Masennuslääkkeet perustuvat aivojen välittäjäaineiden, esimerkiksi serotoniinin ja dopamiinin, säätelyyn. Nämä välittäjäaineet vaikuttavat mielialaan ja tunteisiin, ja niiden tasapainoa säätelemällä voidaan helpottaa masennuksen oireita. Välittäjäaineilla on kuitenkin tehtävänsä myös aivojen kehityksessä sikiövaiheessa. Äidin käyttämän masennuslääkkeen vaikuttavat aineet siirtyvät istukan kautta myös sikiön verenkiertoon, mikä voi vaikuttaa keskushermoston kehitykseen ja mahdollisesti aiheuttaa myöhemmin terveysongelmia.[27, 28]

Raskauden aikana käytetyistä masennuslääkkeistä on tutkittu erityisesti yleisintä ryhmää, selektiivisiä serotoniinin takaisinoton estäjiä (selective serotonin reuptake inhibitors, SSRI-lääkkeet). Jyrsijöillä tehdyissä kokeissa altistumisen SSRI-lääkkeille herkässä kehitysvaiheessa on havaittu vaikuttavan muun muassa käyttäytymiseen ja aivojen toimintaan [28]. Ihmisillä vastaavat kokeelliset tutkimukset lääkkeiden vaikutuksesta eivät ole eettisesti mahdollisia, joten aihetta tutkitaan havainnoivien tutkimusten avulla. Joissain tutkimuksissa SSRI-lääkkeille altistumisella on havaittu olevan yhteyttä lapsuudessa ja varhaisnuoruudessa todettuun masennukseen, mutta toisissa vastaavasta yhteydestä ei ole saatu todisteita. Myös esimerkiksi vaikutuksesta autismiin on ristiriitaisia havaintoja. SSRI-lääkkeille altistuneissa lapsissa on havaittu myös joitain fysiologisia eroja aivoissa verrattuna altistumattomiin. Lisäksi on havaittu yhteyksiä SSRI-lääkkeiden käytön ja lapsen kehityksen ongelmien välillä. Erityisesti altistuminen raskauden ensimmäisen kolmanneksen aikana saattaa lisätä epämuodostumien riskiä, mutta tähänkin päätelmään liittyy epävarmuutta.[29, 30, 31]

Kun tutkitaan raskaudenaikaisen masennuslääkkeiden käytön vaikutuksia, pitäisi pyrkiä erottamaan lääkkeistä johtuvat seuraukset äidin masennuksen vaikutuksista. Äidin masennus nimittäin on myös riskitekijä lapsen fyysisen ja psyykkisen terveyden kannalta. Jos masennuksen lääkehoito keskeytetään raskauden aikana, voi sekä äidille että lapselle aiheutua vakavia seurauksia. Siksi masennuslääkkeiden käytön lopettamista raskauden ajaksi ei suositella ilman neuvottelua lääkärin kanssa.[32]

5.1 Aineisto

Tässä tutkimuksessa pyrittiin selvittämään raskaudenaikaisen masennuslääkealtistuksen vaikutusta lapsuudessa ja nuoruudessa diagnosoituun masenukseen ja ahdistuneisuushäiriöön. Tutkimusaineisto perustui viranomaisten ylläpitämiin rekisteriaineistoihin. Kohderyhmänä olivat kaikki Suomessa vuosina 1999–2016 elävänä syntyneet lapset pois lukien monikkosynnytykset. Tietoja seurattiin rekistereiden perusteella vuoden 2018 loppuun asti eli tutkimushenkilöt olivat seurannan päät-

tyessä 2—19-vuotiaita.

Tietoja kerättiin seuraavista rekistereistä:

- Terveyden ja hyvinvoinnin laitoksen (THL) ylläpitämä syntyneiden lasten rekisteri: Tiedot muun muassa raskauden arvioidusta alkupäivästä, äidin terveydentilasta ja elintavoista raskauden aikana (esimerkiksi tupakointi), aiemmista raskauksista ja synnytyksistä, äidin raskauden tai synnytyksen aikana saamista diagnooseista sekä vanhempien sosioekonomisesta asemasta ja asuinympäristöstä (maaseutu tai kaupunki).

- THL:n hoito- ja poistoilmoitusrekisteri: Rekisteri sisältää tiedot kaikista hoitojaksoista vuodeosastolla erikoissairaanhoidossa ja perusterveydenhuollossa sekä avohoidosta julkisessa erikoissairaanhoidossa. Sekä kohdejoukon lapsille että heidän vanhemmilleen poimittiin rekisteristä tiedot hoitojakson tai käynnin ajankohdasta sekä annetuista pää- ja sivudiagnooseista. Rekisterissä on tietoja vuodesta 1969 alkaen, joten vanhempien tietoja saatiin myös ajalta ennen raskauden alkua. Vuodesta 2011 alkaen saatavilla on myös perusterveydenhuollon avohoidossa annetut diagnoosit. Diagnoosit on koodattu kansainvälisen tautiluokituksen (International Classification of Diseases, ICD) mukaan. Eri versioista käytössä ovat olleet ICD-8 (vuosina 1969-1986), ICD-9 (1987-1995) ja ICD-10 (1996 alkaen).

- Digi- ja väestöviraston (DVV) väestörekisteri: Tiedot seurannan aikana kuolleista tai ulkomaille muuttaneista tutkimushenkilöistä.

- Kansaneläkelaitoksen (KELA) lääketoimitusrekisteri: Tiedot Kela-korvattavien lääkkeiden ostoista apteekeissa vuodesta 1993 alkaen. Lääkkeet tunnistetaan rekisteristä kansainvälisen anatomis-terapeuttis-kemiallisen (Anatomical Therapeutic Chemical, ATC) luokituksen perusteella.

Kaiken kaikkiaan rekistereistä kerättiin 1 002 379 lapsen ja heidän vanhempiensa tietoja. Näiden tietojen perusteella lasten joukosta tunnistettiin kolme altistusryhmää:

- Masennuslääkkeille altistuneiden ryhmään kuuluivat ne, joiden äiti oli käyttänyt jotakin masennuslääkettä raskauden aikana. Äidin tulkittiin käyttäneen lääkkeitä, jos hän oli ostanut niitä vähintään kerran raskauden aikana tai enintään kolme kuukautta ennen raskauden arvioitua alkamispäivää. ATC-luokituksessa masennuslääkkeisiin kuuluvat ne, joiden koodin alku on N06A tai N06CA.

- Verrokkiryhmään 1 kuuluvien äidit eivät olleet ostaneet mitään masennus- tai psykoosilääkkeitä (ATC-koodit N06A, N06CA ja N05A) vuotta ennen raskauden alkua tai sen aikana. He olivat kuitenkin saaneet diagnoosin masennuksesta tai muusta masennukseen läheisesti liittyvästä häiriöstä raskauden aikana tai enintään vuosi ennen sitä. Masennukseen liittyviksi häiriöiksi laskettiin muun muassa ahdistuneisuushäiriöt, kaksisuuntainen mielialahäiriö ja skitsofrenia. ICD-10 luokituksessa masennukseen liittyviä häiriöitä kuvaavat koodit F20–F48.

- Verrokkiryhmän 2 henkilöiden äideillä ei ollut rekisterien kattamalla ajanjaksolla masennus- tai psykoosilääkeostoja eikä diagnoosia masennukseen liittyvistä häiriöistä missään vaiheessa ennen raskautta tai sen aikana. Tämä ryhmä siis edustaa terveiden äitien lapsia.

Altistuneiden ryhmään kuului 32 562 henkilöä ja verrokkiryhmään 1 puolestaan 16 499 henkilöä. Verrokkiryhmän 2 koko oli 824 256 henkilöä. Yhteensä lopulliseen tutkimusaineistoon kuului siis 873 317 vuosina 1999–2016 syntynyttä henkilöä. Lopullisen aineiston ulkopuolelle jäivät muun muassa ne lapset, joiden äiti oli saanut masennukseen liittyvän diagnoosin tai ostanut masennuslääkkeitä ennen raskautta, mutta ei edellä määriteltyjen aikavälien sisällä. Masennuslääkkeiden käytön yleistyminen näkyi tutkimusaineistossa: Koko aineistosta vuonna 1999 syntyneistä lapsista 1,08 % oli altistunut raskauden aikana masennuslääkkeille, mutta vuonna 2016 vastaava luku oli 4,23 %. Suurimmillaan altistuneiden osuus oli vuonna 2011 syntyneiden joukossa: 4,79 %. Masennukseen liittyvälle sairaudelle mutta ei lääkkeille altistuneiden osuus nousi myös, vaikkakaan ei yhtä paljon: vuonna 1999 syntyneistä tähän ryhmään kuului 0,96 % ja vuonna 2016 syntyneistä 2,17 %.

