Lapsiluvun ja sepelvaltimotaudin kausaalisuhteen yleistys yhdysvaltalaisesta populaatiosta jyväskyläläiseen keski-ikäisillä naisilla

Katariina Laaksovuori

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
Kevät 2023

Katariina Laaksovuori, *Lapsiluvun ja sepelvaltimotaudin kausaalisuhteen yleistys yhdysvaltalaisesta populaatiosta jyväskyläläiseen keski-ikäisillä naisilla*
Tilastotieteen pro gradu -tutkielma, 38+7 s.
Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
Kevät 2023

Tiivistelmä

Tämän pro gradu -tutkielman tavoitteena on selvittää, onko lapsiluvun vaikutus sepelvaltimotautiriskiin yleistettävissä yhdysvaltalaisesta populaatiosta jyväskyläläiseen keski-ikäisillä naisilla hyödyntäen kausaalivaikutuksen siirtoa. Tutkielmassa selvitetään, millaisin oletuksin kausaalivaikutuksen siirto on mahdollista, ja mitä tietoa kummastakin populaatiosta tarvitaan kausaalivaikutuksen estimoimiseksi kohdepopulaatiossa. Lopulta halutaan tietää, kuinka hyvin kausaalivaikutuksen estimoimiseksi valittu malli ennustaa sepelvaltimotautitapaukset kohdepopulaatiossa.

Tutkielmassa hyödynnetään lähdepopulaationa yhdysvaltalaista SWAN-tutkimuksen osa-aineistoa ja kohdepopulaationa jyväskyläläistä ERMA-tutkimuksen osa-aineistoa. Aineistojen tutkimushenkilöt ovat keski-ikäisiä naisia.

Tutkielmassa määritellään kausaalimalli lapsiluvun vaikutukselle sepelvaltimotautiriskiin. Mallin graafilla kuvattu kausaalirakenne on samanlainen kummallekin populaatiolle, ja sen avulla kuvataan ilmiöön liittyviä kausaalisuhteita. Kun erot graafin muuttujissa populaatioiden välillä ovat tiedossa, voidaan laskea siirtokaava, jolla kausaalivaikutus kohdepopulaatiossa estimoidaan. Siirtokaavasta nähdään, mitä muuttujia lähdepopulaatiosta tarvitaan mallin sovittamiseksi, ja mitä tietoa kohdepopulaatiosta tarvitaan sepelvaltimotautitapausten lukumäärän estimoimiseksi.

SWAN-aineistoon sovitetaan Coxin malli, jossa sepelvaltimotautia selitetään siirtokaavan mukaan lapsiluvulla, tupakoinnilla, etnisellä taustalla, koulutusasteella, siviilisäädyllä, iällä ja lisäksi painoindeksillä (body mass index, BMI), jos BMI oletetaan havaituksi. Mallilla ennustetaan sepelvaltimotautitapausten lukumäärä ERMA-aineistossa. Ennusteeseen tarvitaan siirtokaavan mukaan ERMA-aineistosta tupakointi, etninen tausta, koulutusaste, siviilisääty, ikä sekä lisäksi BMI, jos BMI oletetaan havaituksi. Mallin sovituksessa on huomioitu iän päivittyminen ajan kuluessa, mutta ennusteessa käytetään ERMA-aineiston seurannan alun ikää. Muut muuttujat ovat sekä mallinnuksessa että ennusteessa seurannan alun arvoja.

Mallin antaman ennusteen mukaan ERMA-aineistossa sepelvaltimotautitapauksia ilmenee 1.50–6.58 (95%:n luottamusvälit välillä [1.41, 6.87]) riippuen lapsiluvusta ja siitä oletetaanko BMI havaituksi vai ei. Todellinen tapausten lukumäärä ERMA-aineistossa on kaksi. Mallin mukaan lapsiluvun kasvaessa riski sairastua sepelvaltimotautiin kasvaa, mikä tukee aikaisempaa tutkimusta aiheesta.

Avainsanat: Kausaalivakutuksen siirto, kausaalivaikutuksen yleistäminen, kausaalimalli, lapsiluvun vaikutus sepelvaltimotautiin

Sisällys

Johdanto

Tämän pro gradu -tutkielman tavoitteena on selvittää, onko lapsiluvun vaikutus sepelvaltimotautiriskiin yleistettävissä yhdysvaltalaisesta populaatiosta jyväskyläläiseen keski-ikäisillä naisilla. Tutkimustuloksen yleistämistä tiettyyn kohdepopulaatioon kutsutaan kausaalivaikutuksen siirtämiseksi. Tutkielmassa selvitetään, millaisin oletuksin kausaalivaikutuksen siirto on mahdollista, ja mitä tietoa kummastakin populaatiosta tarvitaan kausaalivaikutuksen estimoimiseksi kohdepopulaatiossa. Lopulta halutaan tietää, kuinka hyvin kausaalivaikutuksen estimoimiseksi valittu malli ennustaa sepelvaltimotautitapaukset kohdepopulaatiossa. Tutkielmassa hyödynnetään lähdepopulaationa yhdysvaltalaista SWAN-tutkimuksen osa-aineistoa ja kohdepopulaationa jyväskyläläistä ERMA-tutkimuksen osa-aineistoa.

Kausaalivaikutuksen siirto on eräs keino yleistää tutkimuksessa saatu tulos toiseen populaatioon, joka voi ympäristöltään erota alkuperäisen tutkimuksen olosuhteista (Pearl ja Bareinboim (2014)). Pearl ja Bareinboim (2014) määrittelevät artikkelissaan täsmällisesti, millaisin oletuksin kausaalivaikutuksen siirtäminen on mahdollista, ja kuinka nk. siirtokaavan avulla voidaan selvittää, mitä tietoa kummastakin populaatiosta tarvitaan kausaalivaikutuksen estimoimiseksi kohdepopulaatiossa. Siirtokaavan selvittäminen vaatii kausaalilaskentaa, johon Tikka, Hyttinen ja Karvanen (2021a) tarjoavat artikkelissaan työkaluksi do-search -algoritmin. Do-search -algoritmia käytetään myös yleisesti kausaalivaikutuksen identifioimiseen.

Sepelvaltimotauti on länsimaalaisten yleisin kuolinsyy (Oliver-Williams et al. (2019), Psaltopoulou et al. (2017)). Tärkein sepelvaltimotaudin riskitekijä on ylipaino (Ho et al. (2022)). Muita keskeisiä riskitekijöitä ovat kohonnut veren kolesteroli, korkea verenpaine, diabetes, tupakointi ja ikä (Kettunen (2021)). Muita sydän- ja verisuonitauteihin yhteydessä olevia tekijöitä ovat sosioekonominen asema, etninen tausta ja siviilisääty (Psaltopoulou et al. (2017), Meadows et al. (2011), Wong et al. (2018)). Naisilla synnytys lisää sydän- ja verisuonitauteihin sairastumisen riskiä (Li et al. (2019), Oliver-Williams et al. (2019)). Lisäksi lapsiluvun kasvaessa ylipaino on yleisempää (Iversen, Kesmodel ja Ovesen (2018)). Lapsilukuun vaikuttaa olennaisesti hedelmällisyys, johon taas vaikuttavat mm. ikä ja päihteet (Tiitinen (2022)). Naisen sosioekonominen asema voi olla yhteydessä lapsilukuun (Tarkoma (2018)).

Tutkielman ensimmäisessä luvussa kuvaillaan lapsiluvun ja sepelvaltimotaudin yhteyttä kirjallisuuden pohjalta. Toisessa luvussa kuvaillaan tarkemmin tutkimusongelmaa ja tutkielmassa käytettävää aineistoa. Kolmas luku keskittyy kausaalivaikutuksen siirron teoriaan ja luvussa esitellään do-search -algoritmia hyödyntäen saatu siirtokaava. Neljännessä ja viidennessä luvussa määritetään ja sovitetaan tilastollinen malli, ja estimoidaan kausaalivaikutus kohdepopulaatiossa. Viimeisessä luvussa pohditaan mm. kausaalivaikutuksen siirron onnistumista ja tulosten luotettavuutta.

LUKU 1

Lapsiluvun yhteys sepelvaltimotaudin riskiin

Tässä luvussa kuvaillaan yleisesti sepelvaltimotautia, sen yleisyyttä ja riskitekijöitä sekä esitellään naissukupuolen ominaispiirteitä sepelvaltimotautiriskiin. Lisäksi kuvaillaan lapsilukuun vaikuttavia tekjöitä. Näiden tietojen pohjalta voidaan seuraavissa luvuissa kuvailla lapsiluvun ja sepelvalmitotaudin yhteyteen liittyviä aineiston muuttujia sekä määritellä ilmiöön liittyvä kausaalimalli.

1.1. Sepelvaltimotauti ja siihen vaikuttavat tekijät

Sepelvaltimotauti on länsimaalaisten yleisin kuolinsyy (Oliver-Williams et al. (2019), Psaltopoulou et al. (2017)). Se kuuluu sydän- ja verisuonitauteihin ja on myös Suomessa yksi merkittävimmistä kansansairauksista, johon naisia ja miehiä kuolee yhtä paljon (Kettunen (2021)).

Sydämen pinnalla sijaitsevat kaksi sepelvaltimoa - vasen ja oikea - vastaavat sydänlihaksen hapensaannista ja ravitsemuksesta. Valtimokovettumataudissa (ateroskleroosi) kolesterolia ja tulehdussoluja kertyy valtimoiden sisäpinnalle ahtauttaen valtimoita. Pidemmälle edenneessä taudissa myös sepelvaltimoiden seinämät ahtatuvat heikentäen sydänlihaksen hapensaantia. Tällöin puhutaan sepelvaltimotaudista. Heikentynyt sydämen hapensaanti aiheuttaa rintakipua etenkin rasituksen aikana. Sepelvaltimotaudin keskeisimpiä riskitekijöitä ovat kohonnut veren kolesteroli, korkea verenpaine, diabetes ja tupakointi. Myös ikä lisää sepelvaltimotautiin sairastumisen riskiä. Taudin ehkäisyssä ja hoidossa keskeistä on tupakoinnin lopettaminen, liikunnan lisääminen, terveellinen ruokavalio ja painonhallinta. Hoitona on aina myös lääkitys. (Kettunen (2021)).

Sydäninfarktin aiheuttaa äkillinen hapenpuutos sydänlihaksessa ja sen oireena on voimakas ja pitkittynyt rintakipu levossa. Sydäninfarktin taustalla on lähes aina sepelvaltimotauti. (Kettunen (2020)).

Matala sosioekonominen asema on yhteydessä kohonneeseen sydän- ja verisuonitautien riskiin. Matala sosioekonominen asema ennustaa sydän- ja verisuonitaudeille epäedullista ruokavaliota, mikä voi johtua matalasta koulutus- ja tulotasosta. Epäedulliseen ruokavalioon liittyy usein ylipainoa, korkeaa verenpainetta ja tyypin 2 diabetesta. Myös alkoholin kulutus on suurempaa matalan sosioekonomisen aseman omaavalla. Kohtuullisella alkoholin kulutuksella voi olla sydän- ja verisuonitauteja suojaava vaikutus korkeamman veren HDL-tason vuoksi. Suuri alkoholin kulutus sitä vastoin nostaa sairauksien riskiä. (Psaltopoulou et al. (2017)). Naisilla matalan sosioekonomisen aseman vaikutus sepelvaltimotaudin riskiin on voimakkaampi (Backholer et al. (2016)).

Etninen tausta voi olla yhteydessä sydän- ja verisuonitautien riskiin. Tummaihoisilla riski kuolla sydän- ja verisuonitauteihin on suurentunut, ja aasialaisilla se on

pienentynyt verrattuna muihin etnisiin ryhmiin maailmanlaajuisesti. Eteläaasialaisilla esiintyy eniten sepelvaltimotautia, mutta kuolleisuus sydäninfarktiin on pienin. (Meadows et al. (2011)). Tummaihoisilla useammin ilmenevä ylipaino, kohonnut verenpaine ja diabetes voivat selittää tauteihin sairastumisen riskiä. (CDC (2019)). Tupakointi lisää sairastumisen riskiä eniten tumma- ja valkoihoisilla. Eteläaasialaisilla riskiä voi lisätä korkea punasolujen hemoglobiini. (Ho et al. (2022)).

Siviilisääty voi olla yhteydessä sepelvaltimotaudin ilmenemiseen ja kuolleisuuteen sydänkohtauksen jälkeen sekä miehillä että naisilla. Yleisesti naimisissa olevilla on pienempi riski sairastua tai kuolla sydän- tai verisuonitauteihin verrattuna muuhun väestöön. Lisäksi naimisissa olevilla kuoleman riski sydänkohtauksen jälkeen on pienempi. Naimisissa olemisen suojaavan vaikutuksen arvellaan liittyvän puolison läsnäoloon, ja sitä myötä esimerkiksi varhaisempaan hoitoon hakeutumiseen. Lisäksi puolisolla saattaa olla terveisiin elintapoihin ja sairastumisen jälkeen hoitoon sitoutumiseen kannustava vaikutus. (Wong et al. (2018)).

1.2. Sepelvaltimotauti naisilla

Tärkein sepelvaltimotaudin riskitekijä on ylipaino (Ho et al. (2022)). Ylipainon rajana pidetään painoindeksiä (body mass index, BMI) 25 ja lihavuuden 30 (WHO (2021)). Euroopassa ylipainoisia tai lihavia on 36.8 prosenttia kaikista lisääntymisikäisistä naisista. Vuonna 2014 noin 50 prosenttia amerikkalaisista elävän lapsen synnyttäneistä oli ylipainoisia tai lihavia. (Iversen, Kesmodel ja Ovesen (2018)). Suomessa ylipainoisia oli 41.9 ja lihavia 17.0 prosenttia kaikista synnyttäjistä ennen raskautta vuonna 2019 (Kiuru, Gissler ja Heino (2019)).

Raskaus voi johtaa ylipainon tai lihavuuden kehittymiseen, jos paino ei raskauden jälkeen palaa raskautta edeltävälle tasolle. Ylipaino tai lihavuus ennen raskautta ovat yhteydessä painonnousuun raskauksien myötä. Myös liiallinen painonnousu raskausaikana voi johtaa painonnousuun pidemmällä aikavälillä. Lapsiluvun kasvaessa ylipaino on yleisempää. (Iversen, Kesmodel ja Ovesen (2018)). Imetyksellä on vain lievä vaikutus painonkehitykseen, kun taas tupakoinnin lopettamisella raskausaikana vaikutus on merkittävä. Painonnousua on enemmän naisilla, jotka muuttavat syömistottumuksiaan, ateriarytmiään ja fyysistä aktiivisuuttaan raskauksien myötä. (Rössner ja Öhlin (1995)). Lihavuus aiheuttaa raskaudenaikaisia riskejä kuten raskausdiabetesta tai pre-eklampsiaa (Kiuru, Gissler ja Heino (2019)).

Synnytys lisää sydän- ja verisuonitauteihin sairastumisen riskiä (Li et al. (2019), Oliver-Williams et al. (2019)). Li et al. (2019) mukaan lapsiluvulla ja sydän- ja verisuonitaudeilla on merkitsevä yhteys: Synnytysten lukumäärän kasvaessa, lisääntyy sairastumisen riski epälineaarisesti noudattaen J-muotoa, jolloin yhden synnytyksen jälkeen riski on matalin. Taustalla vaikuttavia tekijöitä voivat olla sukupuolihormonitasojen vaihtelut sekä raskauksien ja synnytysten aiheuttamat verenkiertoelimistön muutokset, joilla voi olla pitkäaikaisvaikutuksia. (Li et al. (2019)). Lisäksi raskaudessa paino nousee, vatsanalueen rasva lisääntyy, veren lipiditasot ovat korkeammat ja insuliiniresistenssi kasvaa väliaikaisesti, joilla voi olla pitkäaikaisvaikutuksia naisen sydän- ja verenkiertoelimistön terveyteen. Synnytysten lukumäärän kasvaessa imetyksen sydän- ja verisuonitaudeilta suojaava vaikutus häviää. (Oliver-Williams et al. (2019)). Raskauskomplikaatiot, kuten pre-eklampsia, lisäävät sepelvaltimotautiin sairastumisen riskiä (Haukkamaa et al. (2020)).

Myös synnyttämättömyys voi lisätä sydän- ja verisuonitauteihin sairastumisen riskiä varsinkin, kun se yhdistyy matalaan koulutustasoon. Mahdollisia vaikuttavia tekijöitä ovat polykystiset munasarjat raskauksien puuttuessa sekä keskenmenot tai kohtukuolemat. (Yasukawa et al. (2022)). Aikaisemmat raskaudet ilman synnytyksiä lisäävät sairastumisen riskiä (Oliver-Williams et al. (2019)).

Estrogeenihormoni suojaa sydän- ja verisuonitaudeilta. Siten vaihdevuosien alkaminen naisilla ja sitä myötä estrogeenitasojen väheneminen lisää sairastumisen riskiä etenkin naisilla, joilla on muitakin riskitekijöitä. (Haapalahti ja Mikkola (2015)).

1.3. Lapsilukuun vaikuttavat tekijät

Lapsilukuun olennaisesti vaikuttava tekijä on hedelmällisyys. Naisella ikä on yksi merkittävimmistä hedelmällisyyteen vaikuttavista tekijöistä, sillä lapsettomuus lisääntyy iän myötä etenkin 35 ikävuoden jälkeen. Myös sukupuoliteitse tarttuvien tautien ehkäisy ja hoito on olennaista lapsettomuuden ehkäisyssä. Muita hedelmällisyyteen vaikuttavia tekijöitä ovat elintavat. Elintavoista hedelmällisyyteen vaikuttavat merkittävästi tupakointi ja alkoholi. Lisäksi vaikuttavia tekijöitä ovat ravinto ja liikunta. Myös ali- tai ylipaino vaikuttavat hedelmällisyyteen sitä laskevasti. (Tiitinen (2022)).