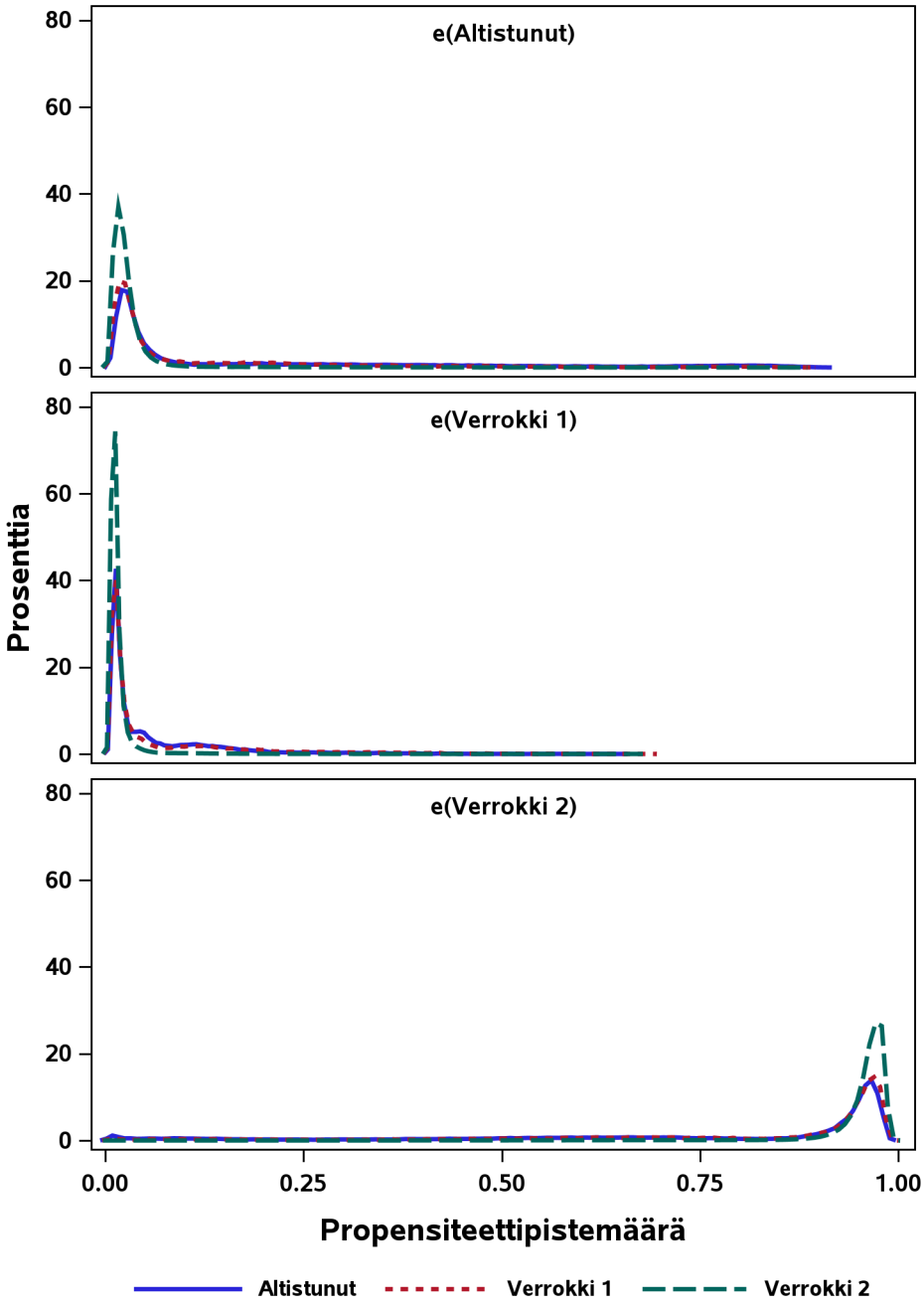
Altistusryhmiin sijoittuminen määriteltiin ainoastaan rekisteritietojen avulla eikä tutkimushenkilöihin tai heidän äiteihinsä otettu henkilökohtaisesti yhteyttä. Niinpä ei voida varmuudella sanoa, käyttivätkö masennuslääkkeitä raskauden aikana ostaneet äidit niitä todella vai eivät. Toisaalta kaikki masennuksesta kärsivät eivät saa diagnoosia, joten osa terveiksi määritellyistä äideistä saattoi todellisuudessa potea masennusta. Rekisteriaineistojen käyttö kuitenkin mahdollistaa hyvin suuren ihmisjoukon tutkimisen, mikä ei onnistuisi esimerkiksi kyselytutkimuksen avulla. Harhaanjohtavien rekisteritietojen ei arvioitu vaikuttavan suuresti tuloksiin.

5.2 Propensiteettipistemäärien estimointi ja painokertoimet

Propensiteettipistemäärien estimointia varten valittiin 22 taustamuuttujaa, joista yksi (lapsen syntymävuosi) oli jatkuva ja loput luokallisia. Ne liittyivät erityisesti äidin terveydentilaan raskauden aikana ja ennen sitä sekä vanhempien sosioekonomiseen asemaan ja asuinympäristöön. Myös vanhempien ja sisarusten ennen lapsen syntymää saamat psykiatriset diagnoosit otettiin huomioon. Muuttujajoukon valitsivat lastenpsykiatrian asiantuntijat. Luettelo taustamuuttujista on liitteessä A.

Kaikki analyysit tehtiin SAS-ohjelmistolla (versio 9.4). Propensiteettipistemäärien laskemiseen käytettiin multinomista logistista regressiota. Näin saatiin jokaiselle tutkimusjoukon henkilölle kolme propensiteettipistemäärää: estimoidut todennäköisyydet kuulua altistuneiden ryhmään, verrokkiryhmään 1 tai verrokkiryhmään 2. Merkitään propensiteettipistemääriä $e(\text{Altistunut})$, $e(\text{Verrokki 1})$ ja $e(\text{Verrokki 2})$.

Taulukossa 2 ja kuvassa 3 ovat kaikkien propensiteettipistemäärien jakaumat eri altistusryhmissä. Kaikissa ryhmissä verrokkiryhmän 2 propensiteettipistemäärä painottui melko lähelle yhtä ja kahden muun ryhmän lähelle nollaa. Altistuneiden ryhmän ja verrokkiryhmän 1 välillä ei ollut suuria eroja pistemäärien jakaumassa. Verrokkiryhmässä 2 jakaumat painottuivat vielä muita ryhmiä voimakkaammin lähelle ääripäitä. Myös verrokkiryhmässä 2 oli joitakin yksilöitä, joilla estimoitu todennäköisyys kuulua jompaan kumpaan muuhun ryhmään oli kohtalaisen suuri. Kuitenkin tästä ryhmästä yli 90 %:lla $e(\text{Verrokki 2}, \boldsymbol{x})$ oli yli 0,9.



Kuva 3: Propensiteettipistemäärien jakaumat altistusryhmittäin. Tutkittavat henkilöt jaettiin rekisteritietojen perusteella altistuneisiin ja kahteen verrokkiryhmään. Estimoidut todennäköisyydet kuulua eri altistusrymiin (propensiteettipistemäärät) saatiin multinomisen logistisen regression avulla, jossa valitut taustamuuttujat olivat selittävinä tekijöinä. Ylimmässä kuvaajassa ovat estimoidut todennäköisyydet kuulua altistuneiden ryhmään, keskimmäisessä verrokkiryhmään 1 ja alimmassa verrokkiryhmään 2. Eri viivat kuvaavat propensiteettipistemäärien jakaumia toteutuneiden altistusryhmien sisällä: yhtenäinen sininen viiva altistuneiden ryhmässä, punainen pisteviiva verrokkiryhmässä 1 ja vihreä katkoviiva verrokkiryhmässä 2.

Taulukko 2: Propensiteettipistemäärien jakaumat eri altistusryhmissä.

	Mediaani	**Minimi**	**1. desiili**	**9. desiili**	**Maksimi**
Altistunut					
e(Altistunut)	0,046	0,0063	0,017	0,54	0,91
e(Verrokki 1)	0,022	0,0030	0,0098	0,16	0,65
e(Verrokki 2)	0,93	0,0011	0,26	0,97	0,99
Verrokki 1					
e(Altistunut)	0,037	0,0040	0,014	0,35	0,89
e(Verrokki 1)	0,021	0,0027	0,0092	0,20	0,69
e(Verrokki 2)	0,94	0,0036	0,44	0,98	0,99
Verrokki 2					
e(Altistunut)	0,021	0,0026	0,011	0,047	0,88
e(Verrokki 1)	0,012	0,0019	0,0077	0,024	0,67
e(Verrokki 2)	0,97	0,0045	0,93	0,98	0,99

Käänteistodennäköisyyspainokertoimien (IPW) lisäksi jokaiselle yksilölle laskettiin muitakin painokertoimia. Suurten painokertoimien eliminoimiseksi kokeiltiin ensin katkaistuja käänteistodennäköisyyspainokertoimia (TIPW). Katkaisu tehtiin luvussa 4.4 esitellyn optimaalisen menetelmän mukaan. Tällä menetelmällä 0,49 % masennuslääkkeille altistuneiden ryhmästä, 0,65 % verrokkiryhmästä 1 ja 2,34 % verrokkiryhmästä 2 sai painokertoimeksi nollan, joten kovin suurta osaa ei menetetty. Käänteistodennäköisyyspainokertoimilla kohdejoukkona oli koko tutkimusjoukko. Haluttiin kuitenkin tarkastella altistuksen vaikutusta myös pelkästään altistuneiden ryhmässä. Sitä varten muodostettiin ATT-painokertoimet niin, että $h(\boldsymbol{x}) = e(\text{Altistunut}, \boldsymbol{x})$. Lisäksi kokeiltiin kaltaistuspainoja (MW) ja päällekkäisyyspainoja (OW), jotka painottavat aineistossa epävarmoja tapauksia.

Taulukkoon 3 on merkitty painokertoimien tunnuslukuja. Koska kaikissa ryhmissä joillakin yksilöillä oman ryhmän propensiteettipistemäärä oli hyvin pieni, aineistossa esiintyi joitakin hyvin suuria käänteistodennäköisyyspainokertoimia $w_i = 1/e(z_i, \boldsymbol{x}_i)$. Katkaistussa käänteistodennäköisyyspainotuksessa suurimmat painot putosivat pois masennuslääkkeille altistuneiden ryhmästä sekä verrokkiryhmästä 1. Verrokkiryhmässä 2 sen sijaan suurin käänteistodennäköisyyspainokerroin on sama sekä ilman katkaisua että sen jälkeen. Vaikka katkaisussa yli 19 000 verrokkiryhmän 2 yksilöä sai painokertoimen 0, ei tämä suurin painokerroin siis ollut tässä joukossa. Katkaisu tehdään, jos summa $1/e(\text{Altistunut}) + 1/e(\text{Verrokki 1}) + 1/e(\text{Verrokki 2})$ ylittää määrätyn luvun. Näin käy, jos minkä tahansa ryhmän propensiteettipistemäärä on hyvin pieni. Verrokkiryhmässä 2 oli paljon yksilöitä, joilla oman ryhmän estimoitu todennäköisyys oli hyvin suuri. Tällöin kaksi muuta propensiteettipistemäärää jäivät pieniksi, ja katkaisuehto täyttyi. Painojen katkaisu ei siis hävitä ainoastaan suurimpia käänteistodennäköisyyspainoja vaan myös pienimpiä.

Määritelmän mukaisesti ATT-paino oli altistuneiden ryhmässä aina 1. Muissa ryhmissä se oli pienempi kuin 1, jos oman ryhmän propensiteettipistemäärä oli suurempi kuin altistusryhmän. Muussa tapauksessa ATT-paino oli suurempi kuin 1. Jos yksilön oman ryhmän propensiteettipistemäärä oli hyvin pieni ja altistuneiden

Taulukko 3: Painokertoimien jakaumat eri ryhmissä. Tarkasteltavana käänteistodennäköisyyspainot (IPW), katkaistut käänteistodennäköisyyspainot (TIPW), ATT-painot, kaltaistuspainot (MW) ja päällekkäisyyspainot (OW). Katkaistuille käänteistodennäköisyyspainoille tunnusluvut on ilmoitettu siinä joukossa, jossa painokerroin ei ole 0.

		Mediaani	**Minimi**	**1. desiili**	**9. desiili**	**Maksimi**
Altistunut	IPW	21,66	1,10	1,84	58,88	159,95
	TIPW	21,56	1,10	1,85	58,13	149,67
	ATT	1,00	1,00	1,00	1,00	1,00
	MW	0,47	0,0012	0,13	0,78	1,00
	OW	0,31	0,0012	0,10	0,42	0,75
Verrokki 1	IPW	48,41	1,45	5,07	109,00	373,45
	TIPW	48,02	1,45	5,06	107,40	181,01
	ATT	1,73	0,23	0,93	2,79	13,67
	MW	1,00	0,016	0,68	1,00	1,00
	OW	0,60	0,016	0,36	0,71	0,86
Verrokki 2	IPW	1,03	1,01	1,02	1,07	222,32
	TIPW	1,03	1,01	1,02	1,08	222,32
	ATT	0,022	0,0026	0,011	0,050	189,55
	MW	0,012	0,0019	0,0078	0,025	1,00
	OW	0,0076	0,0012	0,0047	0,016	0,97

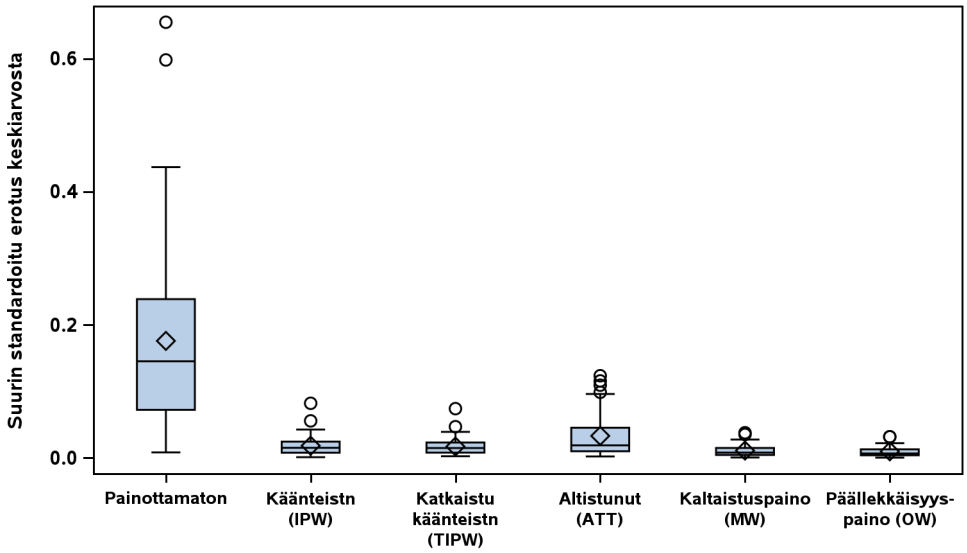
ryhmän pistemäärä lähellä yhtä, voi myös ATT-paino olla hyvin suuri. Verrokkiryhmässä 2 suurin ATT-paino olikin lähes 190, vaikka valtaosa ATT-painoista oli tässä ryhmässä melko lähellä nollaa. Verrokkiryhmässä 1 puolestaan pienimmät ATT-painot eivät ole yhtä pieniä eivätkä suurimmat läheskään yhtä suuria kuin verrokkiryhmässä 2, mikä viittaa siihen, että verrokkiryhmässä 1 altistuneiden ryhmän ja oman ryhmän propensiteettipistemäärät olivat useimmilla yksilöillä lähempänä toisiaan kuin verrokkiryhmässä 2.

Käänteistodennäköisyys- ja ATT-painokertoimien joukossa esiintyi suuriakin arvoja, mutta kaltaistuspainot ja päällekkäisyyspainot olivat, myöskin määritelmän mukaisesti, aina pienempiä tai yhtäsuuria kuin 1. Kaltaistuspainokertoimissa kallistusfunktiona oli $\min\{e(\text{Altistunut}), e(\text{Verrokki 1}), e(\text{Verrokki 2})\}$. Kun painokerroin saatiin jakamalla kallistusfunktio oman ryhmän propensiteettipistemäärällä, painokerroin oli tasan 1, jos oman ryhmän propensiteettipistemäärä oli kaikista pienin. Taulukosta 3 nähdään, että aineistossa verrokkiryhmässä 1 kaltaistuspainojen mediaani oli 1. Tämä tarkoittaa, että ainakin puolella tämän ryhmän yksilöistä $e(\text{Verrokki 1})$ oli kolmesta propensiteettipistemäärästä pienin. Päällekkäisyyspainokertoimet sen sijaan olivat aina pienempiä kuin 1. Päällekkäisyyspainokertoimet olivat taulukon perusteella muutenkin yleensä pienempiä kuin kaltaistuspainot.

5.3 Tasapainoisuus ja tehollinen otoskoko

Taustamuuttujien tasapainoisuutta painotetussa aineistossa arvioitiin luvussa 4.5 esitellyillä menetelmillä. Moniluokkaiset taustamuuttujat muutettiin joukoksi kak-

siluokkaisia muuttujia siten, että esimerkiksi äidin ikää kuvanneen neliluokkaisen muuttujan tilalle otettiin kolme kaksiluokkaista muuttujaa. Muutosten jälkeen tarkasteltavia muuttujia oli yhteensä 41. Tämän jälkeen jokaiselle muuttujalle laskettiin standardoitu erotus keskiarvosta (SEK) kaikissa kolmessa ryhmässä ja parittainen standardoitu erotus (PSE) kaikille kolmelle parille. Näistä luvuista valittiin suurimmat, max(SEK) ja max(PSE), taustamuuttujan tasapainoisuutta kuvaaviksi tunnusluvuiksi. Kuvassa 4 ovat jakaumat tunnusluvusta max(SEK) eri painokertoimilla. Siitä nähdään, että käänteistodennäköisyyspainokertoimet paransivat taustamuuttujien tasapainoisuutta selvästi verrattuna painottamattomaan aineistoon. Optimaalisesti katkaistuilla käänteistodennäköisyyspainoilla tasapainoisuus parani hiukan enemmän ja kaltaistus- ja päällekkäisyyspainoilla vielä lisää. ATT-painoilla ryhmien välille sen sijaan jäi hiukan enemmän eroja. Tulokset olivat samankaltaisia myös parittaisia standardoituja erotuksia tarkasteltaessa. Jos lukua 0,2 pidetään hyväksyttävän standardoidun erotuksen ylärajana, niin myös ATT-painoilla kaikki taustamuuttujat olivat kohtalaisen hyvin tasapainossa. Siksi ei pidetty tarpeellisena muokata propensiteettipistemäärien estimointiin käytettyä mallia.



Kuva 4: Taustamuuttujien tasapainoisuutta mittaavan standardoidun erotuksen keskiarvosta (SEK) jakauma eri painokertoimilla. Mitä pienempi SEK on, sitä lähempänä taustamuuttujan painotetut jakaumat eri altistusryhmissä ovat sen painotettua jakaumaa koko joukossa. Jokaiselle muuttujalle on laskettu erotus jokaisen altistusryhmän ja koko aineiston välillä eli yhteensä kolme erotusta. Näistä suurin on otettu mukaan kuvaajaan. Kuvaajassa yksi havainto vastaa yhtä taustamuuttujaa, ei yksilöä.

Kun katsottiin tarkemmin parittaisia standardoituja erotuksia, huomattiin että painottamattomalla aineistolla suurimmat erot ilmenivät yleensä vertailtaessa verrokkiryhmää 2 ja jompaa kumpaa kahdesta muusta ryhmästä. Myös altistuneiden ryhmän ja verrokkiryhmän 1 välillä oli joidenkin muuttujien kohdalla kohtalaisen suuri ero, mutta useimmat näistä erotuksista olivat myös painottamattomalla ai-

neistolla pienempiä kuin 0,2. Masennuslääkkeille ja pelkälle äidin masennukselle altistuneet muistuttivat siis taustamuuttujien osalta melko paljon toisiaan jo ennen painottamista.

Kun altistuksen vaikutusta estimoidaan painotetun aineiston avulla, esimerkiksi painotetuilla keskiarvoilla, saattaa estimaattien varianssi olla suurempi kuin jos käytettäisiin saman kokoista painottamatonta aineistoa. Tämä vaikuttaa tulosten tarkkuuteen ja voimaan. Suuremman varianssin vaikutusta voidaan arvioida tehollisen otoskoon (*effective sample size, ESS*) avulla. Masennuslääkkeille altistuneiden ryhmälle ja kahdelle verrokkiryhmälle laskettiin ryhmäkohtaiset tehokliset otoskoot artikkelin [5] mukaisesti niin, että

$$ESS_z = \frac{(\sum_{i=1}^{n} D_i(z)w_i)^2}{\sum_{i=1}^{n} D_i(z)w_i^2}.$$

Tarkkuuden heikkenemistä voidaan arvioida tarkastelemalla luvun ESS_z ja ryhmän z koon välistä suhdetta. Sekä lähellä nollaa olevat että hyvin suuret painokertoimet pienentävät tehollista otoskokoa.

Tehollinen otoskoko on tärkeä työkalu etenkin kokeellisen tutkimuksen suunnittelussa. Otoskoon halutaan olevan tarpeeksi suuri, jotta käsittelyn vaikutus voidaan havaita. Liian suuri otoskoko taas voi kuluttaa turhaan resursseja. Tässä tapauksessa, kun aineisto poimittiin rekistereistä, tehollisen otoskoon merkitys ei ole yhtä suuri. Haitallista olisi, jos tehollinen otoskoko jäisi vain muutamaan yksilöön. Nyt tutkimusaineisto kuitenkin oli niin suuri, että tulosten tarkkuus ei kärsinyt merkittävästi, vaikka painokertoimet olivatkin välillä hyvin suuria tai pieniä. Valituilla kriteereillä ei olisi edes voitu poimia suurempaa joukkoa, koska rekisterit kattoivat kaikki Suomessa valittuina vuosina syntyneet. Toisaalta suuren aineiston analysointi ei ollut juurikaan vaivalloisempaa kuin pienemmän, joten liian suurestakaan otoksesta ei tarvinnut huolehtia. Joissain tilanteissa painotettu aineisto voi kuitenkin heikentää analyysien voimakkuutta liikaa myös havainnoivissa tutkimuksissa, joten tehollista otoskokoa on hyvä tarkkailla myös näissä tilanteissa.