Lapsilukuun voi vaikuttaa myös naisen sosioekonominen asema. Suomessa kokonaishedelmällisyysluku on sitä suurempi, mitä korkeampi naisen koulutusaste on. Ennen vuotta 2012 lapsettomien 35-39-vuotiaiden korkeakoulutettujen naisten osuus on ollut suurempi kuin matalamman koulutuksen omaavien. Nykyään lapsettomien osuus on matalin korkekoulutetuilla. (Tarkoma (2018)). Toisaalta matalampi sosioekonominen asema voi olla yhteydessä suurempaan lapsilukuun (Oliver-Williams et al. (2019)).

LUKU 2

Tutkimusongelman ja aineiston kuvaus

Tämän pro gradu -tutkielman tavoitteena on selvittää, onko lapsiluvun vaikutus sepelvaltimotautiriskiin siirrettävissä yhdysvaltalaisesta populaatiosta jyväskyläläiseen keski-ikäisillä naisilla. Lisäksi selvitetään, millaisin oletuksin kausaalivaikutuksen siirto on mahdollista, ja mitä tietoa kummastakin populaatiosta tarvitaan kausaalivaikutuksen estimoimiseksi kohdepopulaatiossa.

Käytännössä kausaalivaikutusta siirrettäessä populaatiosta toiseen määritellään kausaalimalli, jonka graafilla kuvattu kausaalirakenne on identtinen lähde- ja kohdepopulaatiolle. Sen jälkeen kausaalimallin graafiin merkitään oletukset liittyen populaatioiden välisiin eroavaisuuksiin kausaalimallin muuttujissa. Näin saadun valikoitumisgraafin pohjalta voidaan do-search -algoritmia hyödyntäen laskea kaava, jolla estimoidaan haluttu kausaalivaikutus kohdepopulaatiossa. Kausaalivaikutuksen estimoimiseksi valitaan tilastollinen malli, johon muuttujat valitaan siirtokaavan perusteella. Mallin avulla pyritään ennustamaan tautitapaukset kohdepopulaatiossa. Kun saatavilla on tieto myös kohdepopulaation tautitapauksista, voidaan selvittää, kuinka hyvin kausaalivaikutuksen estimoimiseksi valittu malli ennustaa tautitapausten lukumäärän. Kausaalimallille esitetään kaksi vaihtoehtoista ratkaisua. Nämä ratkaisut ovat muutoin identtiset, mutta toisessa BMI oletetaan latentiksi eli havaitsemattomaksi muuttujaksi. Näin tehdään, jotta ennuste saadaan kohdepopulaatiossa mahdollisimman monelle havainnolle, sillä painoindeksissä oli melko paljon puuttuvuutta. Ennusteet raportoidaan kummallakin ratkaisulla.

Tutkielmassa hyödynnetään aineistoja kahdesta populaatiosta: yhdysvaltalaista SWAN-tutkimukseen liittyvää aineistoa sekä jyväskyläläistä ERMA- ja EsmiRs-tutkimuksiin liittyvää aineistoa. ERMA- ja EsmiRs-tutkimuksen osa-aineisto on saatu käyttöön Jyväskylän yliopiston liikuntatieteellisen tiedekunnan luvalla tätä tutkielmaa varten. SWAN-tutkimuksen aineistot tutkimuskäynneiltä 1–10, alkukartoitusaineisto sekä poikkileikkausseulonta-aineisto ovat netistä vapaasti ladattavissa osoitteessa `https://www.icpsr.umich.edu/web/ICPSR/series/00253`. Tässä luvussa kuvaillaan aineistoja tutkielman tavoitteen kannalta olennaisin osin.

2.1. SWAN- ja ERMA-aineisto

SWAN-tutkimus (Study of Women's Health Across the Nation) tutkii keski-ikäisten yhdysvaltalaisten naisten terveyttä. Tutkimuksen tavoitteena on selvittää vaihdevuosien vaikutusta naisten terveyteen ja elämänlaatuun. Tutkimuksessa on selvitetty laajasti naisten fyysisiä ja psyykkisiä muutoksia vaihdevuosien kynnyksellä erilaisin mittauksin ja kyselyin. Aineiston keruu on alkanut vuonna 1994 keskitettynä seitsemään yliopistojen yhteydessä olevaan tutkimuskeskukseen Yhdysvalloissa. (*Study of Women's Health Across the Nation (SWAN) Series* (2019)).

Tutkielmassa hyödynnetään SWAN-tutkimuksen poikkileikkausseulonta-aineistoa (Cross-Sectional Screener Dataset), joka on kerätty vuosina 1995–1997. Tätä tutkielmaa varten aineistosta on poistettu sepelvaltimotautia sairastavat tutkimushenkilöt, sillä lähtötilanteeseen on haluttu vain terveet. Tieto taudin puhkeamisesta on kerätty noin vuoden välein tapahtuvilta käynneilta 1–10. Osalta tutkimushenkilöistä löytyy tiedot kaikilta käynneiltä, osa on jättänyt tutkimuksen kesken, ja joltain saattaa puuttua jokin käynti välistä. Henkilöt, joilta seuranta puuttuu kokonaan, on jätetty pois aineistosta. Siten saadun aineiston 2988 naista ovat 41–52 vuoden ikäisiä mediaanin ollessa 46.00 ja keskihajonnan 2.69.

Etniseltä taustaltaan SWAN-aineiston naiset jakautuvat viiteen eri luokkaan: valkoihoinen – afrikkalaisamerikkalainen – kiinalainen/kiinalaisamerikkalainen – japanilainen/japanilaisamerikkalainen – latinalaisamerikkalainen (engl. Caucasian/White Non-Hispanic – Black/African American – Chinese/Chinese American – Japanese/Japanese American – Hispanic). Valtaosa on valkoihoisia (47.5%) tai afrikkalaisamerikkalaisia (28.2%). 9.2% on japanilaisia/japanilaisamerikkalaisia, 8.0% kiinalaisia/kiinalaisamerikkalaisia ja 7.1% latinalaisamerikkalaisia. Tutkimushenkilöt ovat itse määrittäneet etnisen taustansa. SWAN-aineistoa kuvataan tässä luvussa yleisen tason lisäksi erikseen valkoihoisten osalta, sillä he ovat etniseltä taustaltaan lähinnä ERMA-aineiston naisia, jotka ovat kaikki valkoihoisia.

ERMA-tutkimus (Estrogenic Regulation of Muscle Apoptosis) on Jyväskylän yliopiston liikuntatieteellisen tiedekunnan Gerontologian tutkimuskeskuksessa toteutettava tutkimushanke. Tutkimuksen tavoitteena on tutkia naisten lihasten ikääntymismuutoksia ja vaihdevuosiin liittyviä hormonimuutoksia sekä niiden vaikutuksia naisen fyysiseen ja psyykkiseen toimintakykyyn. Aineiston keruu toteutettiin vuosina 2014–2018 ja kohderyhmänä ovat olleet jyväskyläläiset tai Jyväskylän lähikunnissa asuvat vaihdevuosi-ikää lähestyvät 47–55-vuotiaat naiset, joista tutkimushenkilöt valittiin satunnaisotannalla väestörekisterikeskuksen asukasrekisteristä. Tutkimuksen vaiheissa 1 ja 2 kerättiin poikkileikkausaineisto. Vaihe 1 koostui esitietokyselystä ja vaihe 2 laboratoriotutkimuksista. Laboratoriotutkimusten ja vaiheen 3 laajan terveyskyselytutkimuksen perusteella tutkittavat jaettiin vaihdevuosivaiheiden mukaisiin ryhmiin. Vaihdevuosiin siirtymisvaiheessa olevat naiset jatkoivat vaiheen 4 seurantatutkimukseen, joka kesti vuoden 2018 loppuun. (ERMA (2019)).

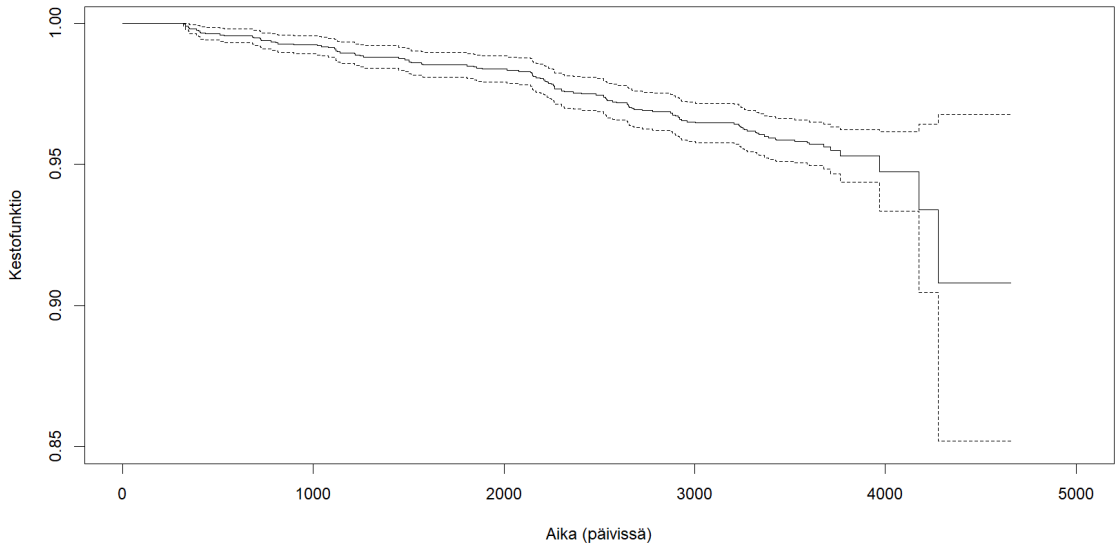
Osa ERMA-tutkimuksen vaiheen 2 naisista osallistui myös ERMA-tutkimusaineistoon pohjautuvaan EsmiRs-tutkimukseen (Estrogeeni, mikro-RNA:t ja metabolisten toimintahäiriöiden riski). EsmiRs-tutkimus selvittää menopaussin negatiivisten terveysvaikutusten taustalla vaikuttavia biologisia mekanismeja. Tutkimuksen aineiston keruu oli kolmivaiheinen: kyselytutkimus, ERMA-tutkimuksen alkumittaukset toistava tutkimuskäynti (ERMA vaihe 3) ja uudet aineenvaihduntamittaukset. (EsmiRs (2021)).

Tässä tutkielmassa hyödynnetään ERMA-tutkimuksen vaiheen 3 laajaa terveystutkimus- ja -kyselyaineistoa, josta on poistettu sepelvaltimotautia sairastavat kuten SWAN-aineistosta. Tieto mahdollisesta tautiin sairastumisesta on saatu ERMA-tutkimuksen loppumittauksista (vaihe 4), sekä EsmiRs-tutkimuksen kyselyaineistosta. Osalta löytyy tiedot vain ERMA-tutkimuksen loppumittauksista tai

EsmiRs-kyselystä. Henkilöt, joilta seurantatietoja ei ole ollenkaan, on jätetty pois aineistosta. Näin saadun seuranta-aineiston 589 naista ovat 47–55-vuotiaita mediaanin ollessa 51.29 ja keskihajonnan 2.00.

2.2. Sepelvaltimotauti aineistoissa

SWAN-tutkimuksessa sepelvaltimotautia on osallistujilta selvitetty taudin oireiden perusteella. Seurannan aikana 119 (4.0%) on vastannut myöntävästi kysymykseen "Onko lääkäri, sairaanhoitaja tai muu terveysalan ammattilainen viime käynnin jälkeen todennut sinulla olevan jokin seuraavista sairauksista tai oletko saanut hoitoa johonkin seuraavista: sydänkohtaus tai rasitusrintakipua?" Valkoihoisista sepelvaltimotauti puhkesi seurannan aikana 40 tutkimushenkilölle (2.8%). Kaplanin-Meierin kestofunktioestimaatin kuvaajasta (Kuva 2.1) nähdään todennäköisyys välttää tautiin sairastuminen ajan (päivissä) kuluessa.



Kuva 2.1. SWAN-aineiston Kaplanin–Meierin estimaatti sepelvaltimotautiin sairastumisesta sekä 95 prosentin luottamusvälit

ERMA- ja EsmiRs-tutkimusten kyselyissä on selvitetty osallistujien sepelvaltimotautia kysymyksellä "Onko Teillä jokin seuraavista sairauksista / häiriöistä, jonka lääkäri on todennut?" johon vastaaja valitsee sairauden kohdalla kyllä tai ei. Sepelvaltimotauti ja sydäninfarkti ovat erillisia vastausvaihtoehtoja, ja tähän tutkielmaan vastaukset on yhdistetty tarkoittamaan sepelvaltimotautia. Sepelvaltimotautitapauksia ilmenee seurannan aikana kaksi (0.3%).

2.3. Lapsiluku aineistoissa

Jälkeläisten lukumäärää on SWAN-tutkimuksessa selvitetty kysymyksellä "Kuinka monta elävää lasta olet synnyttänyt?" ERMA-tutkimuksen terveyskyselyssä vastaajien lapsiluku on taas selvitetty kysymyksellä "Montako synnytystä Teillä on ollut?",

johon vastaaja on voinut vapaasti antaa lukumäärän. Lapsiluku on jaettu SWAN-aineistossa kuuteen luokkaan: Ei lapsia – 1 lapsi – 2 lasta – 3 lasta – 4 lasta – Yli 4 lasta, jonka mukaan myös ERMA-aineiston lapsiluku on muokattu tätä tutkielmaa varten.

Yleisin lapsiluku on kummassakin aineistossa kaksi tai kolme (Taulukko 1). SWAN-aineistossa 2- tai 3-lapsisia onkin yhteensä yli 50% ja ERMA-aineistossa yli 60%. Lapsettomia SWAN-aineistossa on 17.1%, kun taas ERMA-aineistossa 13.1%. Yli 4-lapsisia sen sijaan on SWAN-aineistossa 4.3% ja ERMA-aineistossa 2.5%. Kummassakin aineistossa lapsiluku puuttuu alle yhdeltä prosentilta vastaajista.

Muuttuja	Ryhmä	SWAN lukumäärä (%)	SWAN: Valkoihoinen lukumäärä (%)	ERMA lukumäärä (%)	Sepelvaltimotauti aineistossa SWAN: lukumäärä (%)
Lapsiluku	Ei lapsia	512 (17.1)	342 (24.1)	77 (13.1)	11 (2.1)
	1 lapsi	498 (16.7)	242 (17.1)	82 (13.9)	16 (3.2)
	2 lasta	1011 (33.8)	452 (31.9)	245 (41.6)	34 (3.4)
	3 lasta	572 (19.1)	239 (16.8)	133 (22.6)	23 (4.0)
	4 lasta	263 (8.8)	102 (7.2)	36 (6.1)	23 (8.7)
	Yli 4 lasta	128 (4.3)	41 (2.9)	15 (2.5)	12 (9.4)
	Tieto puuttuu	4 (0.1)	1 (0.1)	1 (0.2)	

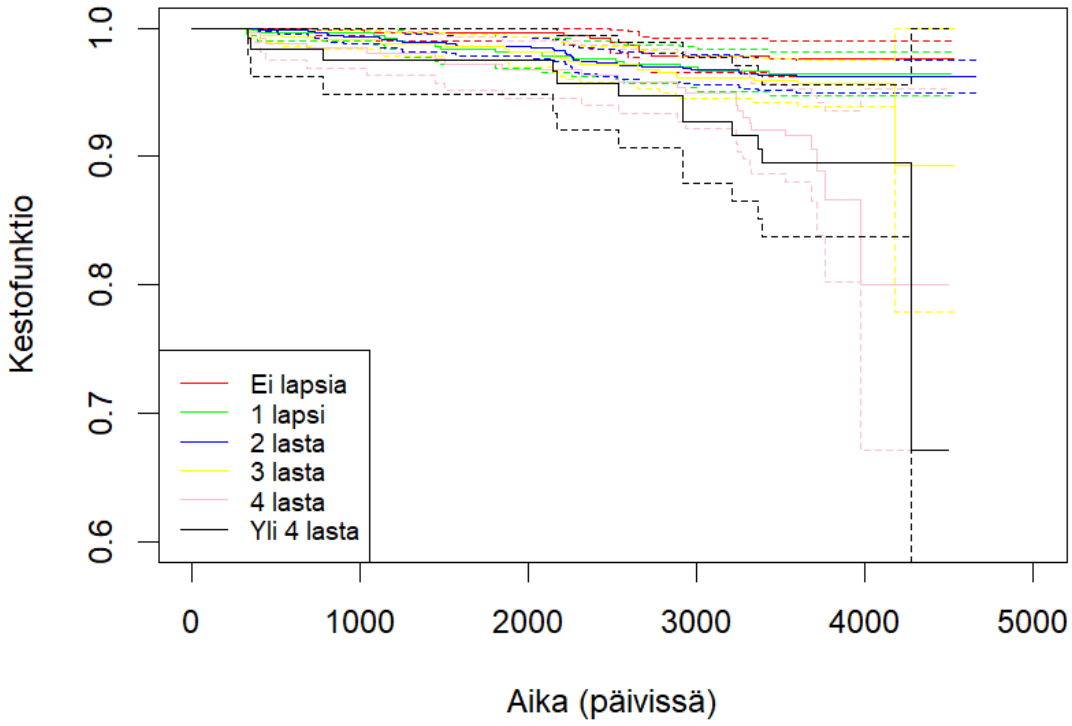
TAULUKKO 1. Lapsiluku aineistoissa ja sepelvaltimotauti lapsiluvuittain SWAN-aineistossa

SWAN-aineistossa sepelvaltimotaudin ilmeneminen vaihtelee jonkin verran lapsiluvun mukaan. Alle 3-lapsisilla sepelvaltimotautia esiintyy vähemmän, sillä esimerkiksi lapsettomilla tauti ilmenee 2.1 prosentille. Kun lapsia on yli kolme, sepelvaltimotauti vaikuttaa olevan yleisempää, sillä 4-lapsisilla ja yli 4-lapsisilla sepelvaltimotauti ilmenee yli kahdeksalle prosentille. Myös Kaplanin-Meierin kestofunktioestimaatin kuvaaja (Kuva 2.2) viittaa siihen, että yli 3-lapsisilla tautiin sairastuminen on todennäköisempää, vaikka 95 prosentin luottamusväleissä on päällekkäisyyttä.

2.4. Koulutusaste aineistoissa

Koulutusaste on SWAN-aineistossa jaettu viiteen eri luokkaan ja ERMA-aineistossa kahdeksaan. Tätä tutkielmaa varten kummankin aineiston koulutusasteita on yhdistelty neljään luokkaan siten, että aineistojen vertailu on mahdollista. Neljä luokkaa ovat peruskoulu (peruskoulu, engl. less than high school) – toinen aste (lukio/ammattikoulu, engl. high school graduate/some college/technical school) – korkeakoulututkinto (opistoasteen ammatillinen koulutus/ammattikorkeakoulu/ alempi korkeakoulututkinto, engl. college graduate) – ylempi korkeakoulututkinto (ylempi korkeakoulututkinto/lisensiaatti- tai tohtorintutkinto, engl. post graduate education).

SWAN-aineistossa suurin osa on toisen asteen suorittaneita (49.2%), kun taas ERMA-aineistossa korkeakoulututkinnon suorittaneita (48.0%) (Taulukko 2). Vain



Kuva 2.2. SWAN-aineiston Kaplanin–Meierin estimaatti sepelvaltimotautiin sairastumisesta 95% luottamusvälillä lapsiluvuittain

peruskoulun käyneitä on SWAN-aineistossa 6.4% ja ERMA-aineistossa 2.0%. Ylemmän korkeakoulututkinnon suorittaneita on SWAN-aineistossa 23.3% ja ERMA-aineistossa 30.1%. SWAN-aineistossa koulutustausta puuttuu 0.5 prosentilta ja ERMA-aineiston koulutustaustassa ei ole puuttuvuutta.

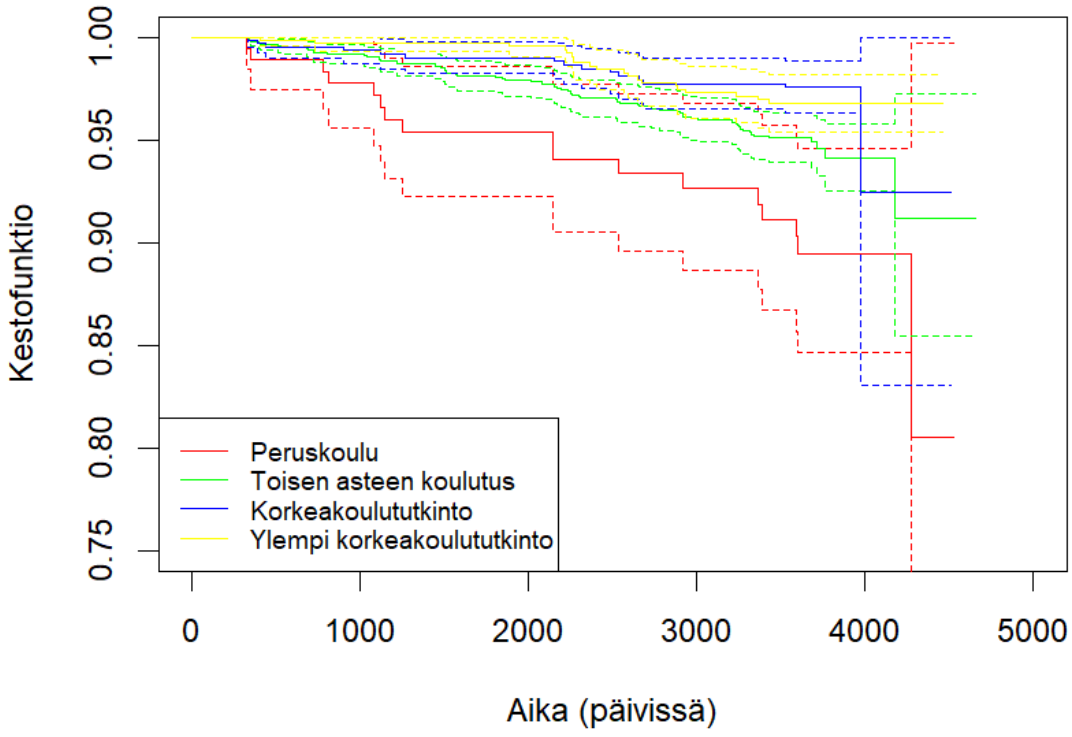
SWAN-aineistossa koulutusasteiden osuudet eroavat hiukan valkoihoisilla yleiseen tasoon verrattuna (Taulukko 2). Myös valkoihoisissa suurin osa on ylioppilaita tai ammatillisen tutkinnon suorittaneita (44.4%). Korkeammin koulutettuja on kuitenkin enemmän valkoihoisissa kuin yleisellä tasolla, kun taas vain peruskoulun käyneitä vähemmän: vain peruskoulun käyneitä on 1.4%, korkeakoulututkinto on 22.2%:lla ja ylempi korkeakoulututkinto 31.7%:lla.

SWAN-aineistossa sepelvaltimotaudin ilmeneminen vaihtelee hiukan koulutusasteen mukaan (Taulukko 2). Sepelvaltimotauti on yleisintä vain peruskoulun käyneillä (8.9%). Ylioppilailla ja ammatillisesti koulutetuilla esiintyvyys on 4.6%, korkeakoulututkinnon suorittaneilla 2.4% ja ylemmän korkeakoulututkinnon suorittaneilla 2.9%. Myös Kaplanin-Meierin kestofunktioestimaatin kuvaaja viittaa eroihin taudin ilmaantuvuudessa (Kuva 2.3).

Lapsiluku vaihtelee aineistoissa koulutustaustan mukaan (Kuva 2.4 ja 2.5). Lapsettomuus lisääntyy koulutusasteen kasvaessa: peruskoulun käyneissä lapsettomia

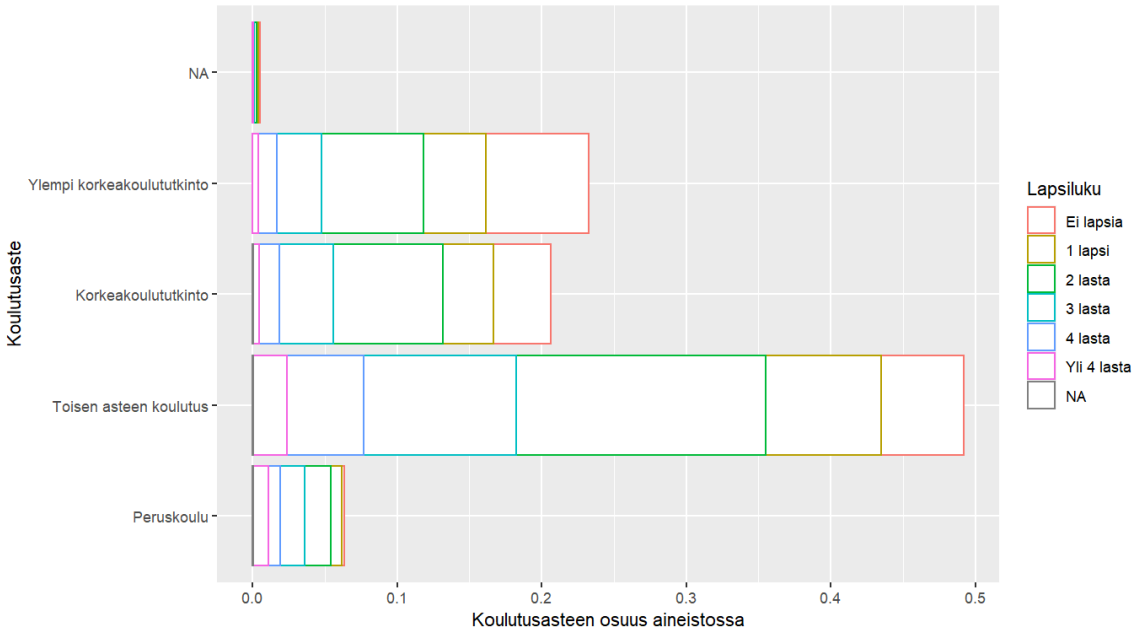
Muuttuja	Ryhmä	SWAN lukumäärä (%)	SWAN: Valkoihoinen lukumäärä (%)	ERMA lukumäärä (%)	Sepelvaltimotauti aineistossa SWAN: lukumäärä (%)
Koulutusaste	Peruskoulu	190 (6.4)	20 (1.4)	12 (2.0)	17 (8.9)
	Toinen aste	1470 (49.2)	630 (44.4)	117 (19.9)	67 (4.6)
	Korkeakoulututkinto	617 (20.6)	315 (22.2)	283 (48.0)	15 (2.4)
	Ylempi korkeakoulututkinto	695 (23.3)	450 (31.7)	177 (30.1)	20 (2.9)
	Tieto puuttuu	16 (0.5)	4 (0.3)	0 (0.0)	

TAULUKKO 2. Koulutusaste aineistoissa ja sepelvaltimotauti koulutusasteittain SWAN-aineistossa



KUVA 2.3. SWAN-aineiston Kaplanin–Meierin estimaatti sepelvaltimotautiin sairastumisesta 95% luottamusvälillä koulutusasteittain

on SWAN-aineistossa 3.2% ja ERMA-aineistossa 8.3%, kun taas ylemmän korkeakoulututkinnon suorittaneissa lapsettomia on SWAN-aineistossa 30.6% ja ERMA-aineistossa 14.1%. Yli 4-lapsisten osuuden trendi on päinvastainen SWAN-aineisossa: peruskoulun käyneissä yli 4-lapsisia on 16.3%, kun taas ylemmän korkeakoulututkinnon suorittaneissa yli 4-lapsisia on 1.7%. ERMA-aineistossa yli 4-lapsisia on vähän, joten osuudet eivät ole vertailukelpoisia.

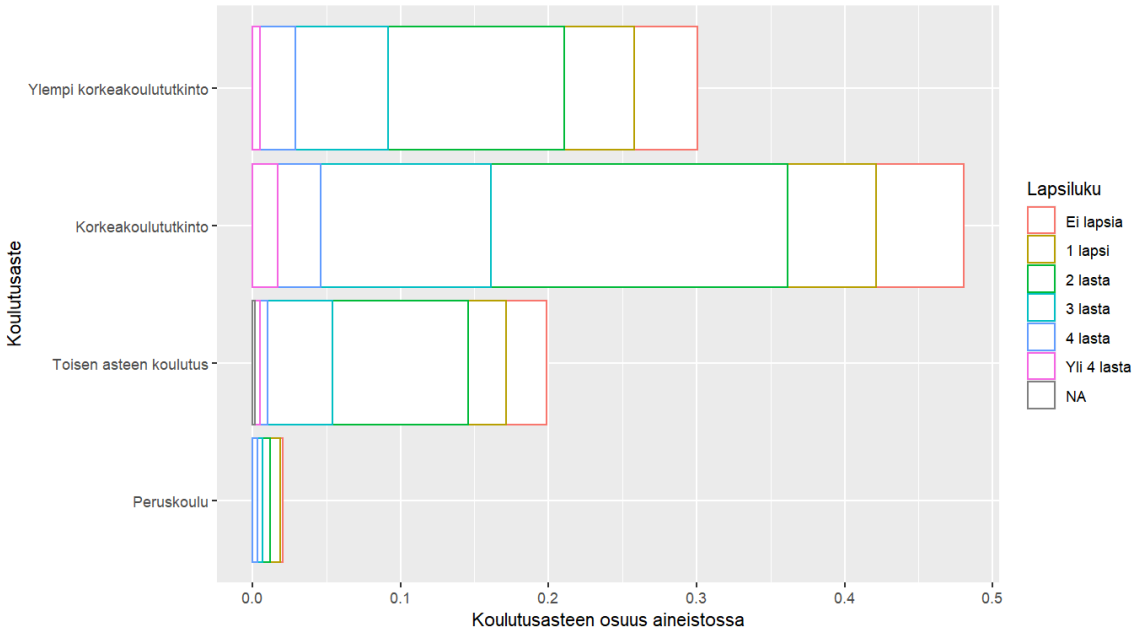


Kuva 2.4. Lapsiluku koulutusasteen mukaan SWAN-aineistossa, NA = tuntematon

2.5. BMI aineistoissa

Aineistoissa BMI on laskettu kaavalla $paino(kg)/pituus(m)^2$. ERMA-tutkimuksessa paino ja pituus on mitattu tutkimuskäynneillä, ja lisäksi BMI on mitattu InBody-kehonkoostumusmittarilla. Tässä tutkielmassa hyödynnetään tutkijoiden mittaamia arvoja, joita on tarvittaessa täydennetty InBody-mittauksilla. Tämän tutkielman analyyseissä hyödynnetään jatkuvia BMI-arvoja, mutta tätä kappaletta varten aineistojen naiset on jaettu painoindeksin mukaan neljään ryhmään: alipainoisiin (BMI <18.5), normaalipainoisiin (18.5 $\leq$ BMI <25), ylipainoisiin (25 $\leq$ BMI <30) ja merkittävästi ylipainoisiin (BMI $\geq$30). Aineistoissa normaalipainoisten osuudet ovat suurinpiirtein samat, noin 44–47% (Taulukko 3). Alipainoisten osuus on SWAN-aineistossa 1.7% kun taas ERMA-aineistossa se on 0.5%. Ylipainoisten tai merkittävästi ylipainoisten osuudet aineistoissa ovat yhteensä noin 43–50%. SWAN-aineistossa vastaajista 117:lla (3.9%) ei ole tietoa BMI:stä, kun taas ERMA-aineistossa tieto BMI:stä puuttuu 53 tutkimushenkilöltä (9.0%).

SWAN-aineistossa sepelvaltimotautin ilmeneminen eroaa etenkin verrataessa merkittävästi ylipainoisia muihin ryhmiin (Taulukko 3). Merkittävästi ylipainoisilla tauti ilmenee 8.2 prosentille, kun muissa ryhmissä se ilmenee alle neljälle prosentille. Myös Kaplanin-Meierin kestofunktioestimaatin kuvaajasta (Kuva 2.6) voidaan havaita, että



KUVA 2.5. Lapsiluku koulutusasteen mukaan ERMA-aineistossa, NA = tuntematon

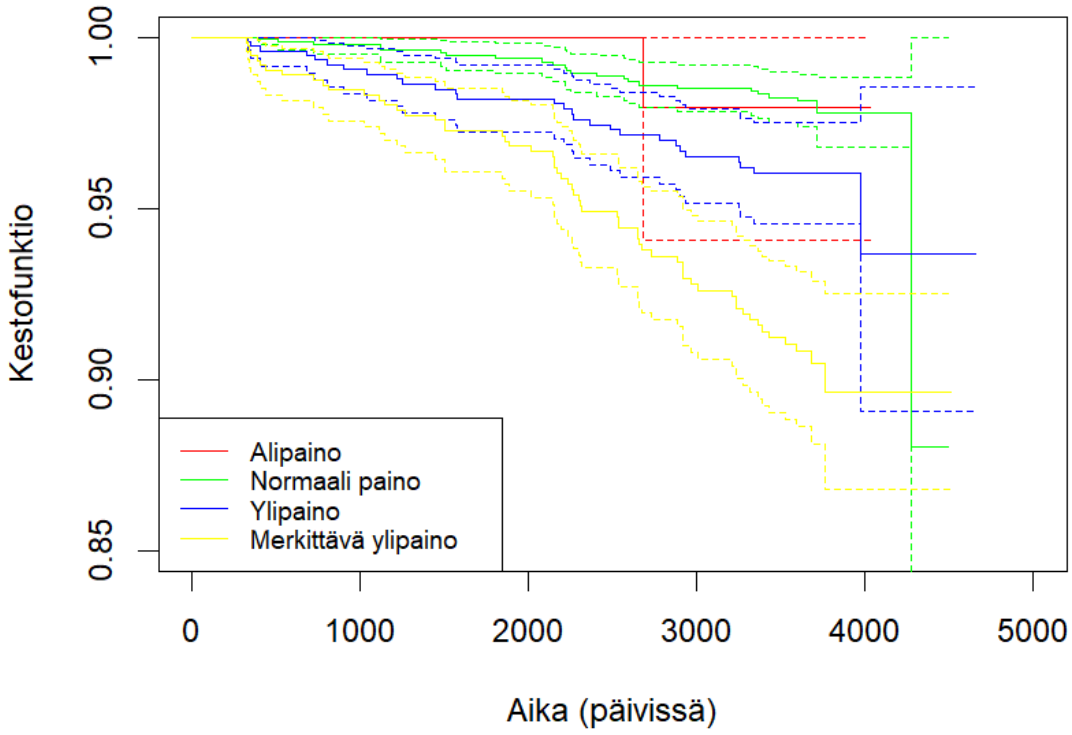
Muuttuja	Ryhmä	SWAN lukumäärä (%)	SWAN: Valkoihoinen lukumäärä (%)	ERMA lukumäärä (%)	Sepelvaltimotauti aineistossa SWAN: lukumäärä (%)
BMI	Alipaino	51 (1.7)	26 (1.8)	3 (0.5)	1 (2.0)
	Normaali paino	1328 (44.4)	673 (47.4)	279 (47.4)	24 (1.8)
	Ylipaino	761 (25.5)	366 (25.8)	180 (30.6)	28 (3.7)
	Merkittävä ylipaino	731 (24.5)	320 (22.6)	74 (12.6)	60 (8.2)
	Tieto puuttuu	117 (3.9)	34 (2.4)	53 (9.0)	6 (5.1)

TAULUKKO 3. BMI aineistoissa ja sepelvaltimotauti BMI-ryhmittäin SWAN-aineistossa

merkittävästi ylipainoisten todennäköisyys välttää tautiin sairastuminen ajan kuluessa on pienempi muihin ryhmiin verrattuna.