Taulukko 4: Tehollinen otoskoko eri ryhmissä ja eri painokertoimilla

	Altistunut	**Verrokki 1**	**Verrokki 2**
Painottamaton	32 562	16 499	824 256
Käänteistn	17 268	10 214	612 899
Katkaistu käänteistn	17 446	10 254	595 288
ATT	32 562	11 255	5 453
Kaltaistuspaino	26 116	15 889	126 645
Päällekkäisyyspaino	27 729	15 546	162 091

Tehokliset otoskoot masennuslääkkeille altistuneiden ryhmälle sekä molemmille verrokkiryhmille ovat taulukossa 4. Kuten jo taulukosta 3 huomattiin, verrokkiryhmästä 2 poistui käänteistodennäköisyyspainokertoimien katkaisussa yksilöitä, joiden painokerroin oli pieni eli lähellä yhtä. Suurimmat painot jäivät aineistoon. Tehollinen otoskoko kuitenkin on sitä suurempi, mitä lähempänä yhtä painokertoimet ovat.

Koska verrokkiryhmästä 2 muutettiin nolliksi juuri pieniä painokertoimia suurimpien pysyessä ennallaan, oli tämän ryhmän tehollinen otoskoko katkaistuilla käänteistodennäköisyyspainoilla pienempi kuin ilman katkaisua. Altistuneiden ryhmästä ja verrokkiryhmästä 1 poistettiin katkaisussa suurimmat painot, jolloin tehollinen otoskoko kasvoi hiukan.

Kaltaistus- ja päällekkäisyyspainoilla altistuneiden ja verrokkiryhmän 1 tehollinen otoskoko oli lähimpänä ryhmien todellista kokoa, kun taas verrokkiryhmässä 2 se oli selvästi pienempi kuin käänteistodennäköisyyspainokertoimilla. Tämä on ymmärrettävää, koska molemmat painokertoimet painottavat eniten epävarmoja tapauksia. Verrokkiryhmässä 2 kuitenkin oman ryhmän propensiteettipistemäärä oli useimmilla yksilöillä hyvin suuri, mikä näkyy sekä taulukossa 2 että kuvassa 3. Tämän vuoksi kaltaistus- ja päällekkäisyyspainokertoimet jäivät tässä ryhmässä usein lähelle nollaa, mikä pienensi myös tehollista otoskokoa. Myös ATT-painot olivat verrokkiryhmässä 2 keskimäärin lähempänä nollaa kuin kahdessa muussa ryhmässä. Mukana oli myös muutama suuri painokerroin. Suuret ja pienet painokertoimet yhdessä pienensivät tehollisen otoskoon alle sadasosaan ryhmän todellisesta koosta.

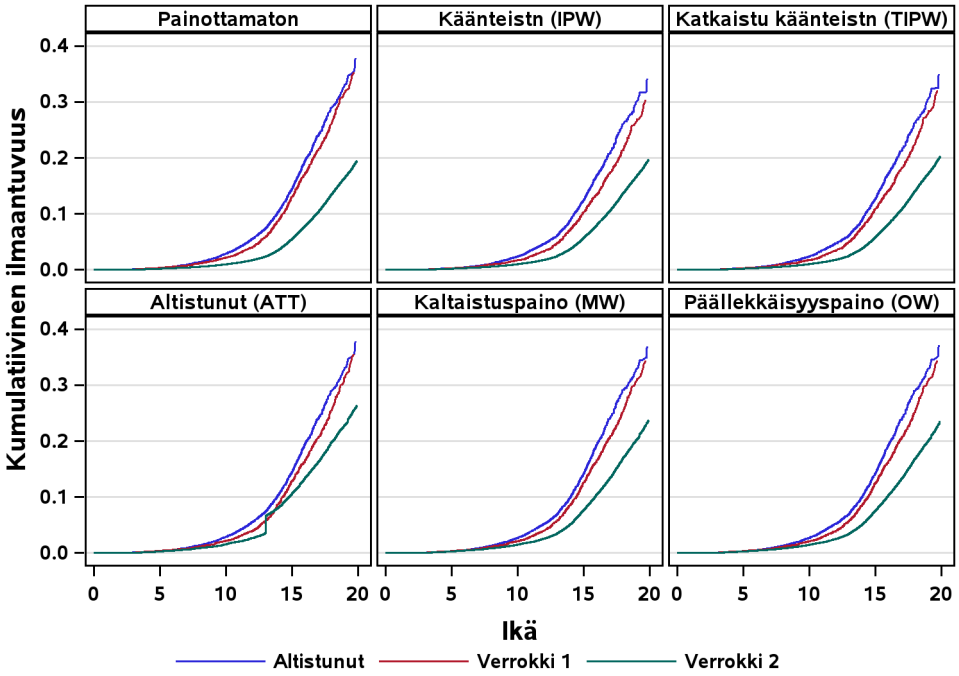
5.4 Analyysin tulokset

Vastemuuttujana oli henkilön sairastuminen masennukseen tai ahdistuneisuushäiriöön. Sairastuminen todettiin THL:n hoitoilmoitusrekisterissä esiintyvistä diagnooseista F32—F41 (ICD-10). Diagnooseja, jotka oli hoitoilmoitusrekisterin mukaan annettu alle kaksivuotiaille lapsille, ei huomioitu. Myös henkilölle itselleen määrätyn masennuslääkkeen katsottiin viittaavan sairastumiseen, joten lääkeostojen ajankohdat otettiin huomioon vaikka aiempaa diagnoosia ei olisi ollut. Lääkeostojen tiedot saatiin Kelan lääketoimitusrekisteristä.

Koska vanhimmista tutkimushenkilöistä oli kerätty tietoja 19 vuoden ajalta ja nuorimmista vain kahden vuoden, sairastumista masennuslääkkeille altistuneissa ja kahdessa verrokkiryhmässä tutkittiin elinaika-analyysin keinoin. Päätetapahtuma oli henkilön ensimmäinen masennus- tai ahdistuneisuusdiagnoosi tai masennuslääkeosto. Tapahtuma-aika T tarkoitti aikaa henkilön syntymästä päätetapahtumaan. Seuranta-aika jatkui vuoden 2018 loppuun. Jos henkilö ei siihen mennessä ollut kohdannut päätetapahtumaa, hänet merkittiin sensuroituneeksi. Sensuroitumiseksi katsottiin myös henkilön kuolema tai pysyvä muutto pois Suomesta.

Kumulatiivista ilmaantuvuutta $P(T \leq t)$ estimoitiin parametrittomasti painotetun Kaplan-Meierin estimaattorin avulla, joka esiteltiin luvussa 4.3.1. Estimaateista saadut ilmaantuvuuskäyrät eri painokertoimilla ovat kuvassa 5. Kaikilla painokertoimilla ilmaantuvuus kasvoi selvästi hitaiten verrokkiryhmässä 2 eli terveiden äitien lasten ryhmässä. Tuloksissa ei muutenkaan ollut kovin suurta eroa eri painokertoimien välillä. Käänteistodennäköisyyspainokertoimilla altistuneiden ryhmän ja verrokkiryhmän 1 estimoidut kumulatiivisen todennäköisyydet saada masennus- tai ahdistuneisuusdiagnoosi olivat koko ajan hiukan alhaisempia verrattuna painottamattomiin estimaatteihin, mutta verrokkiryhmän 2 kumulatiivisessa ilmaantuvuudessa ei näkynyt juuri eroa. ATT-, kaltaistus- ja päällekkäisyyspainoilla sen sijaan verrokkiryhmän 2 sairastumistodennäköisyys oli koko ajan hiukan suurempi kuin painottamattomassa aineistossa, mutta kahden muun ryhmän ilmaantuvuudet pysyivät

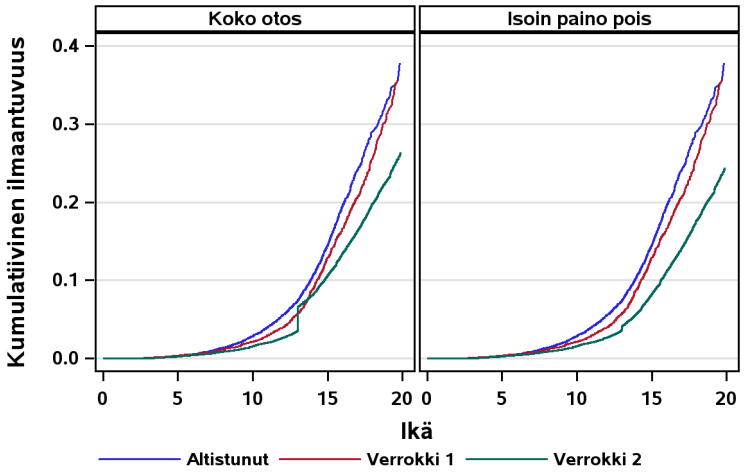
suunnilleen samalla tasolla kuin ilman painoja. Vaikuttaa siis siltä, että taustamuuttujien perusteella masennuslääkkeille altistuneita muistuttavassa joukossa kaikkien ryhmien sairastumistodennäköisyys on suurempi kuin koko tutkimusjoukossa. Sama pätee myös joukossa, jossa painotetaan niitä, jotka voisivat taustamuuttujien perusteella kuulua mihin altistusryhmään tahansa. Ryhmien välisissä eroissa on tulosten perusteella pientä vaihtelua kohdejoukkojen välillä, mutta ryhmien järjestys ei muutu: Masennuslääkkeille altistuneilla masennuksen ja ahdistuneisuushäiriön kumulatiivinen ilmaantuvuus on koko ajan suurin. Äidin masennukselle mutta ei masennuslääkkeille altistuneilla sairastumistodennäköisyys on hiukan pienempi ja masennusta sairastamattomien äitien lapsilla selvästi pienempi.



Kuva 5: Painotettu kumulatiivinen ilmaantuvuus ryhmittäin estimoituna eri painokertoimilla. Kumulatiivinen ilmaantuvuus tarkoittaa todennäköisyyttä, että yksilö sairastuu masennukseen tai ahdistuneisuushäiriöön annettuun ikään mennessä.

Selvimmin kuvassa 5 erottuu verrokkiryhmän 2 käyrä ATT-painoilla verrattuna muihin saman ryhmän käyriin. Siitä huomataan, miten paljon yksi poikkeavan suuri painokerroin voi vaikuttaa tuloksiin. Verrokkiryhmässä 2 ATT-painot olivat pääsääntöisesti hyvin pieniä: 90 %:lla aineistosta painokerroin oli alle 0,05. Yhdellä henkilöllä painokerroin kuitenkin oli suurempi kuin 180, mikä vastaa noin 0,55:ttä % koko verrokkiryhmän 2 ATT-painojen summasta. Kyseinen henkilö oli saanut ensimmäisen masennusdiagnoosinsa noin 13-vuotiaana, joten tapausten painotettu määrä tuolla hetkellä oli erityisen suuri. Sen takia sovellettu Kaplan-Meierin estimaattori teki kyseisen hetken kohdalla selvän hyppäyksen, joka näkyy kuvassa 5. Vaikka käänteistodennäköisyyspainokertoimetkin olivat suuria joillakin henkilöillä, yksittäiset kertoimet eivät kuitenkaan aiheuttaneet samanlaisia hyppyjä kuin ATT-painoilla. Toisin kuin ATT-painot, käänteistodennäköisyyspainot ovat aina suurem-

pia kuin 1. Koska verrokkiryhmään 2 kuului yli 824 000 henkilöä, muiden painokertoimet yhdessä estivät yksittäisten suurten painojen liian suuren vaikutuksen. Kuvasta 6 huomataan, että kun aineistosta poistettiin kaikkein suurimman painokertoimen saanut henkilö, hyppäys käyrässä poistui ja sen jälkeisetkin arvot jäivät hiukan alhaisemmiksi. Käyrä oli kuitenkin edelleen ylempänä kuin painottamattomalla aineistolla tai käänteistodennäköisyyspainoilla.



Kuva 6: Painotettu kumulatiivinen ilmaantuvuus ATT-painoilla. Vasemmanpuoleisessa kuvassa Kaplan-Meierin estimaatit on laskettu koko joukossa. Oikella suurimman painokertoimen saanut henkilö on jätetty pois verrokkiryhmästä 2, jolloin käyrässä näkynyt hyppäys poistui.

Koska pelkkien kuvaajien perusteella ei voitu päätellä, paljonko tilastollista epävarmuutta eroihin liittyi, tehtiin aineistolle myös luvussa 4.3.1 esitelty logrank-testi. Ensin testattiin nollahypoteesia, jonka mukaan ilmaantuvuus on sama kaikissa kolmessa ryhmässä kaikkina ajanhetkinä kahden ja 19 ikävuoden välissä. Testi tehtiin sekä painottamattomalle aineistolle että kaikilla viidellä painokertoimella. Testisuure noudatti χ^2-jakaumaa vapausastein 2, ja sen p-arvo oli kaikilla painokertoimilla pienempi kuin 0,0001. Voidaan siis todeta, että masennuksen ja ahdistuneisuuden kumulatiivisessa ilmaantuvuudessa on eroja masennuslääkkeille altistuneiden, äidin masennukselle mutta ei masennuslääkkeille altistuneiden sekä terveiden äitien lasten välillä. Ero on havaittavissa riippumatta siitä, ovatko kohdejoukkona kaikki vuosina 1999–2016 syntyneet (käänteistodennäköisyyspainot), pelkästään masennuslääkkeille altistuneet (ATT-painot) vai ne, jotka eivät taustamuuttujien perusteella kuulu automaattisesti mihinkään altistusryhmään (kaltaistus- ja päällekkäisyyspainot).

Se, että kaikki kolme ilmaantuvuuskäyrää eivät logrank-testin mukaan ole samanlaisia, ei kuitenkaan ollut yllättävää. Aiempien tutkimusten perusteella tiedetään, että jos äidillä on mieleterveyden ongelmia, niin lapsen psykiatristen häiriöiden todennäköisyys kasvaa. Myös kuvissa 5 ja 6 estimoitu kumulatiivisen ilmaantuvuuden käyrä on selvästi muita alempi niillä henkilöillä, joiden äidillä ei ollut ollenkaan masennusdiagnoosia tai masennuslääkeostoja ennen lapsen syntymää. Masennuslääkkeitä käyttäneiden ja masennusdiagnoosin saaneiden mutta ilman lääkkeitä olleiden

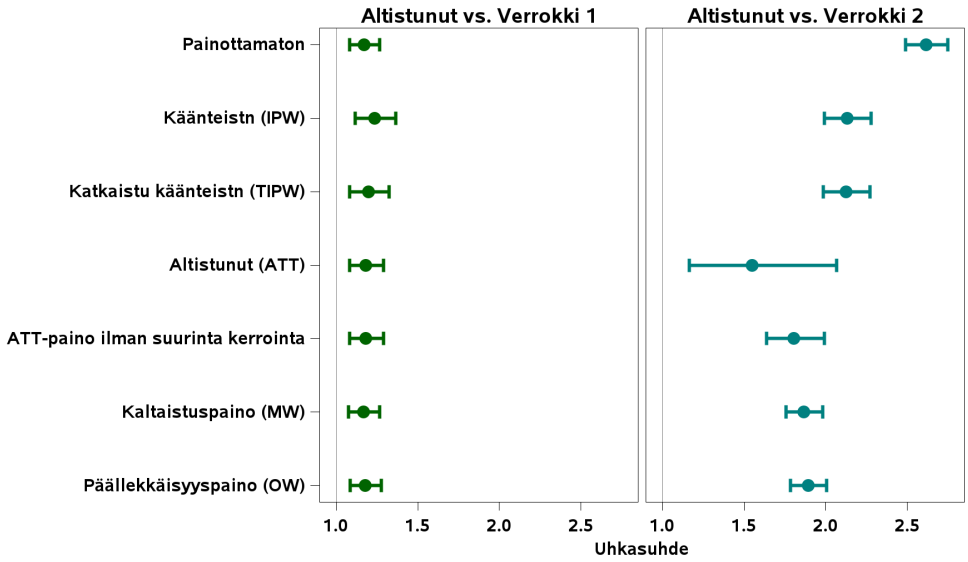
äitien lapsilla käyrät taas olivat melko lähellä toisiaan. Niinpä pidettiin tarpeellisena tarkastella myös pelkästään näiden kahden ryhmän kumulatiivisen ilmaantuvuuden samankaltaisuutta kontrastien avulla. Myös tämän kahden ryhmän testin p-arvo oli alle 0,0001 sekä painottamattomalla aineistolla että kaikilla viidellä painokertoimella. Testien perusteella siis lapsen masennuksen ja ahdistuksen kumulatiivisessa ilmaantuvuudessa on eroja riippuen siitä, onko masennusta tai siihen liittyvää häiriötä sairastanut äiti käyttänyt masennuslääkkeitä vai ei. Toisaalta χ^2-jakaumaa noudattavien testisuureiden tiedetään olevan suurilla aineistoilla usein tilastollisesti merkitseviä, vaikka ryhmien välinen ero olisi todellisuudessa vain pieni.

Logrank-testi testaa absoluuttista eroa estimoitujen Kaplan-Meierin käyrien välillä. Tämän lisäksi testattiin suhteellista eroa uhkafunktioissa verrannollisten uhkien mallin avulla. Malli toteutettiin sekä painottamattomalle aineistolle että eri kertoimilla painotetulle. Selittävänä tekijänä oli altistusryhmä kolmiluokkaisena muuttujana. Sairastumisikä ja sensurointi määriteltiin samoin kuin kumulatiivista ilmaantuvuutta laskettaessa. Estimaattien keskivirheet laskettiin vakaan sandwich-estimaattorin avulla, jotta painottamisen aiheuttama epävarmuus saatiin huomioitua. Lisäksi saman äidin lasten oletettiin muistuttavan toisiaan jossain määrin. Keskivirheen vakaan estimaattorin avulla voitiin huomioida myös tämä korrelaatio. Verrannollisten uhkien mallin avulla estimoitiin uhkasuhde altistuneiden ja verrokkiryhmän 1 sekä altistuneiden ja verrokkiryhmän 2 välillä. Lisäksi estimoitiin uhkasuhteen 95 %:n luottamusväli. Tulokset ovat taulukossa 5 ja kuvassa 7. Koska ATT-painoilla suurimman kertoimen oli havaittu vaikuttavan vahvasti estimoituun kumulatiiviseen ilmaantuvuuteen verrokkiryhmässä 2, estimoitiin altistuneiden ryhmän ja verrokkiryhmän 2 välinen uhkasuhde sekä ilman tätä suurinta painoa että sen kanssa.

Taulukko 5: Estimoidut uhkasuhteet (HR) ja niiden 95 %:n luottamusvälit (LV). Luottamusvälit on laskettu vakaan sandwich-estimaattorin avulla.