SWAN-aineistossa lapsiluku vaihtelee jonkin verran BMI-ryhmien mukaan (Kuva 2.7). Lapsettomuus on yleisintä alipainoisten ryhmässä (23.5%). Muissa ryhmissä lapsettomuus vaihtelee 14–20% välillä. 4-lapsisten osuus ja yli 4-lapsisten osuus on isompi ylipainoisten ja merkittävästi ylipainoisten ryhmässä kuin normaalipainoisissa.

ERMA-aineistossa lapsiluvuissa ei ole kovinkaan paljon eroja eri BMI-ryhmien välillä. Alipainoisten ryhmä on pieni, joten lapsilukujen osuudet eivät välttämättä ole kovin vertailukelpoisia. Myös 4-lapsisia tai yli 4-lapsisia on melko vähän, joten



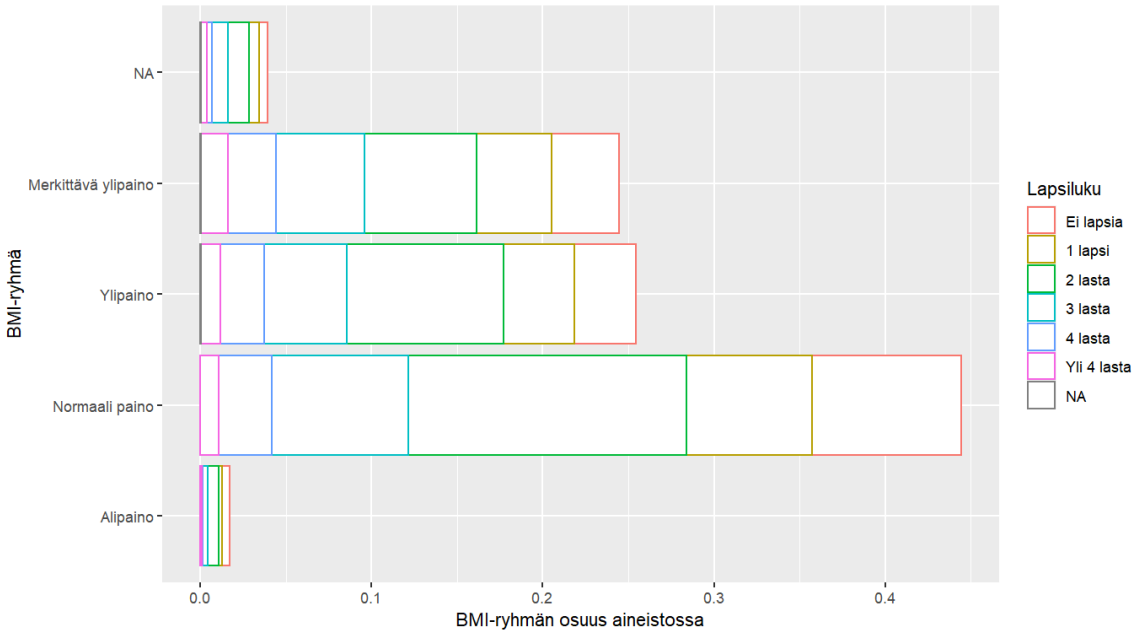
KUVA 2.6. SWAN-aineiston Kaplanin–Meierin estimaatti sepelvaltimotautiin sairastumisesta 95% luottamusvälillä BMI:n mukaan

on vaikeaa sanoa, onko heitä enemmän ylipainoisten ja merkittävästi ylipainoisten ryhmässä kuin normaalipainoisissa.

2.6. Tupakointi aineistoissa

Tupakointia on SWAN- ja ERMA-tutkimuksissa selvitetty usealla eri kysymyksellä, joiden perusteella tutkittavat on jaettu kolmeen ryhmään: ei koskaan – tupakoinut menneisyydessä – tupakoi nykyään. Ryhmistä on muodostettu 3-luokkainen muuttuja, jota voidaan hyödyntää tässä tutkielmassa. Luokkien osuudet ovat SWAN-aineistossa 58.7%, 24.8% ja 15.8%, kun taas ERMA-aineistossa ne ovat 66.7%, 27.2% ja 5.6% (Taulukko 4). Tieto tupakoinnista puuttuu alle prosentissa tapauksista kummassakin aineistossa.

SWAN-aineistossa sepelvaltimotaudin ilmeneminen vaikuttaisi olevan yleisempää nykyään tupakoivien ryhmässä verrattuna muihin ryhmiin. Myös Kaplanin-Meierin kestofunktioestimaatin kuvaajan (Kuva 2.8) mukaan nykyään tupakoivien todennäköisyys välttää tautiin sairastuminen voisi olla pienempi kuin muiden ryhmien. Lapsiluvut eivät aineistoissa vaikuta vaihtelevan eri tupakointiluokkien välillä.



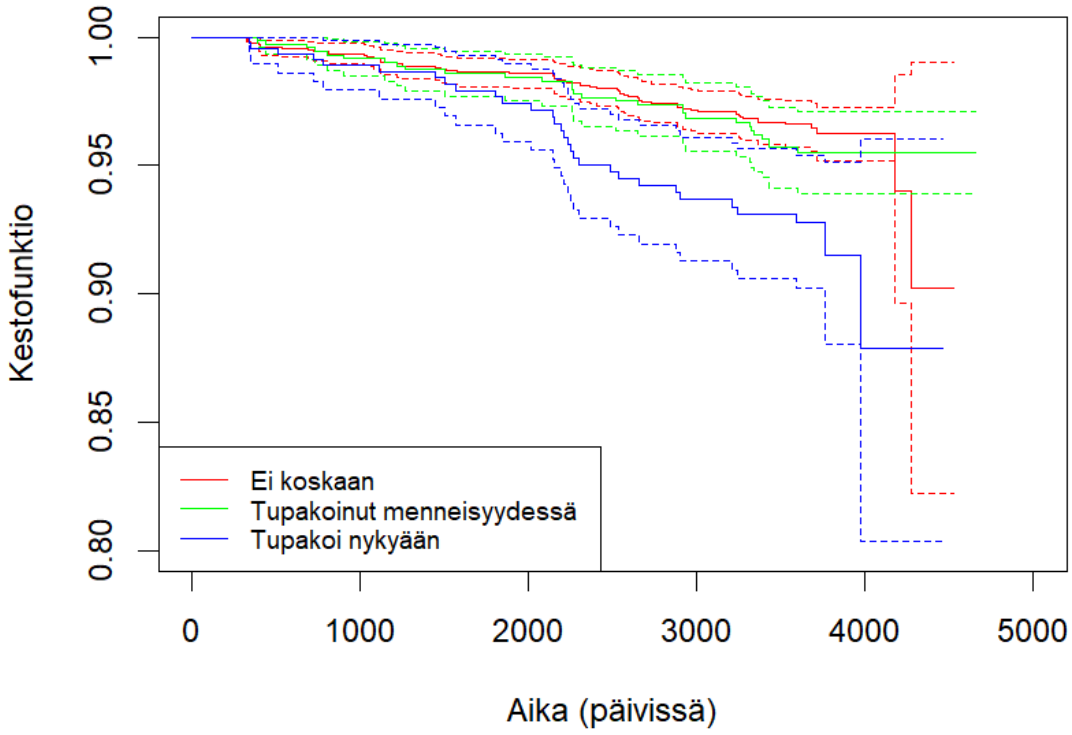
KUVA 2.7. Lapsiluku BMI-ryhmän mukaan SWAN-aineistossa, NA = tuntematon

Muuttuja	Ryhmä	SWAN lukumäärä (%)	SWAN: Valkoihoinen lukumäärä (%)	ERMA lukumäärä (%)	Sepelvaltimotauti aineistossa SWAN: lukumäärä (%)
Tupakointi	Ei koskaan	1753 (58.7)	751 (52.9)	393 (66.7)	58 (3.3)
	Tupakoinut menneisyydessä	742 (24.8)	459 (32.3)	160 (27.2)	29 (3.9)
	Tupakoi nykyään	472 (15.8)	208 (14.7)	33 (5.6)	30 (6.4)
	Tieto puuttuu	21 (0.7)	1 (0.1)	3 (0.5)	2 (9.5)

TAULUKKO 4. Tupakointi aineistoissa ja sepelvaltimotauti tupakoinnin mukaan SWAN-aineistossa

2.7. Siviilisääty aineistoissa

SWAN-aineistossa siviilisäätyä kuvaa neliluokkainen muuttuja, jonka vaihtoehtoja ovat naimaton – avioliitossa/avoliitossa – eronnut – leski (engl. single/never married – currently married/living as married – separated/divorced – widowed). Suurin osa aineiston naisista on avio- tai avoliitossa (67.3%) (Taulukko 5). Eronneita on 17.4%, naimattomia 13.3% ja leskiä 1.9%. Tieto siviilisäädystä puuttuu alle prosentilta. SWAN-aineistossa sepelvaltimotauti vaikuttaisi puhkeavan useammin leskille verrattuna muihin ryhmiin. Kaplanin-Meierin kestofunktioestimaatin 95 prosentin luottamusvälit ovat kuitenkin melko päällekkäiset (Kuva 2.9).

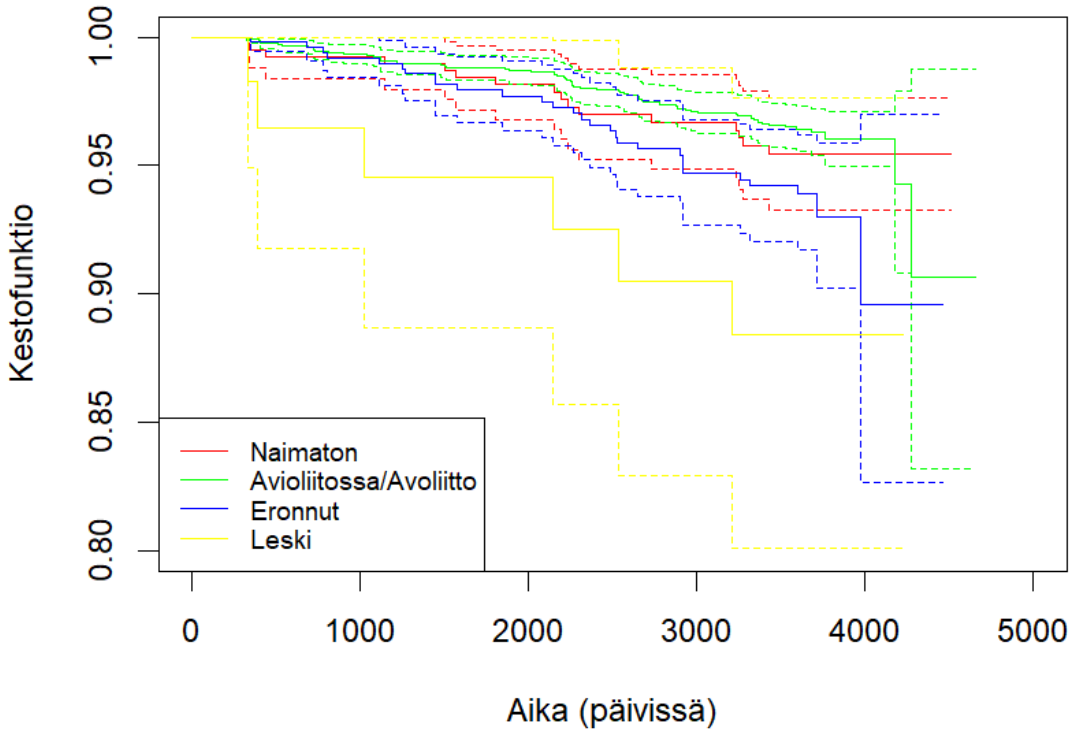


KUVA 2.8. SWAN-aineiston Kaplanin–Meierin estimaatti sepelvaltimotautiin sairastumisesta 95% luottamusvälillä tupakoinnin mukaan

Muuttuja	Ryhmä	SWAN lukumäärä (%)	SWAN: Valkoihoinen lukumäärä (%)	ERMA lukumäärä (%)	Sepelvaltimotauti aineistossa SWAN: lukumäärä (%)
Siviilisääty	Naimaton	398 (13.3)	170 (12.0)	56 (9.5)	16 (4.0)
	Avio-/avoliitossa	2011 (67.3)	1029 (72.5)	451 (76.6)	68 (3.4)
	Eronnut	519 (17.4)	202 (14.2)	78 (13.2)	29 (5.6)
	Leski	57 (1.9)	16 (1.1)	3 (0.5)	6 (10.5)
	Tieto puuttuu	3 (0.1)	2 (0.1)	1 (0.2)	

TAULUKKO 5. Siviilisääty aineistoissa ja sepelvaltimotauti siviilisäädyn mukaan SWAN-aineistossa

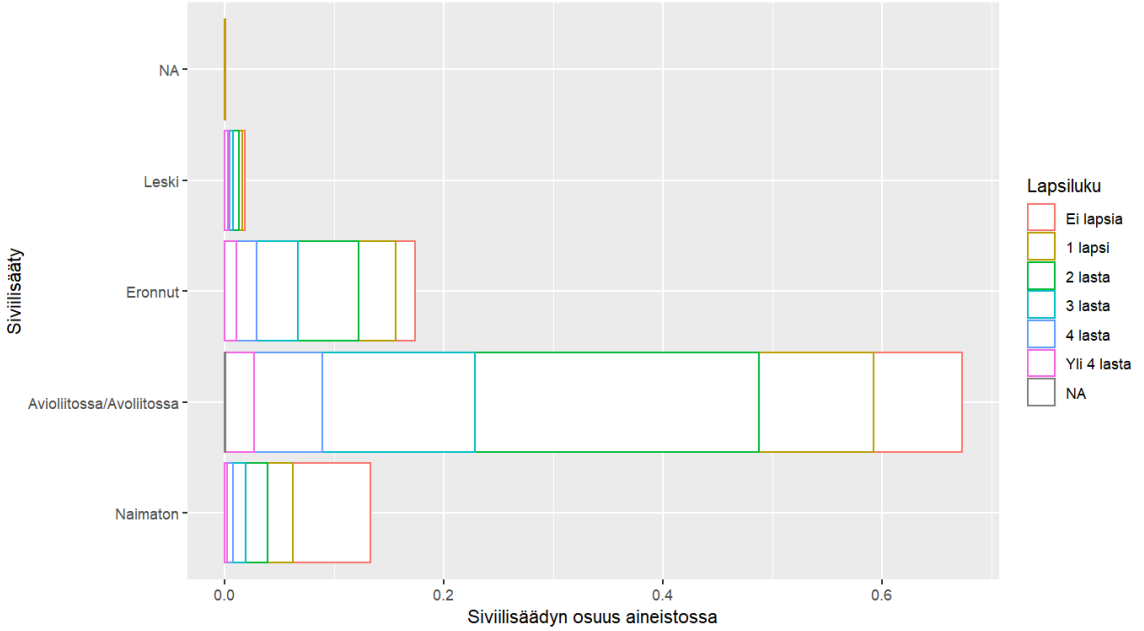
ERMA-tutkimuksen terveyskyselyssä vastaaja on valinnut siviilisäätynsä vaihtoehdoista naimaton – avioliitossa tai rekisteröidyssä parisuhteessa – uudessa avioliitossa – avoliitossa – eronnut tai asumuserossa – leski – muu. Tässä tutkielmassa



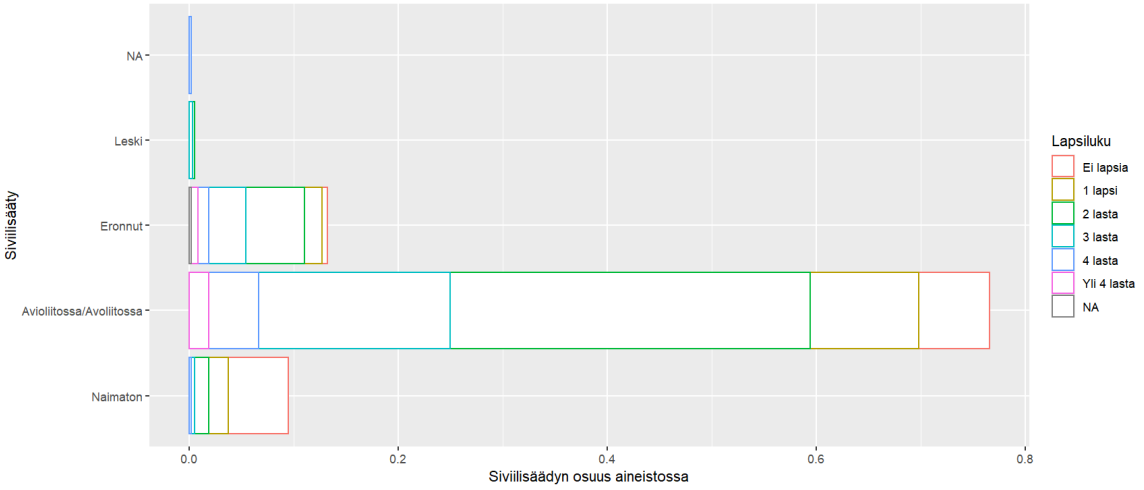
Kuva 2.9. SWAN-aineiston Kaplanin–Meierin estimaatti sepelvaltimotautiin sairastumisesta 95% luottamusvälillä siviilisäädyn mukaan

vaihtoehto "muu" on tulkittu puuttuvaksi tiedoksi, sillä heidän siviilisäädystä ei ole tietoa, eikä SWAN-aineistossa tätä vastausvaihtoehtoa ole. Siviilisäädyt on jaoteltu tätä tutkielmaa varten neljään ryhmään kuten SWAN-aineistossa: naimaton – avioliitossa/avoliitossa – eronnut – leski. Ryhmien osuudet aineistossa ovat 9.5%, 76.6%, 13.2% ja 0.5% (Taulukko 5). Avioliitossa/avoliitossa -ryhmään on luettu myös rekisteröidyt parisuhteet ja uusi avioliitto.

Aineistojen lapsiluvut vaihtelevat siviilisäädyn mukaan (Kuva 2.10 ja 2.11). Lapsettomuus on yleisintä naimattomien ryhmässä (noin 50–60%). Yli 4-lapsisuus on SWAN-aineistossa yleisintä leskien parissa (17.5%). ERMA-aineistossa leskien ryhmä on pieni, joten osuudet eivät välttämättä ole vertailukelpoisia.



KUVA 2.10. Lapsiluku siviilisäädyn mukaan SWAN-aineistossa, NA = tuntematon



KUVA 2.11. Lapsiluku siviilisäädyn mukaan ERMA-aineistossa, NA = tuntematon

LUKU 3

Kausaalivaikutuksen siirto populaatiosta toiseen

Tutkielman tavoitteena on selvittää, onko lapsiluvun vaikutus sepelvaltimotautiin siirrettävissä yhdysvaltalaisesta populaatiosta jyväskyläläiseen keski-ikäisillä naisilla. Lisäksi selvitetään, millaisin oletuksen kausaalivaikutuksen siirto on mahdollista, ja mitä tietoa populaatioista tarvitaan kausaalivaikutuksen estimoimiseksi kohdepopulaatiossa. Ensin määritellään ilmiöön liittyvä kausaalimalli, ja selvitetään, onko kiinnostava kausaalivaikutus identifioituva. Tämän jälkeen määritellään omassa kappaleessaan kausaalivaikutuksen siirto, ja sitä varten valikoitumismuuttujat ja -graafi. Kausaalimallille esitetään kaksi vaihtoehtoa, joista toisessa BMI on oletettu havaitsemattomaksi (latentiksi) muuttujaksi.

Luku perustuu pääosin artikkeliin "External validity: From do calculus to transportability across populations" (Pearl ja Bareinboim (2014)). Artikkelissaan Pearl ja Bareinboim (2014) tarjoavat muodollisen tavan esittää kahden populaation väliset erot. Lisäksi he näyttävät tavan selvittää, voidaanko kausaalisuhteita kohdepopulaatiossa tulkita lähdepopulaation kokeellisen tutkimusaineiston perusteella, ja mitä havainnoivaa aineistoa kohdepopulaatiosta ja kokeellista aineistoa lähdepopulaatiosta tarvitaan kausaaliyhteyden määrittämiseksi.

Kuten Pearl ja Bareinboim (2014) artikkelissaan toteavat, kausaaliyhteyden siirto populaatiosta toiseen on mahdollista myös, kun halutaan yleistää yhdessä populaatiossa tehdyn tutkimuksen tulokset toiseen populaatioon tilanteessa, jossa kummastakin populaatiosta on tarjolla vain havainnoivaa aineistoa. Kausaalisuhde voitaisiin tietenkin määrittää suoraan kohdepopulaation havainnoivaa aineistoa hyödyntäen. Kuitenkin, jos lähdepopulaation saatavilla olevan aineiston pohjalta voidaan selvittää, mitä tietoa kohdepopulaatiosta tarvitaan kausaalisuhteen selvittämiseksi jo ennen aineiston keruuta, voidaan resursseja säästää. Tässä tutkielmassa lähde- ja kohdepopulaatiosta on saatavilla vain havainnoivaa aineistoa. Selvitetään, mitä tietoja kummastakin populaatiosta tarvitaan lapsiluvun ja sepelvaltimotaudin kausaaliyhteyden määrittämiseksi kohdepopulaatiossa, jos oletetaan, että sepelvaltimotaudista ei ole tietoa kohdepopulaatiossa.

3.1. Kausaalimalli ja kausaalivaikutuksen identifioituvuus

Kiinnostuksen kohteena oleva kausaalivaikutus on lapsiluvun vaikutus sepelvaltimotautiin. Ilmiöön liittyvää kausaalimallia merkitään kirjaimella M. Tässä kappaleessa määritellään kausaalimalli, ja kausaalivaikutuksen identifioituvuus.

3.1.1. Kausaalimalli. Kausaalimalli voidaan määritellä rakenneyhtälöiden avulla.

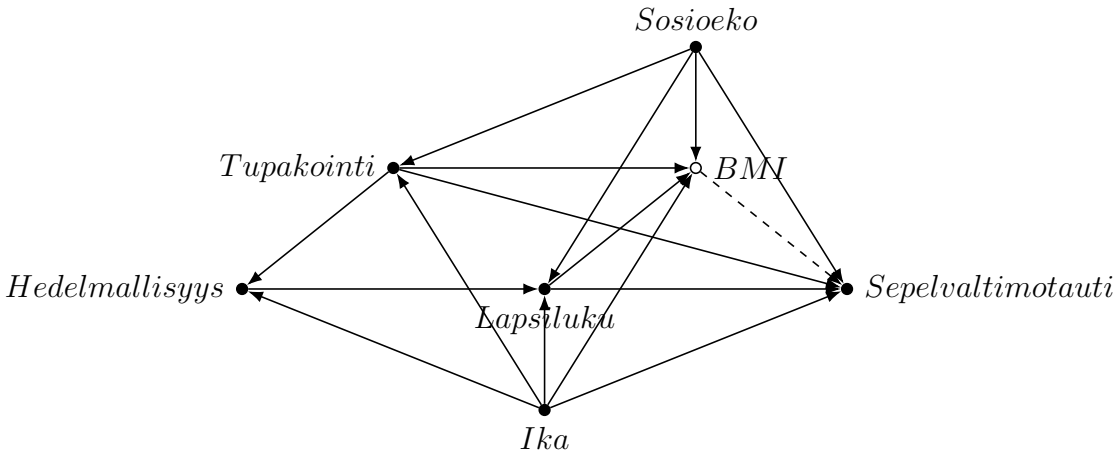
Määritelmä 1. (Kausaalimalli, Pearl (2009) 7.1.1). Kausaalimalli on nelikko $< \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(u) >$, missä

(1) $\boldsymbol{U}$ on joukko taustamuuttujia tai ulkoisia muuttujia, jotka edustavat mallin ulkopuolisia tekijöitä, ja jotka kuitenkin vaikuttavat mallin suhteisiin.
(2) $\boldsymbol{V} = V_1, \ldots, V_n$ on joukko mallin sisäisiä muuttujia, jotka oletetaan havaituiksi. Jokainen näistä muuttujista on toiminnallisesti riippuvainen jostain joukon $\boldsymbol{U} \cup \boldsymbol{V}$ alijoukosta PA_i.
(3) $\boldsymbol{F} = f_1, \ldots, f_n$ on joukko funktioita siten, että jokainen funktio f_i määrää muuttujan $V_i \in \boldsymbol{V}$ arvon, $v_i = f_i(pa_i, u)$. Tätä kutsutaan rakenneyhtälöksi.
(4) $P(u)$ on yhteistodennäköisyysjakauma yli joukon $\boldsymbol{U}$.

Kausaalimallia voidaan havainnollistaa graafin avulla. Graafin perusteella voidaan päätellä parametrittomat rakenneyhtälöt kausaalimallissa, ja rakenneyhtälöiden perusteella voidaan päätellä kausaalimallin graafi. Kausaalimallia rakenneyhtälöillä tai graafilla kuvatessa ei muuttujien jakaumista tarvitse tehdä parametrisia oletuksia.

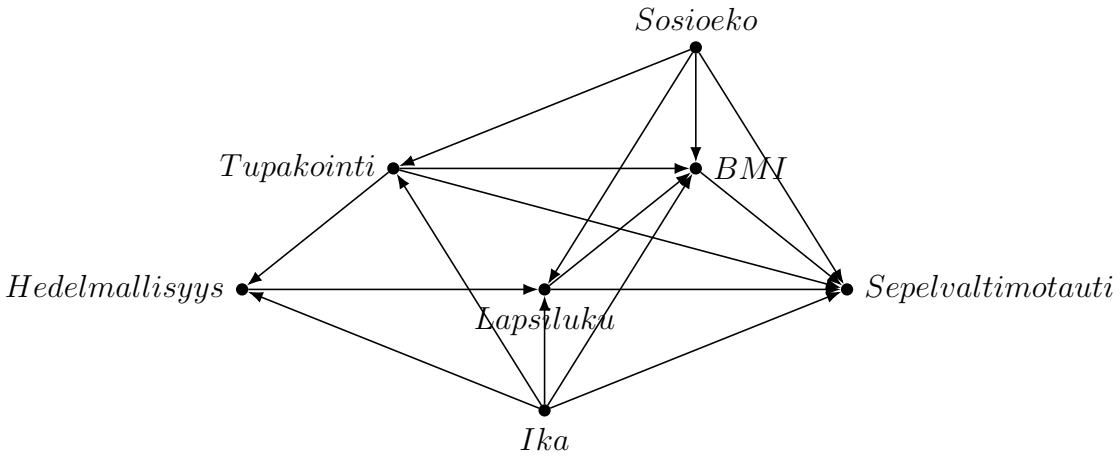
Graafissa solmu voi kuvata yhtä muuttujaa, tai muuttujien joukkoa. Graafista nähdään muuttujien välisiä mahdollisia kausaaliyhteyksiä, joita kuvataan nuolilla eli särmillä. Graafin särmät kertovat kausaalimalliin liittyvät oletukset, jotka perustellaan yleensä aikaisemman tiedon pohjalta. Särmillä kuvattuja muuttujien tai muuttujajoukkojen välisiä yhteyksiä kutsutaan poluiksi.

Tässä tutkielmassa lapsiluvun vaikutus sepelvaltimotautiin esitetään kahden vaihtoehtoisen graafin avulla (Kuva 3.1 ja 3.2). Kausaalimalliin M liittyviä oletuksia havainnollistavat graafiin lisätyt taustatekijät, jotka voidaan perustella ilmiöön liittyvän tiedon (kts. luku 1) ja tutkielman aineiston perusteella (kts. luku 2).



Kuva 3.1. Kausaalimalli lapsiluvun vaikutukselle sepelvaltimotaudin riskiin, jossa BMI on oletettu latentiksi muuttujaksi

Graafista nähdään, mitä oletuksia ilmiöstä on tehty. Esimerkiksi oletetaan, että hedelmällisyys vaikuttaa lapsilukuun, ja että hedelmällisyyteen vaikuttavia tekijöitä ovat taas ikä ja tupakointi. “Sosioeko” kuvaa koulutusastetta, siviilisäätyä ja etnistä taustaa, sillä niiden suhteet toisiin muuttujiin ovat identtiset. Kuvan 3.1 graafissa BMI oletetaan latentiksi muuttujaksi ja kuvan 3.2 graafissa BMI on havaittu. Kausaalimallin M graafia vastaava rakenneyhtälö on



KUVA 3.2. Kausaalimalli lapsiluvun vaikutukselle sepelvaltimotaudin riskiin, jossa kaikki graafin muuttujat on havaittu

$$
\begin{aligned}
Sosioeko &= f_{Sosioeko}(U_{Sosioeko}),\\
Ika &= f_{Ika}(U_{Ika}),\\
Tupakointi &= f_{Tupakointi}(Sosioeko, Ika, U_{Tupakointi}),\\
Hedelmallisyys &= f_{Hedelmallisyys}(Tupakointi, Ika, U_{Hedelmallisyys}),\\
Lapsiluku &= f_{Lapsiluku}(Hedelmallisyys, Sosioeko, Ika, U_{Lapsiluku}),\\
(BMI &= f_{BMI}(Sosioeko, Ika, Tupakointi, Lapsiluku, U_{BMI},)\\
Sepelvaltimotauti &= f_{Sepelvaltimotauti}(Lapsiluku, BMI, Sosioeko, Ika, Tupakointi, U_{Sepelvaltimotauti}),
\end{aligned}
$$

missä BMI:lle on oma funktio, jos se oletetaan havaituksi.

3.1.2. Kausaalivaikutuksen identifioituvuus. Kiinnostuksen kohteena oleva kausaalivaikutus on lapsiluvun vaikutus sepelvaltimotautiin. Halutaan tietää, mikä on sepelvaltimotaudin riski lapsiluvun ollessa vakio. Lapsiluvun kausaalivaikutusta sepelvaltimotautiin selvitettäessä korvataan muuttujan *Lapsiluku* funktio vakiolla, esimerkiksi $Lapsiluku = 2$. Tätä muuttujan *Lapsiluku* kontrollointia, interventiota, merkitään do-operaattorilla eli $do(Lapsiluku)$. Siten kausaalimallissa M kiinnostava kausaalivaikutus on $Q(M) = P(Sepelvaltimotauti|do(Lapsiluku))$, ja halutaan tietää, onko kausaalivaikutus identifioituva.

MÄÄRITELMÄ 2. (Identifioituvuus, Pearl (2009) 3.2.3). Olkoon A joukko oletuksia. Kausaalivaikutus $Q(M)$ on identifioituva, jos mille tahansa parille kausaalimalleja M_1 ja M_2, joille oletukset A ovat voimassa, pätee

$$P_{M_1}(V) = P_{M_2}(V) \Rightarrow Q(M_1) = Q(M_2),$$

missä $P_M(V)$ on yhteisjakauma muuttujajoukon V yli.

Toisin sanoen, jos mallien M_1 ja M_2 havaintojakaumat ovat samat, myös interventiojakaumat ovat samat. Kun kiinnostuksen kohteena oleva kausaalivaikutus sisältää do-operaattorin, voidaan identifioituvuus selvittää myös systemaattisesti kolmea laskusääntöä hyödyntäen. Laskusääntöjen hyödyntäminen vaatii työvälineenä d-separoituvuutta, jolla voidaan todeta ehdollinen riippumattomuus.

MÄÄRITELMÄ 3. (d-separoituvuus, Pearl (2009) 1.2.3). Solmujoukko $\boldsymbol{S}$ d-separoi polun p, jos

(1) p sisältää ketjun $i \to m \to j$ tai haarukan $i \leftarrow m \to j$ siten, että keskisolmu m kuuluu solmujoukkoon $\boldsymbol{S}$ tai
(2) p sisältää ainakin yhden käänteisen haarukan $i \to j \leftarrow k$ siten, että $\boldsymbol{S}$ ei sisällä solmua j eikä yhtään solmun j jälkeläistä.

Jos joukko $\boldsymbol{S}$ d-separoi kaikki polut muuttujajoukosta $\boldsymbol{X}$ muuttujajoukkoon $\boldsymbol{Y}$, joukko $\boldsymbol{S}$ d-separoi joukot $\boldsymbol{X}$ ja $\boldsymbol{Y}$, ja muuttujajoukot $\boldsymbol{X}$ ja $\boldsymbol{Y}$ ovat riippumattomia ehdolla $\boldsymbol{S}$, eli $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{S}$.

Oletetaan, että $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{W}$, ja $\boldsymbol{Z}$ erään suunnatun graafin G solmujoukkoja, jotka edustavat satunnaisesti jakautuneita, erillisiä muuttujia. Merkinnällä $G_{\bar{\boldsymbol{X}}}$ tarkoitetaan graafia G, josta on poistettu solmujoukkoon $\boldsymbol{X}$ tulevat nuolet, ja merkinnällä $G_{\underline{\boldsymbol{X}}}$ graafia G, josta on poistettu solmujoukosta $\boldsymbol{X}$ lähtevät nuolet. Kun graafista on poistettu sekä tulevat että lähtevät nuolet, käytetään merkintää $G_{\bar{\boldsymbol{X}}\underline{\boldsymbol{Z}}}$.

Kausaalilaskennan kolme laskusääntöä (Pearl, 1995) ovat

SÄÄNTÖ 1. (Havaintojen lisääminen ja poistaminen).

$$P(\boldsymbol{y}|do(\boldsymbol{x}), \boldsymbol{z}, \boldsymbol{w}) = P(\boldsymbol{y}|do(\boldsymbol{x}), \boldsymbol{w}),$$

jos $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ graafissa $G_{\bar{\boldsymbol{X}}}$.

SÄÄNTÖ 2. (Toiminnan ja havainnon vaihtaminen).

$$P(\boldsymbol{y}|do(\boldsymbol{x}), do(\boldsymbol{z}), \boldsymbol{w}) = P(\boldsymbol{y}|do(\boldsymbol{x}), \boldsymbol{z}, \boldsymbol{w}),$$

jos $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ graafissa $G_{\bar{\boldsymbol{X}}\underline{\boldsymbol{Z}}}$.

SÄÄNTÖ 3. (Toiminnan lisääminen ja poistaminen).

$$P(\boldsymbol{y}|do(\boldsymbol{x}), do(\boldsymbol{z}), \boldsymbol{w}) = P(\boldsymbol{y}|do(\boldsymbol{x}), \boldsymbol{w}),$$

jos $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ graafissa $G_{\overline{\boldsymbol{X}Z(W)}}$, jossa $Z(W)$ on joukko Z-solmuja, jotka eivät ole minkään W-solmun jälkeläisiä graafissa $G_{\overline{\boldsymbol{X}}}$.

Do-search on algoritmi, joka hyödyntää edellä esitettyjä kolmea laskusääntöä (Tikka, Hyttinen ja Karvanen, 2021a). Algoritmin avulla voidaan selvittää, onko tietty kausaalivaikutus identifioituva kausaalisuhteista saatavilla olevaa tietoa (oletuksia) edustavan graafin perusteella. Kun kausaalivaikutus on identifioituva, algoritmi palauttaa kausaalivaikutuksen estimoimiseen tarvittavan lausekkeen. Jos kausaali- ja todennäköisyyslaskentaa hyödyntäen ei voida saada yksikäsitteistä lauseketta, kausaalivaikutus ei ole identifioituva ja algoritmi palauttaa NA (=ei saatavilla). Joissain harvinaisissa tilanteissa on mahdollista, että algoritmi ei löydä ratkaisua, vaikka kausaalivaikutus olisi identifioituva. Algoritmia voidaan hyödyntää myös tilanteissa, jossa datalähteitä on useita, ja tilanteissa, joissa esiintyy valikoitumisharhaa. Do-search

-algoritmia voidaan käyttää työvälineenä myös siirrettäessä kausaalivaikutusta populaatiosta toiseen, mitä sovelletaan tässä tutkielmassa.