	Altistunut vs. Verrokki 1		**Altistunut vs. Verrokki 2**	
	HR	**95 %:n LV**	**HR**	**95 %:n LV**
Painottamaton	1,17	(1,08;1,26)	2,62	(2,49;2,75)
Käänteistn (IPW)	1,23	(1,12;1,37)	2,13	(1,99;2,28)
Katkaistu käänteistn (TIPW)	1,20	(1,08;1,32)	2,12	(1,99;2,27)
ATT-paino koko aineistolla	1,18	(1,08;1,29)	1,55	(1,16;2,07)
ATT-paino ilman suurinta painokerrointa	1,18	(1,08;1,29)	1,80	(1,64;1,99)
Kaltaistuspaino (MW)	1,17	(1,08;1,27)	1,86	(1,76;1,98)
Päällekkäisyyspaino (OW)	1,18	(1,07;1,28)	1,89	(1,79;2,01)

Taustamuuttujien tasapainoisuutta tarkasteltaessa huomattiin, että useimmissa taustamuuttujissa ei ollut kovin suuria eroja masennuslääkkeille altistuneiden ja verrokkiryhmän 1 välillä painottamattomassakaan aineistossa. Siitä saattaa johtua, että näiden ryhmien väliset uhkasuhteet olivat eri painokertoimilla melko lähellä toisiaan. Uhkasuhde vaihteli 1,17:n ja 1,23:n välillä eli sairastumistodennäköisyys kasvoi lyhyellä aikavälillä altistuneiden ryhmässä noin 20 % nopeammin kuin ver-



Kuva 7: Uhkasuhde eli hasardisuhde ja sen 95 %:n luottamusväli estimoituna eri tavoin painotetussa aineistossa. Luottamusvälit laskettiin vakaan sandwich-estimaattorin avulla, joka ottaa huomioon sekä painokertoimista johtuvan epävarmuuden että saman äidin lasten välisen korrelaation.

rokkiryhmässä 1 kaikissa tarkastelluissa kohdejoukoissa. Myös 95 %:n luottamusvälit olivat hyvin samankaltaiset kaikilla painokertoimilla. Luottamusvälien alarajat olivat suurempia kuin yksi eli tasolla 0,05 altistuneiden ryhmän ja verrokkiryhmän 1 ero on tilastollisesti merkitsevä. Eri asia on, pidetäänkö noin 20 %:n eroa merkittävänä vai ei. Uhkasuhteista ei voida päätellä varsinaisia sairastumistodennäköisyyksiä, vaan kysessä on suhteellinen ero kahden ryhmän välillä. Koska painotettu verrannollisten uhkien malli estimoi marginaalista vaikutusta, uhkasuhteet kertovat todennäköisyyksien keskimääräisistä eroista koko kohdejoukossa, ei yhden yksilön kohdalla.

Altistuneiden ja verrokkiryhmän 2 välinen ero oli huomattavasti suurempi, ja myös eri painokertoimien välillä oli enemmän eroja. Kaikkein suurin estimoitu uhkasuhde (2,62) oli painottamattomalla aineistolla. Poikkeavan suuri ATT-paino vaikutti vahvasti myös estimoituun uhkasuhteeseen. Muuten uusien sairastuneiden määrä kasvoi lyhyellä aikavälillä selvästi hitaammin terveiden äitien lapsilla kuin altistuneilla, mutta yksi suuren painoarvon saanut sairastunut yksilö terveiden äitien lasten joukossa painoi eroa pienemmäksi. Kyseisen yksilön ollessa mukana uhkasuhde oli 1,55 kun taas ilman sitä suhde oli 1,80. Jälkimmäinen luku oli huomattavasti lähempänä muilla painokertoimilla saatuja tuloksia. Myös estimaatin keskivirhe oli suurimman painon kanssa suurempi kuin ilman sitä, mikä näkyi leveämpänä luottamusvälinä. Kummassakin tapauksessa estimoitu uhkasuhde oli pienempi kuin muilla painokertoimilla. Kaltaistus- ja päällekkäisyyspainoilla saatujen estimaattien (1,86 ja 1,89) ero verrattuna ATT-painoon ilman suurinta kerrointa ei tosin ollut kovin suuri. Käänteistodennäköisyyspainokertoimilla uhkasuhde sen sijaan oli 2,13 (katkaisun kanssa 2,12). Tulosten perusteella vaikuttaa siis siltä, että koko joukossa ma-

sennukseen tai ahdistuneisuushäiriöön sairastumisen uhka eli hasardi on masennuslääkkeille altistuneiden ryhmällä vähän yli kaksinkertainen verrattuna niihin, joiden äidit eivät ole sairastaneet masennusta tai käyttäneet masennuslääkkeitä raskauden aikana tai ennen sitä. Joukossa, jossa korostuvat epävarmimmat yksilöt, vastaava ero on vähän pienempi mutta silti selvä: altistuneilla uhka on noin 90 % suurempi kuin altistumattomilla.

6 Yhteenveto ja pohdintaa

Tässä tutkielmassa perehdyttiin propensiteettipistemäärään eli taustamuuttujien perusteella estimoituun todennäköisyyteen, että yksilö kuuluu tiettyyn altistusryhmään. Propensiteettipistemäärä tiivistää taustamuuttujien informaation yhteen tai muutamaan lukuun altistusryhmien määrästä riippuen. Havainnoivassa tutkimuksessa propensiteettipistemääriä käytetään tasapainottamaan taustamuuttujien jakaumia eri altistusryhmien välillä. Kaltaistamalla, osittamalla tai painottamalla tutkimusaineistoa propensiteettipistemäärien avulla pyritään tilanteeseen, jossa eri tavalla altistuneet eivät eroa toisistaan taustamuuttujien suhteen. Tutkielmassa keskityttiin erityisesti propensiteettipistemääriin perustuviin painokertoimiin. Kun tutkimusaineistoa painotetaan näillä kertoimilla, muodostuu synteettinen aineisto, jossa taustamuuttujat ovat kaikissa altistusryhmissä jakautuneet samoin kuin valitussa kohdejoukossa. Tämän jälkeen altistuksen vaikutusta voidaan arvioida vertailemalla eri ryhmien vasteita tässä painotetussa aineistossa. Kohdejoukoksi voidaan valita esimerkiksi koko tutkittava populaatio, yksi altistusryhmistä tai joukko, jossa korostuvat erityisesti yksilöt, jotka voisivat taustamuuttujien perusteella kuulua mihin altistusryhmään tahansa. Yleisimmin propensiteettipistemääriä käytetään tutkimuksissa, joissa tutkittavat voivat kuulua jompaan kumpaan kahdesta ryhmästä: altistuneisiin tai altistumattomiin. Menetelmät voidaan kuitenkin yleistää koskemaan myös tapauksia, joissa altistusryhmiä on enemmän kuin kaksi.

Perinteinen tapa huomioida taustamuuttujien vaikutus on lisätä ne selittävinä muuttujina regressiomalliin. Jos muuttujia on paljon, mallissa on paljon estimoitavia parametreja, mikä voi aiheuttaa laskennallisia ongelmia. Voi myös olla, että joitain taustamuuttujien yhdistelmiä ei esiinny aineistossa. Koska regressiomallit yleensä estimoivat yhden selittävän tekijän vaikutusta siinä tapauksessa, että kaikkien muiden arvot pysyvät samoina, voi yhdistelmien puuttuminen aiheuttaa epävarmuutta tuloksissa. Propensiteettipistemäärien etu onkin, että yksi tai muutama luku edustaa kaikkien taustamuuttujien vaikutusta. Muutaman taustamuuttujan tiivistämisellä ei todennäköisesti ole kovin suurta merkitystä, mutta suuren muuttujajoukon kanssa propensiteettipistemäärät voivat olla hyvin hyödyllisiä.

Propensiteettipistemääriä hyödyntävien menetelmien joukossa painokertoimien hyvä puoli on, että valitsemalla erilaisia kallistusfunktioita voidaan määritellä, missä kohdejoukossa altistuksen vaikutusta tarkastellaan. Kaltaistaminen ja osittaminen eivät mahdollista tällaista valintaa. Lisäksi jos altistusryhmiä on monta, voi kaltaistaminen tai osittaminen kaikkien propensiteettipistemäärien perusteella olla vaikeaa. Painokertoimien laskemista useat ryhmät sen sijaan eivät hankaloita yhtä paljon.

Painokertoimia käytettäessä ongelma on, että oman ryhmän propensiteettipistemäärän ollessa pieni voi painokerroin olla hyvin suuri. Tällöin yksi poikkeava yksilö voi vaikuttaa tuloksiin paljonkin. Tämän tutkielman esimerkissä tosin havaittiin, että muutama erityisen suuri painokerroin ei vaikuta tuloksiin kovin suuresti, jos tutkimusjoukko on suuri ja kaikkien painokertoimet ovat suurempia kuin 1. Tämä toteutuu käänteistodennäköisyyspainokertoimilla (IPW). Toisaalta huomattiin, että jos yhdessä ryhmässä on paljon lähellä nollaa olevia painokertoimia, voi yksikin erityisen suuri paino muuttaa tuloksia huomattavasti. Tämä näkyi ATT-

painoilla estimoiduista tuloksista verrokkiryhmän 2 kohdalla sekä selvänä hyppäyksenä estimoidussa kumulatiivisessa ilmaantuvuudessa että muutoksena uhkasuhteen estimaatissa ja sen keskivirheessä. ATT-painoilla suuri aineisto ei siis tasapainota poikkeavien yksilöiden vaikutusta yhtä hyvin kuin käänteistodennäköisyyspainoilla. Painokertoimien jakaumaa on näin ollen syytä tarkastella ennen kuin altistuksen vaikutusta ryhdytään arvioimaan. Hyvin suurten painokertoimien aiheuttamia ongelmia voidaan pyrkiä ehkäisemään käänteistodennäköisyyspainojen katkaisulla tai käyttämällä kallistusfunktiota, joka pakottaa kaikki painokertoimet nollan ja yhden välille (kaltaistus-, päällekkäisyys- ja entropiapainot). Molempien menetelmien kohdalla on kuitenkin muistettava, että niiden kohdejoukko ei ole koko tutkimusjoukko. Tämä täytyy ottaa huomioon tuloksia tulkittaessa.

Kaikissa propensiteettipistemääriä hyödyntävissä menetelmissä haasteena on taustamuuttujien valinta. Sekoittumattomuusoletuksen mukaan havaitsemattomia sekoittavia tekijöitä ei pitäisi olla, mutta tätä on käytännössä hyvin vaikea varmistaa. Myös luvun 5 esimerkin yhteydessä voidaan pohtia, olisivatko jotkin muut tekijät voineet vaikuttaa masennuslääkkeiden käyttöön. On esimerkiksi mahdollista, että lääkkeitä käyttäneet äidit kärsivät vakavammasta masennuksesta kuin diagnoosin saaneet mutta ilman lääkkeitä olleet äidit. Sairauden vaikeudesta on kuitenkin vaikea saada tietoa pelkkien rekisteriaineistojen perusteella. Rekisterit eivät myöskään kerro sitä, saiko äiti masennukseen jotain muuta hoitoa kuten terapiaa. Havaitsemattomien sekoittamien tekijöiden ongelma on toki olemassa myös silloin, kun taustamuuttujat ovat mukana regressiomallissa selittävinä tekijöinä.

Esimerkissä käytettyä tutkimusjoukkoa olisi mielenkiintoista tarkastella myöhemmin uudestaan pidemmällä seuranta-ajalla. Nyt altistuneiden ryhmässä painottuivat enemmän myöhempinä vuosina syntyneet, jotka olivat seurannan päättyessä vuonna 2018 hyvin nuoria. Pienillä lapsilla masennus- ja ahdistuneisuusdiagnoosit kuitenkin ovat varsin harvinaisia. Myös kumulatiivisen ilmaantuvuuden käyristä nähtiin, että sairastumistodennäköisyys alkaa todella kasvaa noin 13-vuotiailla. Näin ollen altistuneiden ryhmässä oli paljon sellaisia, jotka olivat seurannan päättyessä vielä liian nuoria sairastumaan, kun taas teini-ikään ehtineitä oli enemmän terveiden äitien lasten joukossa. Olisikin mielenkiintoista, jos tutkimus pystyttäisiin toistamaan myöhemmin niin, että seurantaa jatkettaisiin esimerkiksi kymmenellä vuodella vuoteen 2028 asti. Tällöin nuorimmistakin lapsista saataisiin tietoa 12-vuotiaiksi asti.

Viitteet

[1] Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. Statistics in Medicine, 26:20–36.

[2] Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70:41–55.

[3] Greenland, S.; Robins, J. M. & Pearl, J. (1999). Confounding and Collapsibility in Causal Inference. Statistical Science, 14(1):29–46.

[4] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46:399–424.

[5] McCaffrey, D. F.; Griffin B. A.; Almirall, D.; Slaughter, M. E.; Ramchand, R. & Burgette L. F. (2013). A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. Statistics in Medicine, 32:3388–3414.

[6] Imbens, G. W.(2000). The Role of the Propensity Score in Estimating Dose-Response Functions. Biometrika, 87:706–710.

[7] Li, F. (2019). Propensity Score Weighting for Causal Inference with Multiple Treatments. The Annals of Applied Statistics, 13(4):2389–2415.

[8] Yang, S.; Imbens, G. W.; Cui, Z.; Faries, D. E. & Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. Biometrics, 72:1055–1065.

[9] Lopez, M. J. & Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. Statistical Science, 32:432–454.

[10] Pearl, J. (2000). Causality: Models, Reasoning, and Inference. 2nd edition. Cambridge: Cambridge University Press. s. 348–352.

[11] Rosenbaum, P. R. (1987). Model-based direct adjustment. The Journal of American Statistician, 82:387–394.

[12] Morgan, S. L. & Todd, J. L. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. Sociological Methodology, 38:231–281.

[13] Yoshida, K.; Hernández-Díaz, S.; Solomon, D. H.; Jackson, J. W.; Gagne, J. J. ; Glynn, R. J. & Franklin, J. M. (2017). Matching Weights to Simultaneously Compare Three Treatment Groups. Epidemiology, 28(3):387–395.

[14] Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Statistics in medicine, 33:1242–1258.

[15] Li, F; Morgan, K. L. & Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. Journal of the American Statistical Association, 113:390–400.

[16] Hirano, K.; Imbens, G. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4):1161–1189.

[17] Shu, D.; Young, J. G.; Toh, S. & Wang, R. (2021). Variance estimation in inverse probability weighted Cox models. Biometrics, 77(3):1101–1117.

[18] Xie, J. & Liu, C. (2005). Adjusted Kaplan-Meier Estimator and Log-Rank Test with Inverse Probability of Treatment Weighting for Survival Data. Statistics in medicine 24:3089–3110.

[19] Sugihara, M. (2010). Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score. Pharmaceutical statistics, 9(1):21–34.

[20] Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B, Methodological,34(2):187–220.

[21] Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. Biometrika, 79(1):139—147.

[22] Buchanan, A. L.; Hudgens, M. G.; Cole, S. R.; Lau, B. & Adimora, A. A. (2014). Worth the weight: using inverse probability weighted Cox models in AIDS research. AIDS Research and Human Retroviruses, 30(12):1170–1177.

[23] Zhou, Y.; Matsouaka, R. A. & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. Statistical Methods in Medical Research, 29(12):3721–3756.

[24] Li, L. & Greene, T. (2013). Weighting Analogue to Pair Matching in Propensity Score Analysis. The International Journal of Biostatistics, 9(2):215—234.

[25] Robins, J. M.; Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89:846—66.

[26] Glynn, A. & Quinn, K. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. Political Analysis, 18:36–56.

[27] Tampereen yliopistollinen sairaala. Masennus- eli mielialalääkkeet. Potilasohje 90.00.02. Saatavilla verkossa: https://www.tays.fi/fi-FI/Ohjeet/Potilasohjeet/Psykiatria/Masennus_eli_mielialalaakkeet(25726) (Viitattu 18.3.2023).

[28] Ansorge, M. S.; Zhou, M.; Lira, A.; Hen, R. & Gingrich J. A. (2004). Early-life blockade of the 5-HT transporter alters emotional behavior in adult mice. Science, 306:879—881.

[29] Malm, H.; Brown, A. S.; Gissler, M.; Gyllenberg, D.; Hinkka-Yli-Salomäki, S.; McKeague, I.W.; et al. (2016). Gestational exposure to selective serotonin reuptake inhibitors and offspring psychiatric disorders: A National Register-Based Study. J Am Acad Child Adolesc Psychiatry, 55:359–366.

[30] Moreau, A. L.; Voss, M.; Hansen, I.; Paul, S. E.; Barch, D. M.; Rogers, C. E. & Bogdan, R. (2022). Prenatal Selective Serotonin Reuptake Inhibitor Exposure, Depression, and Brain Morphology in Middle Childhood: Results From the ABCD Study, Biological Psychiatry Global Open Science.

[31] Furu, K.; Kieler, H.; Haglund, B.; Engeland, A.; et al. (2015). Selective serotonin reuptake inhibitors and venlafaxine in early pregnancy and risk of birth defects: population based cohort study and sibling design. BMJ, 350:h1798.

[32] Suomalaisen Lääkäriseuran Duodecimin ja Suomen Psykiatriyhdistys ry:n asettama työryhmä. Depressio. Käypä hoito -suositus. 29.6.2022 (päivityksessä). Suomalainen Lääkäriseura Duodecim. Saatavilla verkossa: https://www.kaypahoito.fi/hoi50023#R97 (Viitattu 18.3.2023).

A Propensiteettipistemäärien estimoinnissa käytetyt taustamuuttujat

Lapsen sukupuoli	Tyttö / Poika
Lapsen syntymävuosi	Jatkuva
Äidin ikä lapsen syntyessä	19 tai alle / 20–29 / 30–39 / 40 tai yli
Isän ikä lapsen syntyessä	19 tai alle / 20–29 / 30–39 / 40 tai yli / Tieto puuttuu
Äidin asuinseutu lapsen syntyessä	Kaupunki / Taajama / Maaseutu / Tieto puuttuu
Äiti avio- tai avoliitossa	Kyllä / Ei
Äidin sosioekonominen asema	Ylempi toimihenkilö / Alempi toimihenkilö / Työntekijä / Muu (mm. opiskelijat, työttömät) / Tieto puuttuu
Isän sosioekonominen asema	Ylempi toimihenkilö / Alempi toimihenkilö / Työntekijä / Muu (mm. opiskelijat, työttömät) / Tieto puuttuu
Äidin koulutus lapsen syntymävuotena	Vain peruskoulu / Toinen aste / Korkea-aste
Isän koulutus lapsen syntymävuotena	Vain peruskoulu / Toinen aste / Korkea-aste / Tieto puuttuu
Äiti synnyttänyt aiemmin	Kyllä / Ei / Tieto puuttuu
Äiti tupakoinut raskauden aikana	Kyllä / Ei / Tieto puuttuu
Hedelmöityshoitoja käytetty kyseisessä raskaudessa	Kyllä / Ei
Äidin painoindeksi raskauden alussa	Alle 18,5 / 18,5–24,9 / 25–29,9 / 30 tai yli / Tieto puuttuu
Äiti ostanut sikiövaurioita aiheuttavia (teratogeenisiä) lääkkeitä raskauden aikana tai enintään kuukausi ennen sitä	Kyllä / Ei
Äiti ostanut ahdistusta lievittäviä tai epilepsialääkkeitä raskauden aikana tai enintään kuukausi ennen sitä	Kyllä / Ei
Äidillä diagnosoitu lääkkeiden tai päihteiden käytön aiheuttama häiriö (ICD-10: F10–F19) ennen lapsen syntymää	Kyllä / Ei
Äidillä diagnosoitu muu kuin päihteisiin tai masennukseen liittyvä psykiatrinen häiriö (ICD-10: F50–F99) ennen lapsen syntymää	Kyllä / Ei

Isällä diagnosoitu skitsofrenia, mania tai kaksisuuntainen mielialahäiriö (ICD-10: F20–F31) ennen lapsen syntymää	Kyllä / Ei / Tieto puuttuu
Isällä diagnosoitu masennus tai ahdistuneisuushäiriö (ICD-10: F32–F41) ennen lapsen syntymää	Kyllä / Ei / Tieto puuttuu
Isällä diagnosoitu muu psykiatrinen häiriö ennen lapsen syntymää	Kyllä / Ei / Tieto puuttuu
Sisaruksella diagnosoitu jokin psykiatrinen häiriö (ICD-10: F10–F99) ennen lapsen syntymää	Kyllä / Ei