Kuvan 3.1 tai 3.2 mukaisen kausaalimallin graafin perusteella kausaalivaikutus $P(Sepelvaltimotauti|do(Lapsiluku))$ on identifioituva kaavalla

$$\begin{aligned} &P(Sepelvaltimotauti|do(Lapsiluku)) \\ &= \sum_{Tupakointi,Ika,Sosioeko} p(Tupakointi, Ika, Sosioeko) \\ &\quad p(Sepelvaltimotauti|Lapsiluku, Tupakointi, Ika, Sosioeko) \end{aligned} \tag{3.1}$$

Lauseke identifioituvuudelle on siis sama riippumatta siitä, oletetaanko BMI latentiksi vai ei.

3.2. Siirrettävyys

Tässä kappaleessa esitellään kausaalivakutuksen siirrettävyyteen liittyvää teoriaa. Ensin määritellään valikoitumisgraafi, ja sitten siirrettävyys.

3.2.1. Valikoitumismuuttujat ja valikoitumisgraafi. Kausaalivaikutuksen siirtäminen populaatiosta toiseen edellyttää tietoa populaatioiden yhtäläisyyksistä ja eroavaisuuksista. Erojen taustalla olevia mekanismeja kuvataan valikoitumismuuttujilla. Yksittäisen muuttujan suhteen eroja aiheuttavia mekanismeja kuvataan yhdellä valikoitumismuuttujalla $t_i, i = 1, .., n$, ja joukko $\boldsymbol{T}$ sisältää kaikki valikoitumismuuttujat eli tekijät, joiden suhteen populaatiot eroavat toisistaan.

Määritelmä 4. (Valikoitumisgraafi, Pearl ja Bareinboim (2014)). Olkoot $\langle M, M^* \rangle$ populaatioita $\langle \Pi, \Pi^* \rangle$ vastaavat kausaalimallit, ja joita vastaa sama graafi G. Kausaalimallien parista $\langle M, M^* \rangle$ saadaan valikoitumisgraafi D, jos D rakentuu seuraavasti:

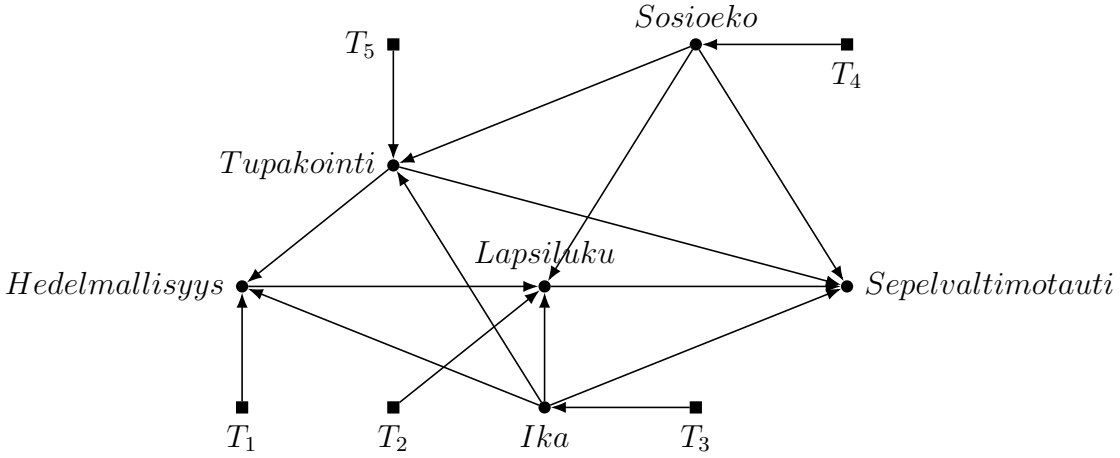
(1) Jokainen särmä graafissa G on myös valikoitumisgraafissa D.
(2) Valikoitumisgraafissa D on lisäksi särmä $T_i \to V_i$ aina, kun oletetaan erisuuruus $f_i \neq f_i^*$ tai $P(U_i) \neq P^*(U_i)$ kausaalimallien M ja M^* välillä.

Edellä esitetyn määritelmän merkinnöillä f_i, V_i, U_i ja $P(U_i)$ tarkoitetaan määritelmässä 1 esitettyjä muuttujia, funktioita ja todennäköisyysjakaumia.

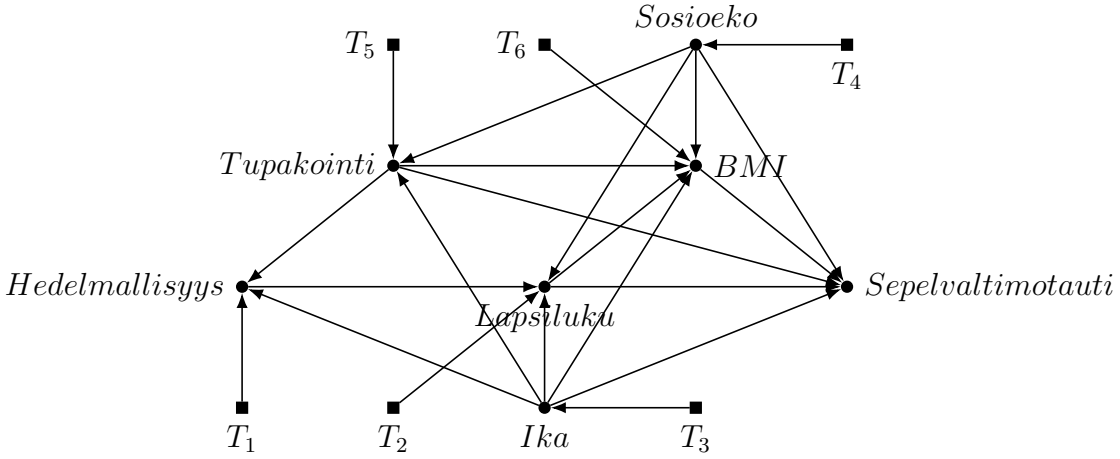
Koska kuvan 3.1 graafissa BMI on latentti muuttuja, saadaan graafista latentti projektio (Verma (1992)). Tämä tarkoittaa, että BMI, ja siihen tulevat ja siitä lähtevät särmät poistetaan. Latentista projektiosta saadaan valikoitumisgraafi (Kuva 3.3) siirtokaavaa varten.

Jos BMI oletetaan havaituksi, saadaan kuvan 3.2 graafista vaihtoehtoinen valikoitumisgraafi (Kuva 3.4). Valikoitumisgraafissa on lisäksi valikoitumismuuttuja T_6, josta on särmä muuttujaan BMI.

Valikoitumismuuttujia on korkeintaan graafissa esiintyvien muuttujien verran. Siinä tapauksessa oletetaan, että muuttujien jakaumia generoivat mekanismit ovat jokaisen muuttujan osalta erilaiset, eikä siirto ole mahdollista. Valikoitumisgraafissa on särmä valikoitumismuuttujasta T_i muuttujaan V_i aina, kun yhdysvaltalaisen ja jyväskyläläisen populaation välillä oletetaan ero kyseisen muuttujan jakaumassa. Siten oletetaan, että populaatioiden jakaumat lapsiluvun, hedelmällisyyden, tupakoinnin,



KUVA 3.3. Yhdysvaltalaisen ja jyväskyläläisen populaation eroja kuvaava valikoitumisgraafi



KUVA 3.4. Yhdysvaltalaisen ja jyväskyläläisen populaation eroja kuvaava valikoitumisgraafi, jossa BMI on mukana

iän, painoindeksin ja sosioekonomisten tekijöiden suhteen ovat erilaiset. Valikoitumismuuttujan puuttuminen taas kuvaa oletusta, että mekanismi, jolla saadaan muuttujan arvo, on sama kummassakin populaatiossa. Kun tällaisia oletuksia on riittävästi, kausaalivaikutuksen siirto mahdollistuu.

3.2.2. Siirrettävyys ja siirtokaava. Se, voidaanko kausaalivaikutus siirtää populaatiosta toiseen riippuu muuttujien välisistä kausaalisuhteista eli oletuksista, jotka on kuvattu graafissa. Muuttujien tilastollisilla ominaisuuksilla ei siten ole väliä, vaan siirtämisen onnistuminen voidaan todeta parametrittomasti.

MÄÄRITELMÄ 5. (Triviaali siirrettävyys, Pearl ja Bareinboim (2014)). Kausaalisuhde R on triviaalisti siirrettävissä populaatiosta Π populaatioon Π^*, jos $R(\Pi^*) = P^*(y|do(x))$ on identifioituva graafin G^* ja (havainto)jakauman P^* perusteella.

Edellisen määritelmän perusteella kausaalivaikutus kohdepopulaatiossa Π^* voidaan estimoida suoraan kohdepopulaation havainnollista aineistoa hyödyntäen ilman

tietoja lähdepopulaatiosta Π. Kohdepopulaatiosta ei kuitenkaan aina ole saatavilla kaikkea tarvittavaa tietoa. Tässä tutkielmassa oletetaan, että tieto kohdepopulaation sepelvaltimotaudista puuttuu. Seuraava lause (Pearl ja Bareinboim (2014)) antaa konkreettisemmat välineet siihen, kuinka kahden populaation tietoja voidaan hyödyntää kausaalivaikutuksen estimoimiseksi kohdepopulaatiossa.

LAUSE 1. Olkoon D valikoitumisgraafi, joka luonnehtii kahta populaatiota Π ja Π^*, ja $\boldsymbol{T}$ joukko valikoitumismuuttujia valikoitumisgraafissa D. Kausaalisuhde $R = P^*(y|do(x))$ on siirrettävissä populaatiosta Π populaatioon Π^*, jos lauseke $P(y|do(x), t)$ voidaan sieventää kausaalilaskennan sääntöjä käyttäen muotoon, jossa $\boldsymbol{T}$ esiintyy vain ehdollistavana muuttujana termeissä, joissa ei ole do-operaattoria.

Saatua lauseketta kutsutaan siirtokaavaksi. Edellisen lauseen kausaalilaskennassa voidaan hyödyntää do-search -algoritmia. Kausaalisuhde $P^*(Sepelvaltimotauti|do(Lapsiluku))$ on siirrettävissä yhdysvaltalaisesta populaatiosta jyväskyläläiseen, sillä lauseen 1 mukaan lauseke $P(Sepelvaltimotauti|do(Lapsiluku), t)$ voidaan sieventää kausaalilaskennan sääntöjä käyttäen muotoon, jossa $\boldsymbol{T}$ esiintyy vain ehdollistavana muuttujana termeissä, joissa ei ole do-operaattoria. Kun BMI oletetaan latentiksi, saatu lauseke eli siirtokaava on

$$
\begin{aligned}
&P^*(Sepelvaltimotauti|do(Lapsiluku)) \\
&\quad = P(Sepelvaltimotauti|do(Lapsiluku), t) \\
&\quad = \sum_{Tup,Sos,Ika} p(Tup, Sos, Ika) \\
&\quad * p(Sepelvaltimotauti|Lapsiluku, Tup, Sos, Ika, T_1, T_2, T_3, T_4, T_5)
\end{aligned} \tag{3.2}
$$

Jos BMI oletetaan havaituksi, siirtokaava on

$$
\begin{aligned}
&P^*(Sepelvaltimotauti|do(Lapsiluku)) \\
&\quad = P(Sepelvaltimotauti|do(Lapsiluku), t) \\
&\quad = \sum_{BMI,Tup,Sos,Ika} p(Tup, Sos, Ika) \\
&\quad * p(BMI|Laps, Tup, Sos, Ika) \\
&\quad * p(Sepelvaltimotauti|Lapsiluku, BMI, Tup, Sos, Ika, T_1, T_2, T_3, T_4, T_5, T_6)
\end{aligned} \tag{3.3}
$$

Siirtokaavan laskemiseksi on hyödynnetty do-search -algoritmia, joka on toteutettu R-paketissa "dosearch" (Tikka, Hyttinen ja Karvanen (2021b)). Käytetty R-koodi kaavalle 3.2 löytyy liitteistä (LIITE A). Kun BMI oletetaan latentiksi, siirtokaavan 3.2 mukaan yhdysvaltalaisesta populaatiosta tarvitaan tieto sepelvaltimotaudista, lapsiluvusta, sosioekonomisista tekijöistä sekä iästä kausaalivaikutuksen $P^*(Sepelvaltimotauti|do(Lapsiluku))$ estimoimiseksi. Jyväskyläläläisestä populaatiosta sen sijaan riittää tieto tupakoinnista, sosioekonomisista tekijöistä ja iästä. Siirtokaava tässä tapauksessa on samankaltainen kuin lauseke, joka saadaan kausaalivaikutuksen identifioimiseksi (kaava 3.1). Jos BMI taas oletetaan havaituksi, myös tieto painoindeksistä tarvitaan kummastakin populaatiosta siirtokaavan 3.3 mukaan.

Lapsiluku käsitetään ennusteessa ikään kuin interventiona, joten tietoa lapsiluvusta ei tarvita jyväskyläläisestä populaatiosta, vaikka se siirtokaavassa 3.3 onkin.

LUKU 4

Mallin valinta ja estimointi

Tässä luvussa määritellään malli, jolla lapsiluvun vaikutus sepelvaltimotautiin estimoidaan. Malli sovitetaan lähdepopulaation aineistoon eli SWAN-aineistoon. Edellisessä luvussa saatiin kaksi vaihtoehtoista siirtokaavaa (kaava 3.2 ja kaava 3.3), joita hyödyntäen määritellään kaksi vaihtoehtoista mallia. Ensimmäiseen malliin valitaan siirtokaavan 3.2 perusteella lähdepopulaatiosta eli SWAN-aineistosta selittäjiksi lapsiluku, ikä, tupakointi, koulutusaste, siviilisääty ja etninen tausta. Toiseen malliin valitaan siirtokaavan 3.3 perusteella SWAN-aineistosta selittäjiksi lapsiluvun, iän, tupakoinnin, koulutusasteen, siviilisäädyn ja etnisen taustan lisäksi BMI.

SWAN-aineisto on muokattu elinaika-aineistoksi, jossa seuranta alkaa päivästä 0 ja tutkimuskäynnit ovat noin vuoden välein riippuen tutkimushenkilöstä. Kiinnostava tapahtuma on sepelvaltimotautiin sairastuminen, mikä tässä tapauksessa tarkoittaa tutkimushenkilön ensimmäistä kyllä-vastausta kysymykseen "Onko lääkäri, sairaanhoitaja tai muu terveysalan ammattilainen viime käynnin jälkeen todennut sinulla olevan jokin seuraavista sairauksista tai oletko saanut hoitoa johonkin seuraavista: sydänkohtaus tai rasitusrintakipua?".

Tutkielmassa tapahtuma-aika on tutkimuskäyntipäivä, vaikka todellisuudessa tapahtuma-aika on jokin päivä käyntien välissä. Tutkimuskäyntipäivän käyttö tapahtuma-aikana on siten yksinkertaistava oletus. Tutkielmassa käytettävässä aineistossa seuranta päättyy joko sairastumiseen tai käyntiin 10, jos sairastumista ei tapahdu. Seuranta-ajaksi saadaan siten kulunut aika päivissä seurannan alusta joko sairastumiseen tai käyntiin 10.

Osa tutkimushenkilöistä on jättänyt tutkimuksen kesken tai osa seurannan käynneistä puuttuu. Näille henkilöille seuranta-ajaksi asetetaan aika päivissä seurannan alusta viimeisimpään käyntiin. Jokaisen tutkimushenkilön kohdalla käynnit tallennetaan omaksi havainnoikseen, jolloin sekä tutkimuskäyntien välinen aika että iän päivittyminen tulee huomioitua. Seurannan keskeyttäneiden tutkimushenkilöiden keskeytyksen syy ei ole tiedossa, joten syynä voi olla esimerkiksi kuolema.

Seuranta-aika on siis jaettu noin vuoden mittaisiin osiin, joiden aikana kahden sovitettavan mallin selittäjistä ikä oletetaan vakioksi. Muut mallien selittäjät ovat seurannan alun mittauksista. Seuranta-ajan määritys ja aineiston muokkaus siten, että iän päivittymien ajan kuluessa tulee huomioitua, on esitetty R-koodissa liitteessä A.

Aineisto on oikealta sensuroitua. Tässä tapauksessa se tarkoittaa sitä, että jos yksilö ei ole seurannan aikana sairastunut, ei voida tietää, sairastuuko yksilö käynnin 10 (tai viimeisemmän käynnin) jälkeen, ja jos sairastuu, milloin se tapahtuu.

4.1. Coxin malli

Lapsiluvun vaikutus sepelvaltimotautiin estimoidaan hyödyntäen Coxin mallia (Cox (1972)). Coxin mallissa uhkafunktio on muotoa

$$h(t;\boldsymbol{x}) = h_0(t)e^{\boldsymbol{\beta}^T\boldsymbol{x}},$$

missä $h_0(t)$ on tuntematon funktio, joka antaa uhan referenssitasolle, eli kun $\boldsymbol{x} = 0$. $\boldsymbol{\beta}^T$ on vektori mallin regressiokertoimia ja $\boldsymbol{x}$ mallin selittäjiä. Malli sovitetaan osittaisuskottavuuspäättelyn avulla. Osittaisuskottavuus (Cox (1975)) on muotoa

$$\prod_{i=1}^{n} \frac{e^{\beta^T\boldsymbol{x}_{(i)}}}{\sum_{j\in R(t_i)} e^{\beta^T\boldsymbol{x}_j}},$$

missä $\boldsymbol{x}_j$ on muuttujan x saama arvo j:nnelle yksilölle ja $\boldsymbol{x}_{(i)}$ arvo yksilölle, joka epäonnistuu ajanhetkellä t_i. $R(t)$ käsittää yksilöt, jotka eivät ole sensuroituneet tai jotka eivät ole epäonnistuneet ajanhetkeen t mennessä eli

$$R(t) = \{i : t_i \geq t\}$$

Epäonnistumisella tarkoitetaan kiinnostuksen kohteena olevaa tapahtumaa, joka tässä tutkielmassa on sepelvaltimotaudin puhkeminen.

SWAN-aineistoon sovitetaan Coxin malli, johon ikä lisätään ajasta riippuvana kovariaattina (T. Therneau, Crowson ja Atkinson (2023)). Malli on

$$h(t;\boldsymbol{x}) = h_0(t)e^{\boldsymbol{\beta}^T\boldsymbol{x}(t)},$$

missä t on aika päivissä ja mallissa, jossa BMI on oletettu latentiksi,

$$\begin{aligned}\boldsymbol{\beta}^T\boldsymbol{x} = {} & \beta_{ika}ika(t) + \beta_{lapsia1}lapsia1 + \beta_{lapsia2}lapsia2 + \beta_{lapsia3}lapsia3 + \beta_{lapsia4}lapsia4 \\ & + \beta_{lapsiaYli4}lapsiaYli4 + \beta_{tupakoiEnnen}tupakoiEnnen + \beta_{tupakoiNyt}tupakoiNyt \\ & + \beta_{kouluPerus}kouluPerus + \beta_{koulu2aste}koulu2aste + \beta_{kouluYlempi}kouluYlempi \\ & + \beta_{naimaton}naimaton + \beta_{eronnut}eronnut + \beta_{leski}leski \\ & + \beta_{afrikkalaisamerikkalainen}afrikkalaisamerikkalainen \\ & + \beta_{kiinalaisamerikkalainen}kiinalaisamerikkalainen \\ & + \beta_{japanilaisamerikkalainen}japanilaisamerikkalainen \\ & + \beta_{latinalaisamerikkalainen}latinalaisamerikkalainen\end{aligned}$$

Mallissa, jossa BMI on mukana,

$$\begin{aligned}\boldsymbol{\beta}^T\boldsymbol{x} = {} & \beta_{ika}ika(t) + \beta_{lapsia1}lapsia1 + \beta_{lapsia2}lapsia2 + \beta_{lapsia3}lapsia3 + \beta_{lapsia4}lapsia4 \\ & + \beta_{lapsiaYli4}lapsiaYli4 + \beta_{tupakoiEnnen}tupakoiEnnen + \beta_{tupakoiNyt}tupakoiNyt \\ & + \beta_{kouluPerus}kouluPerus + \beta_{koulu2aste}koulu2aste + \beta_{kouluYlempi}kouluYlempi \\ & + \beta_{naimaton}naimaton + \beta_{eronnut}eronnut + \beta_{leski}leski \\ & + \beta_{afrikkalaisamerikkalainen}afrikkalaisamerikkalainen \\ & + \beta_{kiinalaisamerikkalainen}kiinalaisamerikkalainen \\ & + \beta_{japanilaisamerikkalainen}japanilaisamerikkalainen \\ & + \beta_{latinalaisamerikkalainen}latinalaisamerikkalainen \\ & + \beta_{bmi}bmi\end{aligned}$$

Referenssitasona kummassakin versiossa on lapseton, ei koskaan tupakoinut, korkeakoulun suorittanut, avio- tai avoliitossa oleva ja valkoihoinen.

4.2. Estimointi

Mallin avulla on tarkoitus ennustaa sepelvaltimotautitapausten lukumäärä kohdepopulaatiossa koko seuranta-ajalle. Ennuste lasketaan perusuhkakertymän avulla hyödyntäen Breslowin estimaattoria, joka määritellään seuraavasti (Lin (2007)):

$$\hat{H}_{0,i} = \sum_{i=1}^{n} \hat{h}_{0,i} = \sum_{i=1}^{n} \frac{\delta_{(i)}}{\sum_{j \in R(t_{(i)})} e^{\hat{\beta^T} x_j(t_{(i)})}},$$

missä $\hat{h}_{0,i}$ on aikavälin $t_{(i)} \leq t < t_{(i+1)}$ perusuhka ja $\delta_{(i)}$ on tapahtumien lukumäärä kyseisellä aikavälillä.

Coxin malli voidaan sovittaa aineistoon R:ssä survival-paketin coxph-funktiolla. Odotetut tapahtumien lukumäärät seuranta-ajalle saadaan antamalla predict.coxph-funktiolle argumentiksi type="expected". (T. M. Therneau et al. (2023)). Tällä funktiolla saadut ennusteet summataan, jolloin saadaan odotettu tapahtumien lukumäärä aineistossa seuranta-ajalla.

Luottamusväli tapahtumien lukumäärän odotusarvolle estimoidaan uusio-otantamenetelmällä. Uusio-otantamenetelmän vaiheet 95 prosentin luottamusvälin laskemiseksi keskiarvolle on muokattu tähän tutkielmaan soveltuviksi (Ramachandran ja Tsokos (2021) 13.3.1). Vaiheet ovat

(1) Arvotaan palauttaen N otosta alkuperäisestä aineistosta.
(2) Summataan jokaisen otoksen odotetut tapahtumien lukumäärät, jolloin saadaan odotettu tapahtumien lukumäärä otoksessa. Odotetut tapahtumien lukumäärät on laskettu R:llä predict.coxph-funktiolla asettamalla argumentiksi type="expected".
(3) Järjestetään otosten odotetut tapahtumien lukumäärät suuruusjärjestykseen.
(4) Keskellä olevat 95 prosenttia otosten odotetuista tapahtumien lukumääristä ovat paikkojen (0.025)(N + 1) ja (0.975)(N + 1) välissä. Jos paikat eivät ole kokonaislukuja, ne pyöristetään lähimpään kokonaislukuun. Näiden paikkojen arvot ovat uusioluottamusvälin ylä- ja alaraja tapahtumien lukumäärän odotusarvolle.

Uusio-otantamenetelmällä luottamusväliä laskettaessa saattaa tulos toistettaessa olla hieman erilainen. Tässä tutkielmassa arvotaan $N = 1000$ otosta alkuperäisestä ERMA-aineistosta. Otosten koot ovat 589, kuten ERMA-aineiston koko.

LUKU 5

Tulokset

Kiinnostuksen kohteena oleva kausaalivaikutus on lapsiluvun vaikutus sepelvaltimotautiin eli $P(Sepelvaltimotauti|do(Lapsiluku))$. Edellisessä luvussa määritettiin kausaalivaikutuksen estimointiin käytettävä malli. Tässä luvussa sovitetaan malli SWAN-aineistoon ja estimoidaan lapsiluvun vaikutus sepelvaltimotautiin ERMA-aineisossa.

5.1. Kausaalivaikutuksen estimointi SWAN-aineistolla

Edellisessä luvussa määritetty Coxin malli sovitetaan SWAN-aineistoon. Mallin sovitukseen käytetty R-koodi löytyy liitteestä A. Taulukossa 1 on mallin, jossa BMI on oletettu latentiksi, mukaiset estimaatit β-kertoimille, niiden keskivirheet, uhkasuhteet ja 95 prosentin luottamusvälit uhkasuhteille.

β-kerroin	Estimaatti	Keskivirhe	Uhkasuhde	Luottamusväli
β_{ika}	0.01	0.03	1.01	[0.94, 1.07]
$\beta_{lapsia1}$	0.25	0.42	1.28	[0.57, 2.90]
$\beta_{lapsia2}$	0.41	0.38	1.51	[0.71, 3.19]
$\beta_{lapsia3}$	0.44	0.41	1.56	[0.70, 3.45]
$\beta_{lapsia4}$	1.12	0.41	3.07	[1.36, 6.90]*
$\beta_{lapsiaYli4}$	0.99	0.47	2.69	[1.06, 6.78]*
$\beta_{tupakoiEnnen}$	0.26	0.23	1.29	[0.82, 2.05]
$\beta_{tupakoiNyt}$	0.54	0.24	1.71	[1.08, 2.71]*
$\beta_{kouluPerus}$	1.00	0.40	2.72	[1.23, 5.97]*
$\beta_{koulu2aste}$	0.45	0.30	1.56	[0.87, 2.80]
$\beta_{kouluYlempi}$	0.28	0.35	1.33	[0.67, 2.64]
$\beta_{naimaton}$	0.09	0.31	1.09	[0.60, 2.00]
$\beta_{eronnut}$	0.25	0.23	1.28	[0.82, 2.02]
β_{leski}	0.64	0.44	1.90	[0.80, 4.53]
$\beta_{afrikkalaisamerikkalainen}$	0.68	0.22	1.98	[1.28, 3.07]*
$\beta_{kiinalaisamerikkalainen}$	-0.17	0.49	0.84	[0.32, 2.21]
$\beta_{japanilaisamerikkalainen}$	-2.01	1.01	0.13	[0.02, 0.98]*
$\beta_{latinalaisamerikkalainen}$	0.29	0.40	1.34	[0.62, 2.91]

TAULUKKO 1. Estimaatit mallille, jossa BMI latentti, *-merkillä tilastollisesti merkitsevät 5 prosentin riskitasolla

Mallin mukaan β-kertoimista $\beta_{lapsia4}$, $\beta_{lapsiaYli4}$, $\beta_{tupakoiNyt}$, $\beta_{kouluPerus}$, $\beta_{afrikkalaisamerikkalainen}$ ja $\beta_{japanilaismaerikkalainen}$ ovat merkitseviä.

Taulukossa 2 on taas mallin, jossa BMI on oletettu havaituksi, mukaiset estimaatit β-kertoimille, niiden keskivirheet, uhkasuhteet ja 95 prosentin luottamusvälit uhkasuhteille.

β-kerroin	Estimaatti	Keskivirhe	Uhkasuhde	Luottamusväli
β_{ika}	-0.01	-0.33	0.99	[0.92, 1.06]
$\beta_{lapsia1}$	0.20	0.47	1.22	[0.53, 2.82]
$\beta_{lapsia2}$	0.51	1.31	1.66	[0.78, 3.56]
$\beta_{lapsia3}$	0.49	1.18	1.64	[0.72, 3.70]
$\beta_{lapsia4}$	1.12	2.66	3.08	[1.35, 7.04]*
$\beta_{lapsiaYli4}$	1.09	2.26	2.96	[1.16, 7.57]*
$\beta_{tupakoiEnnen}$	0.26	1.07	1.30	[0.81, 2.08]
$\beta_{tupakoiNyt}$	0.61	2.53	1.84	[1.15, 2.94]*
$\beta_{kouluPerus}$	0.96	2.35	2.60	[1.17, 5.77]*
$\beta_{koulu2aste}$	0.34	1.12	1.40	[0.78, 2.52]
$\beta_{kouluYlempi}$	0.36	1.01	1.43	[0.71, 2.85]
$\beta_{naimaton}$	0.05	0.17	1.05	[0.57, 1.95]
$\beta_{eronnut}$	0.13	0.53	1.14	[0.71, 1.82]
β_{leski}	0.55	1.23	1.73	[0.72, 4.12]
$\beta_{afrikkalaisamerikkalainen}$	0.50	2.22	1.65	[1.06, 2.57]*
$\beta_{kiinalaisamerikkalainen}$	0.10	0.20	1.10	[0.42, 2.91]
$\beta_{japanilaisamerikkalainen}$	-1.75	-1.72	0.17	[0.02, 1.28]
$\beta_{latinalaisamerikkalainen}$	0.06	0.14	1.06	[0.45, 2.53]
β_{bmi}	0.06	4.82	1.06	[1.03, 1.08]*

TAULUKKO 2. Estimaatit mallille, jossa BMI havaittu, *-merkillä tilastollisesti merkitsevät 5 prosentin riskitasolla

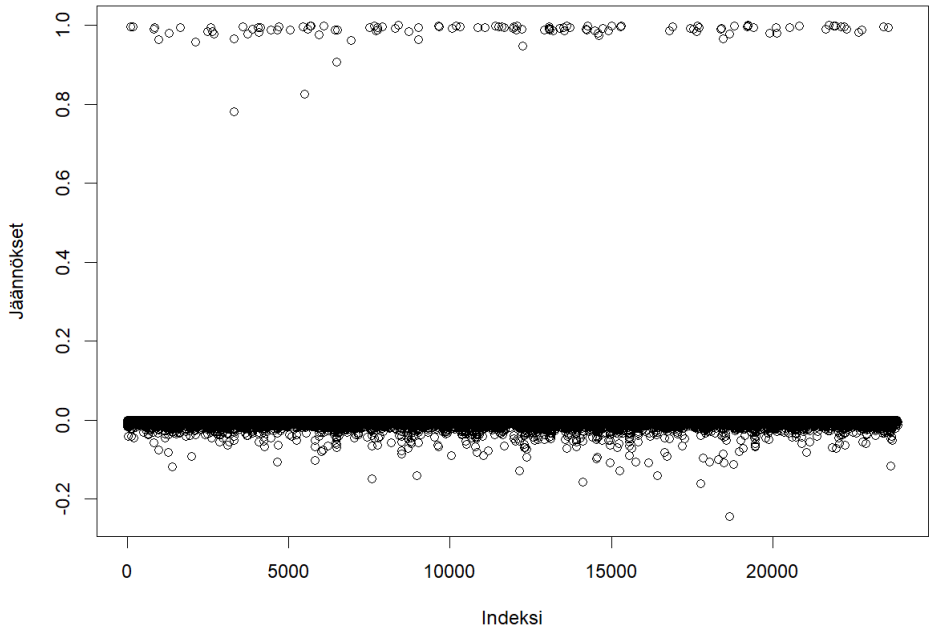
Mallin mukaan β-kertoimista $\beta_{lapsia4}$, $\beta_{lapsiaYli4}$, $\beta_{tupakoiNyt}$, $\beta_{kouluPerus}$, $\beta_{afrikkalaisamerikkalainen}$, $\beta_{japanilaismaerikkalainen}$ ja β_{bmi} ovat merkitseviä.

5.2. Mallin sopivuustarkastelut

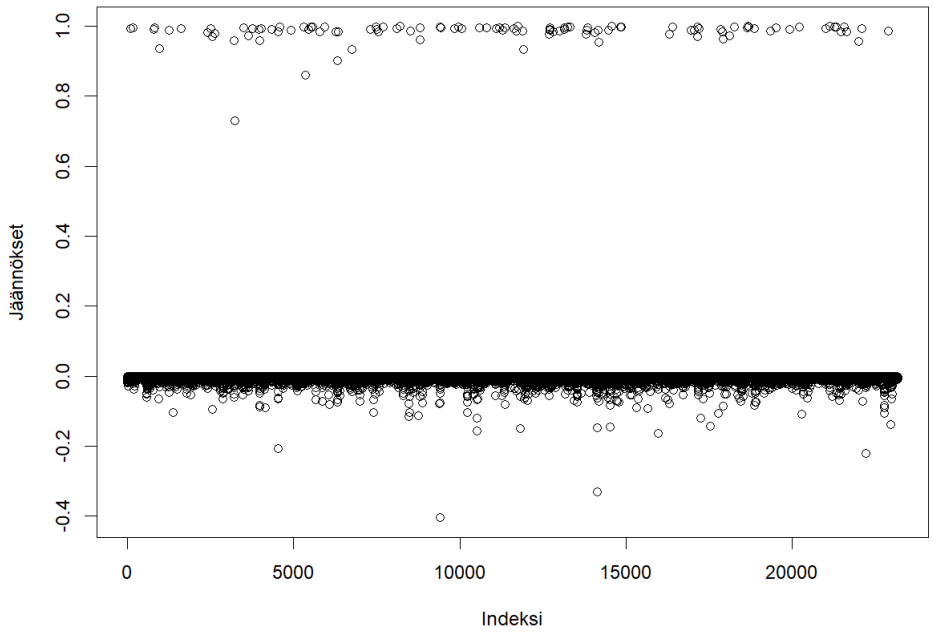
Tässä kappaleessa tarkastellaan sovitettujen mallien sopivuutta aineistoon. Mallit on sovitettu SWAN-aineistoon. Mallien jäännöksiä tutkitaan martingaaliresiduaalien avulla ja verrannollisten uhkien mallin oletuksen voimassaoloa testillä, joka hyödyntää Schoenfeldin residuaaleja.

Martingaaliresiduaalit tulkitaan havaitun tapahtumien lukumäärän ja odotetun tapahtumien lukumäärän erotuksena jokaisella ajanhetkellä t (Terry M. Therneau (1990)). Martingaaliresiduaalien jakauma on vino, sillä residuaalit jakautuvat välille $[-\infty, 1]$ keskittyen kuitenkin nollan ympäristöön. Kuvien 5.1 ja 5.2 perusteella kummassakin sovitetussa mallissa residuaalit keskittyvät voimakkaasti nollan ympärille.

Verrannollisten uhkien mallin sopivuutta testataan Schoenfeldin residuaalien avulla (Patricia M. Grambsch (1994)). Schoenfeldin residuaalit muuttujittain summautuvat nollaan, kun verrannollisten uhkien malli on sopiva. Vastahypoteesina on mahdollisten aikariippuvien β-kertoimien lisääminen malliin. Ainoa merkitsevä testien p-arvo saadaan etnisyydelle mallissa, jossa BMI on latentti (Taulukko 3). Testi voidaan tehdä R:llä survival-paketin cox.zph-funktiolla (T. M. Therneau et al. (2023)).