B SAS-koodia

```
***
Lasketaan propensiteettipistemäärät multinomisella
    logistisella regressiolla
***;
* exposure_group = altistusryhmä kolmiluokkaisena muuttujana
                (arvot: Altistunut/Verrokki1/Verrokki2);
proc logistic data=data1;
class gender age_m_4cat age_f_4cat population marital_status
                ses_m ses_f educ_m educ_f prev_births smoking artif
                bmi_cat teratogens medicine_m subs_m other_nodep_m
                f20_31_f f32_41_f otherpsych_f psych_sibl;
model exposure_group = gender birthyear age_m_4cat age_f_4cat
                population marital_status ses_m ses_f educ_m educ_f
                prev_births smoking artif bmi_cat
                teratogens medicine_m subs_m other_nodep_m
                f20_31_f f32_41_f otherpsych_f psych_sibl
/ link=glogit;
output out=pred_prob predicted=ps;
run;
* Output-rivi tuottaa datan, jossa on jokaiselle yksilölle kolmen
        altistusryhmän estimoidut todennäköisyydet
        eli propensiteettipistemäärät ;


***
Muokataan logistisesta mallista saatua dataa
***;
data depression_psdata;
set pred_prob;
by lapsi_tnro; * Lapsen yksilökohtainen tunniste ;

* Saadussa datassa kolme prop.pistemäärää omilla riveillään.
* Siirretään kaikki samalle ;
retain ps_exposed ps_comp1 ps_comp2;
if first.lapsi_tnro then do;
        ps_exposed=0;
        ps_comp1=0;
        ps_comp2=0;
end;
if _level_=1 then ps_exposed=ps;*e(Altistunut);
if _level_=2 then ps_comp1=ps;  *e(Verrokki 1);
if _level_=3 then ps_comp2=ps;  *e(Verrokki 2);

if last.lapsi_tnro; * Jätetään dataan vain yksi rivi per lapsi ;
```

```
* Katkaisua varten ;
inv_propensity_sum=sum(1/ps_exposed,1/ps_comp1,1/ps_comp2);
* Kallistusfunktiot ;
* Katkaistu käänteistn.
        Katkaisupiste laskettu optimaalisen menetelmän mukaan ;
if inv_propensity_sum<=273 then h_tipw=1;else h_tipw=0;
* ATT-paino (kallistusfunktio on altistuneiden ryhmän prop.pist.) ;
h_att=ps_exposed ;
h_mw=min(ps_exposed,ps_comp1,ps_comp2); * Kaltaistuspaino ;
h_ow=1/sum(1/ps_exposed,1/ps_comp1,1/ps_comp2); * Päällekkäisyyspaino;

* Käänteistn ;
if exposure_group=1 then IPW=1/ps_exposed;
else if exposure_group=2 then IPW=1/ps_comp1;
else if exposure_group=3 then IPW=1/ps_comp2;

* Muut painokertoimet ;
TIPW=h_tipw*IPW;
ATT=h_att*IPW;
MW=h_mw*IPW;
OW=h_ow*IPW;

* Painokertoimien toiset potenssit (ESS:n laskemista varten);
IPW2=IPW**2;
TIPW2=TIPW**2;
ATT2=ATT**2;
MW2=MW**2;
OW2=OW**2;

run;


***
Taustamuuttujien tasapainoisuuden tarkastelu
***;
* Tässä vaiheessa kaikki moniluokkaiset muuttujat on muunnettu
        dummy-muuttujien joukoksi;

* Esimerkkinä ATT-painot ;
* Painotetut keskiarvot altistusryhmittäin;
proc means data=depression_psdata noprint ;
var gender birthyear age_m_19 age_m_30_39 age_m_40
        age_f_19 age_f_30_39 age_f_40 father_missing
        pop_semiurban pop_rural pop_unknown marital_status
        ses_m_lowerw ses_m_blue ses_m_other ses_m_missing
```

```
        ses_f_lowerw ses_f_blue ses_f_other ses_f_missing
        educ_m_basic educ_m_2nd educ_f_basic educ_f_2nd
        prevb_0 prevb_missing smoke_yes smoke_missing artif
        bmi_18 bmi_25 bmi_30 bmi_miss teratogens medicine_m
        subs_m other_nodep_m f20_31_f f32_41_f otherpsych_f psych_sibl;
class exposure_group;
weight att; * Tässä valitaan painokerroin ;
output out=groupmean1 mean=/autoname;
run;
* Edellä output-komento antaa keskiarvot.
        Data on leveässä muodossa, transponoidaan se niin
        että jokaisen muuttujan ka omalla rivillään ;
proc transpose data=groupmean1 out=groupmean_att suffix=_mean;
where _type^=0;
var gender_mean--psych_sibl_mean;
id exposure_group;
run;
* Erotetaan muuttujien nimet yhteen sarakkeeseen ;
data groupmean_att;
length Covariate $ 20;
set groupmean_att;
Covariate=substr(_name_,1,find(_name_,'_Mean')-1);
run;
* Samaan tapaan lasketaan myös painotetut keskiarvot koko joukossa
        ja hajonnat;
* Tallennetaan ne datoihin totalmean_att ja groupsd;

* Lasketaan tasapainoisuusarvot ;
data balance_att;
merge groupmean_att totalmean_att groupsd;
by Covariate;
drop _name_;

*Standardisoidut erotukset keskiarvosta;
SEK1=abs(altistunut_mean-total_mean)/total_sd;
SEK2=abs(verrokki1_mean-total_mean)/total_sd;
SEK3=abs(verrokki2_mean-total_mean)/total_sd;
max_SEK=max(SEK1,SEK2,SEK3);

*Parittaiset erotukset;
PSE12=abs(altistunut_mean-verrokki1_mean)/total_sd;
PSE13=abs(altistunut_mean-verrokki2_mean)/total_sd;
PSE23=abs(verrokki1_mean-verrokki2_mean)/total_sd;
max_PSE=max(PSE12,PSE13,PSE32);
run;
```

```
* Samalla lailla lasketaan tasapainoisuusluvut myös muilla painoilla;
* Kootaan kaikki tulokset dataan nimeltä balance.
* Lisätään siihen sarake 'paino', jossa painokertoimen nimi;
* Piirretään laatikkokuvaaja;
proc template;
define statgraph balancebox;
begingraph;
layout overlay/
        xaxisopts=(discreteopts=(tickvaluefitpolicy=split))
        yaxisopts=(label='Suurin standardisoitu erotus keskiarvosta');
boxplot y=max_sek x=paino;
endlayout;
endgraph;
end;
run;
proc sgrender data=balance template=balancebox;run;


***
Tehollisen otoskoon laskeminen
***;
proc means data=depression_psdata;
output out=summat sum=/autoname;
var IPW IPW2 TIPW TIPW2 ATT ATT2 MW MW2 OW OW2;
class exposure_group;
run;
data ess;
set summat;
where _type_=1;
* Tehollliset otoskoot kaikille painoille ;
ESS_IPW=(IPW_sum**2)/IPW2_sum;
ESS_TIPW=(TIPW_sum**2)/TIPW2_sum;
ESS_ATT=(ATT_sum**2)/ATT2_sum;
ESS_MW=(MW_sum**2)/MW2_sum;
ESS_OW=(OW_sum**2)/OW2_sum;
run;


***
Kaplan-Meierin estimaatit
***;

* Esimerkkinä käänteistn-painot (IPW);
* end_time = ikä sairastumisen tai sensuroinnin hetkellä ;
* Outcome=1, jos henkilö on sairastunut. Muuten Outcome=0;
proc lifetest data=depression_psdata method=km;
```

```
time end_time*Outcome(0);
* Estimaatit erikseen kaikille altistusryhmille.
* Logrank-testi halutaan myös parittaisena
                altistuneille ja verrokkiryhmälle 1,
                joten asetetaan verrokki 1 kontrolliluokaksi;
strata exposure_group/test=logrank diff=control('Verrokki1');
weight=IPW; * Painotetut estimaatit;
* Tallennetaan estimaatit ;
ods output ProductLimitEstimates=ple_ipw(where=(survival^=.));
run;

* Samalla lailla muille painoille ja painottamattomana;
* Yhdistetään tulokset yhteen dataan;
data surv;
set ple_unw(in=a) ple_ipw(in=b) ple_tipw(in=c)
        ple_att(in=d) ple_mw(in=e) ple_ow(in=f);
if a then paino=1;
if b then paino=2;
if c then paino=3;
if d then paino=4;
if e then paino=5;
if f then paino=6;
format paino painof.;
* 'painof' on formaatti, jossa jokaista numeroa
                vastaa painokertoimen nimi;
run;
* Kuvaajat ;
proc sgpanel data=surv;
panelby paino/novarname columns=3 onepanel;
series x=end_time y=failure/
        group=exposure_group lineattrs=(pattern=solid);
colaxis label='Ikä';
rowaxis label='Kumulatiivinen ilmaantuvuus' grid;
run;


***
 Verrannollisten uhkien malli
***;

* Esimerkkinä kaltaistuspaino (MW);
proc phreg data=depression_psdata covs(aggregate);
                * covs-komento laskee robustit keskivirheet;
class id_mother exposure_group/ref=last order=internal;
model end_time*Outcome(0)=exposure_group;
weight=MW;
```

```
hazardratio exposure_group/diff=pairwise;
* Robusteissa keskivirheissä huomioidaan saman äidin lasten korrelaatio;
id id_mother;
* Tallennetaan tulokset;
ods output hazardratios=HR_mw;
run;

* Lasketaan uhkasuhteet myös muille painokertoimille
                ja kootaan kaikki dataan hazard;
* Kuvaajaa varten erotetaan omiksi sarakkeikseen uhkasuhteet
                altistuneilla verrattuna kahteen muuhun;
data comp1 comp2;
set hazard;
if find(description,'Altistunut vs Verrokki1') then output comp1;
if find(description,'Altistunut vs Verrokki2') then output comp2;
run;
data hr_yhd;
merge comp1(rename=(hazardratio=hr_comp1
                                    robustwaldlower=lower_comp1
                                    robustwaldupper=upper_comp1))
          comp2(rename=(hazardratio=hr_comp2
                                    robustwaldlower=lower_comp2
                                    robustwaldupper=upper_comp2));
by paino;
keep paino hr_: lower: upper: ;
run;

* Kuvaaja uhkasuhteista luottamusväleineen;
proc template;
define statgraph hr_forestplot;
begingraph;
layout lattice/columns=2 rowdatarange=union columndatarange=unionall ;
column2headers; entry 'Altistunut vs. Verrokki 1';
                                entry 'Altistunut vs. Verrokki 2';
endcolumn2headers;
rowaxes; rowaxis/display=(line ticks tickvalues);
endrowaxes;
layout overlay/xaxisopts=(display=(line ticks tickvalues));
        scatterplot x=hr_comp1 y=paino/
                        xerrorlower=lower_comp1 xerrorupper=upper_comp1;
        referenceline x=1;
endlayout;
layout overlay/xaxisopts=(display=(line ticks tickvalues));
        scatterplot x=hr_comp2 y=paino/
                        xerrorlower=lower_comp2 xerrorupper=upper_comp2;
        referenceline x=1;
```

```
endlayout;
sidebar/align=bottom; entry 'Uhkasuhde';endsidebar;

endlayout;
endgraph;
end;
run;

proc sgrender data=hr_yhd template=hr_forestplot;run;
```