KUVA 5.1. Martingaaliresiduaalit mallille, jossa BMI latentti



KUVA 5.2. Martingaaliresiduaalit mallille, jossa BMI havaittu

$\beta(t)$-termin lisäys	P-arvot mallille, jossa BMI latentti	P-arvot mallille, jossa BMI havaittu
Ikä	0.413	0.854
Lapsiluku	0.103	0.123
Tupakointi	0.806	0.837
Koulutusaste	0.502	0.409
Siviilisääty	0.672	0.665
Etnisyys	0.034*	0.059
BMI	-	0.534

TAULUKKO 3. Verrannollisten uhkien mallin sopivuustestien p-arvot Schoenfeldin residuaalien avulla, *-merkillä tilastollisesti merkitsevät 5 prosentin riskitasolla

5.3. Kausaalivaikutuksen siirto ERMA-aineistoon

SWAN-aineistoon sovitetulla mallilla estimoidaan sepelvaltimotautitapausten lukumäärä ERMA-aineistossa. Seuranta-ajaksi asetetaan kunkin ERMA-aineiston tutkimushenkilön yksilöllinen seuranta-aika, jolloin on mahdollista verrata ennustettua sepelvaltimotautitapausten lukumäärää todelliseen lukumäärään, joka on tiedossa (kaksi). Siirtokaavan 3.2 mukaan kohdepopulaatiosta tarvitaan tieto iästä, tupakoinnista, koulutusasteesta, siviilisäädystä ja etnisestä taustasta mallilla, jossa BMI on latentti. Mallilla, jossa BMI on havaittu, tarvitaan ennusteeseen siirtokaavan 3.3 mukaan edellisten lisäksi BMI. Ennuste ajatellaan interventiona, jossa lapsiluku on sama jokaiselle kohdepopulaation jäsenelle. Siten ennuste saadaan erikseen kullekin lapsiluvulle. Ennusteen muuttujat ovat ERMA-aineiston seurannan alun mittauksista. Etniseksi taustaksi asetetaan ERMA-aineiston yksilöille valko-ihoinen. Käytetty R-koodi ennusteiden ja uusioluottamusvälien laskemiseksi löytyy liitteestä A.

Taulukossa (4) on ennusteet mallilla, jossa BMI on oletettu latentiksi muuttujaksi. Ennusteen mukaan sepelvaltimotautitapausten lukumäärä ERMA-aineistossa seuranta-ajalla on 2.15–6.58 riippuen lapsiluvusta. 95%:n luottamusvälit vaihtelevat välillä [2.04, 6.87]. Korkein tapausten lukumäärä saadaan lapsiluvulla 4. Ennuste puuttuu neljältä yksilöltä.

Lapsiluku	Odotettu sairastuneiden lukumäärä	Luottamusväli
Ei lapsia	2.15	[2.04, 2.25]
1 lapsi	2.75	[2.62, 2.89]
2 lasta	3.23	[3.08, 3.39]
3 lasta	3.34	[3.17, 3.51]
4 lasta	6.58	[6.29, 6.87]
Yli 4 lasta	5.76	[5.50, 6.05]

TAULUKKO 4. Odotettu sairastuneiden lukumäärä ERMA-aineistossa 95 prosentin luottamusvälillä kullakin lapsiluvulla mallilla, jossa BMI latentti

Taulukossa (5) taas on ennusteet mallilla, jossa BMI on oletettu havaituksi. Sepelvaltimotautitapausten lukumäärä ERMA-aineistossa seuranta-ajalla on 1.50–4.61 riippuen lapsiluvusta. 95%:n luottamusvälit vaihtelevat välillä [1.41, 4.88]. Korkein tapausten lukumäärä saadaan lapsiluvulla 4. Ennuste puuttuu 56 yksilöltä, sillä BMI:ssä on melko paljon puuttuvuutta (9.0%).

Lapsiluku	Odotettu sairastuneiden lukumäärä	Luottamusväli
Ei lapsia	1.50	[1.41, 1.59]
1 lapsi	1.83	[1.71, 1.94]
2 lasta	2.50	[2.35, 2.64]
3 lasta	2.45	[2.31, 2.59]
4 lasta	4.61	[4.33, 4.88]
Yli 4 lasta	4.44	[4.17, 4.71]

TAULUKKO 5. Odotettu sairastuneiden lukumäärä ERMA-aineistossa 95 prosentin luottamusvälillä kullakin lapsiluvulla mallilla, jossa BMI havaittu

LUKU 6

Pohdinta

Lapsiluvun vaikutus sepelvaltimotautiriskiin on siirrettävissä yhdysvaltalaisesta populaatiosta jyväskyläläiseen keski-ikäisillä naisilla tutkielmassa tehtyjen oletusten perusteella. Kun lähdepopulaation aineistoon eli yhdysvaltalaiseen SWAN-aineistoon sovitetulla mallilla ennustettiin sepelvaltimotapaukset kohdepopulaation aineistossa eli jyväskyläläisessä ERMA-aineistossa, saatiin tapauksia 1.50–6.58 (95%:n luottamusvälit välillä [1.41, 6.87]) riippuen lapsiluvusta ja käytetystä mallista. ERMA-aineiston todellinen sepelvaltimotautitapausten lukumäärä on kaksi. Mallin mukaan lapsiluvun kasvaessa riski sairastua sepelvaltimotautiin kasvaa.

Ilmiötä kuvaavan kausaalimallin (Kuva 3.1 ja Kuva 3.2) avulla on kuvattu malliin liittyviä kausaalisuhteita koskevat oletukset, jotka on perusteltu aiemman tiedon ja käytössä olevan aineiston pohjalta. Yhteys, joka olisi aiemman tiedon pohjalta ollut mahdollisesti perusteltua, mutta joka kausaalimallista jäi pois, on painoindeksin vaikutus hedelmällisyyteen. Tämä yhteys olisi aiheittanunt graafiin silmukan, jolloin lapsiluvun vaikutus sepelvaltimotautiin ei olisi ollut identifioituva. Silloin myöskään kausaalivaikutuksen siirto ei olisi ollut mahdollista.

Perinteisesti on ollut tyypillistä nimetä syitä, miksi tutkimustuloksiin on suhtauduttava varauksella. Kausaalivaikutusta siirrettäessä populaatiosta toiseen yleistys kuitenkin sallitaan tehdyin oletuksin populaatioiden samankaltaisuudesta. (Pearl ja Bareinboim (2014)). Valikoitumisgraafissa (Kuva 3.3 ja Kuva 3.4) on nähtävissä oletukset, jotka on tehty liittyen yhdysvaltalaisen ja jyväskyläläisen populaatioiden samankaltaisuuteen. Samankaltaisuutta ei oletettu missään muuttujassa, sillä vaikka sepelvaltimotauti ei saanut omaa valikoitumismuuttujaansa, käytännössä myös sepelvaltimotaudin jakauma populaatioissa on eri, sillä kausaalimallin muut muuttujat vaikuttavat siihen. Populaatioiden samankaltaisuus olisi vaatinut mahdollisesti enemmän perusteluja.

Valikoitumisgraafin perusteella laskettiin do-search -algoritmia hyödyntäen siirtokaava, josta nähtiin, mitä tietoa kummastakin populaatiosta tarvitaan kausaalivaikutuksen estimoimiseksi kohdepopulaatiossa. Siirtokaavan (kaava 3.2 ja kaava 3.3) perusteella tilastolliseen malliin valikoitui lähdepopulaatiosta sepelvaltimotaudin selittäjiksi lapsiluku, ikä, tupakointi, koulutusaste, siviilisääty, etninen tausta ja lisäksi BMI vaihtoehdossa, jossa BMI oletettiin havaituksi. Kohdepopulaatiosta tietoa sepelvaltimotaudista ei tarvittu, ja oletus olikin, että tietoa sepelvaltimotaudista ei ole saatavilla, sillä muutoin kausaalivaikutus kohdepopulaatiossa olisi voitu estimoida suoraan hyödyntäen kohdepopulaation aineistoa ilman tietoja lähdepopulaatiosta. Sen lisäksi tietoa lapsiluvusta ei tarvittu, sillä lapsiluku ajateltiin interventiona, jossa se on sama jokaiselle kohdepopulaation yksilölle. SWAN-aineistoon sovitetulla mallilla

ennustettiin sepelvaltimotautitapaukset ERMA-aineistossa luvussa 5. Sepelvaltimotautitapauksia ilmenee ERMA-aineistossa seuranta-ajalla korkeintaan kuusi (6.58), ja jos BMI oletetaan havaituksi, korkeintaan neljä (4.61).

Mallin antamat ennusteet sepelvaltimotautitapauksille kohdepopulaatiossa ovat lähellä todellista lukumäärää, joten kausaalivaikutuksen siirron voi sanoa onnistuneen hyvin. Tämä voi olla yllättävää, sillä sepelvaltimotauti on melko harvinainen keski-ikäisillä naisilla, vaikka vaihdevuosien aikana riskissä pieni hyppäys tapahtuukin. ERMA-aineistossakin sairaus on harvinainen.

Ennusteen osuminen lähelle oikeaa voi olla yllättävää myös siksi, että moni asia on voinut vaikuttaa tuloksen tarkkuuteen. Ensinnäkin ERMA-aineiston painoindeksissä oli paljon puuttuvuutta, jonka vuoksi se olisi kannattanut imputoida. Tämän sijaan päädyttiin kahteen vaihtoehtoiseen tapaan estimoida kausaalivaikutus kohdepopulaatiossa. Ensimmäisessä kausaalimallissa BMI oletettiin latentiksi muuttujaksi, minkä seurauksena BMI jäi kokonaan pois siirtokaavasta 3.2. Toisessa kausaalimallissa BMI on havaittu, jolloin tieto myös siitä siirtokaavan 3.3 mukaan tarvitaan kummastakin populaatiosta kausaalivaikutuksen estimoimiseksi. Tässä tapauksessa ennusteita ei siis saada niille, keneltä BMI puuttuu. Sitä, miksi siirtokaavat ovat erit, voisi tutkia enemmän. Oikeastaan on mielenkiintoista, miksi BMI tulisi siirtokaavaan missään tilanteessa, sillä kausaalivaikutuksen identifioimiseen sitä ei tarvita, ja sen lisääminen malliin voisi jopa aiheuttaa harhaa. Sen valossa malli, jossa BMI on latentti, voi tuntua jopa luotettavammalta vaihtoehdolta. Jos BMI kuitenkin pidetään mallissa mukana, olisi sen imputoinnin lisäksi voinut lisätä malliin aikariippuvana muuttujana iän tavoin.

Lapsiluvun ja sepelvaltimotaudin yhteyteen liittyvän ilmiön yksinkertaistaminen voi aiheuttaa tuloksiin epätarkkuutta. Ilmiöön liittyvän kausaalimallin muotoilu voi olla vaikeaa, vaikka apuna käyttäisikin asiantuntijatietoa aiheesta. Kausaalimalli saattaa siitä huolimatta yksinkertaistaa ilmiöön liittyvien tekijöiden suhteita. Yksilö, jolla on kaikki tämän tutkielman mallissa määritellyt riskitekijät, voi välttyä tautiin sairastumiselta. Toisaalta yksilö, jolla ei ole yhtäkään tunnettua riskitekijää, voi sairastua. Tämä viittaa siihen, että on olemassa muuttujia, joita ei tämän tutkielman kausaalimallissa ole pystytty esittämään. Tutkielmassa kausaalimalliin on käytännössä jätetty muuttujat, jotka vaikuttavat sekä sepelvaltimotautiin että lapsilukuun (tai hedelmällisyyteen). Siten esimerkiksi kolesteroli, joka on keskeinen tekijä sepelvaltimotaudin synnyssä, on jätetty kausaalimallista pois. Tällaisten muuttujien ottaminen mukaan malliin voisi tarkentaa tulosta, mutta toisaalta niiden pois jättämisen ei pitäisi aiheuttaa harhaa. Liitteessä B on kuvattu kausaalimalli (Kuva B.1), joka ainakin joiltain osin saattaa kuvata ilmiötä kattavammin. Lisäksi jyväskyläläisten etninen tausta oletettiin samaksi, kuin yhdysvaltalaisten valko-ihoisten, mikä on yksinkertaistava oletus.

Muutama asia voi aiheuttaa harhaa tuloksiin. SWAN-aineistossa seurannan keskeyttäneiden poisjäämisen syy ei ole tiedossa. On mahdollista, että osa on jäänyt seurannasta pois kuoleman vuoksi. Ottaen huomioon, että kuolema sepelvaltimotaudin seurauksena sydänkohtaukseen on hyvinkin mahdollinen, tämä tieto olisi ollut tärkeä tämän tutkielman kannalta. Todennäköisyys kuolla sepelvaltimotautiin Yhdysvalloissa voi olla suurempi kuin Suomessa erilaisen terveyhenhuoltojärjestelmän ja sosioekonomisen aseman jakauman takia.

Harhaa tuloksiin voi aiheuttaa myös aikariippuvien muuttujien käyttäminen mallissa siten, että vain niiden lähtöarvo on huomioitu. Mallinnuksessa iän aikariippuvuus huomioitiin, mutta ennusteessa vain yksilöiden ikä seurannan alussa otettiin mukaan. Mallissa, jossa BMI oletettiin havaituksi, huomioitiin vain seurannan alun BMI kummassakin aineistossa, vaikka tieto painoindeksistä eri käynneillä olisi ollut saatavilla ja sen muuttuminen ajassa olisi voitu ottaa huomioon.

Sepelvaltimotaudin puhkeamisen tapahtuma-aika on välisensuroitua. Tapahtuma-ajaksi on tässä tutkielmassa asetettu SWAN-aineistossa tutkimuskäyntipäivä, vaikka oikeasti tapahtuma-aika on käyntien välissä. Myös ikä on SWAN-aineistossa välisensuroitua. Ikä on ilmoitettu kokonaislukuina, jolloin yksilön tarkka ikä on jotain kahden ikävuoden välistä. Tämä voi aiheuttaa harhaa tuloksiin.

Aineistojen koulutustaustat jouduttiin yhtenäistämään, jotta kausaalivaikutuksen siirto onnistuisi, mikä voi aiheuttaa harhaa tuloksiin. ERMA-aineistossa lukio ja ammattikoulu on yhdessä, mutta SWAN-aineistossa erikseen. Siten SWAN-aineiston lukio ja ammattikoulu yhdistettiin, mikä voi olla kyseenalaista, sillä ammattikoulut Yhdysvalloissa yleensä vaativat, että hakija on lukiosta valmistunut (Parker (2022)). Suomessa ammattikouluun voi kuitenkin pyrkiä peruskoulun suorittaneena. Joka tapauksessa Yhdysvalloissa ammattikoulun suorittanut ei ole korkeakoulututkinnon suorittanut, joten sen yhdistäminen lukioon ei ehkä ole tulosten kannalta haitallista.

Lisäksi harhaa voi aiheuttaa se, että lapsiluku on ERMA-tutkimuksessa ja SWAN-tutkimuksessa selvitetty erilaisella kysymyksenasettelulla. ERMA-tutkimuksessa lapsiluku on naisen synnytysten lukumäärä, kun taas SWAN-tutkimuksessa elävinä syntyneiden lasten lukumäärä. Toisin sanoen ERMA-tutkimuksessa lapsilukuun lukeutuvat myös kuolleena syntyneet lapset. Kaksoset ym. monikkoraskaudet lukeutuvat SWAN-aineistossa omiksi luvuikseen, kun taas ERMA-tutkimuksessa on vastaajan tulkinnan varassa, lukeeko hän esimerkiksi kaksossynnytyksen yhdeksi vai kahdeksi synnytykseksi. Edellä esitetyillä tulkintaeroilla voidaan kuitenkin ajatella olevan vain pieni vaikutus tutkielman tuloksiin. Vuonna 2021 kuolleena syntyneitä oli Suomessa vain 0.3% ja monikkosynnytyksiä 1.3% kaikista synnytyksistä (THL, 2021).

Tutkielmassa käytetään Coxin mallia, joka ei välttämättä ole ainoa vaihtoehto mallinnuksessa. Koska aineisto on elinaika-aineiston kaltainen, jokin elinaikamalli on luonteva valinta. Coxin malliin on mahdollista lisätä aikariippuvia kovariaatteja, joten se valikoitui ensisijaikseksi malliksi. Mallin sopivuutta tarkasteltiin martingaaliresiduaalien ja Schoenfeldin residuaaleihin pohjautuvan testin avulla. Koska martingaaliresiduaalien jakauma on vino, niiden tulkinta voi olla haastavaa (Terry M. Therneau (1990)). Tutkielman kahden mallin martingaaliresiudaalit näyttivät kuitenkin keskittyvän pääasiassa nollan ympäristöön (Kuva 5.1 ja Kuva 5.2). Schoenfeldin residuaaleihin pohjautuvassa testissä tarkasteltiin verrannollisten uhkien mallin sopivuutta aineistoon. Testien tulosten mukaan mallissa, jossa BMI oletettiin latentiksi, voisi etnisyyttä vastaavan β-kertoimen lisätä malliin aikariippuvaksi. Muutoin mallit vaikuttivat testin mukaan sopivilta.

Tutkielman tulos on merkityksellinen, sillä tietyin oletuksin tehty kausaalivaikutuksen yleistys populaatiosta toiseen on melko uusi tutkimuskohde, mitä on sovellettu käytäntöön todellisella aineistolla hyvin vähän, jos ollenkaan. Tutkielma osoittaa, että kausaalivaikutuksen siirto voi olla mahdollista, vaikka kysessä olisi monimuotoinen tai kohdepopulaatiossa harvinainen ilmiö, ja vaikka populaatioilla olisi paljonkin

eroavaisuuksia. Lisäksi mallin mukaan näyttää siltä, että lapsiluvun kasvaessa, myös riski sepelvaltimotaudille kasvaa, mikä tuo lisätukea aikaisemmalle tutkimukselle aiheesta (Li et al. (2019), Oliver-Williams et al. (2019)). Koska painoindeksin tiedetään olevan sepelvaltimotaudin keskeinen riskitekijä (Ho et al. (2022)), ja toisaalta tiedetään, että lapsiluvun kasvaessa BMI kasvaa (Iversen, Kesmodel ja Ovesen (2018))), olisi lapsiluvun vaikutus painoindeksiin ollut myös mielenkiintoinen vaihtoehto tutkielman aiheeksi.

LIITE A

```
# Siirtokaava:
R> library(dosearch)
R> graph <- "
+   Hed -> Laps
+   Laps -> Sepel
+   Ika -> Hed
+   Ika -> Laps
+   Ika -> Sepel
+   Ika -> Tup
+   Tup -> Hed
+   Tup -> Sepel
+   Sos -> Laps
+   Sos -> Tup
+   Sos -> Sepel
+   T_1 -> Hed
+   T_2 -> Laps
+   T_3 -> Ika
+   T_4 -> Sos
+   T_5 -> Tup
+ "
R> data <- "
+   p(Sepel, Laps, Ika, Sos, Tup | T_1, T_2, T_3, T_4, T_5)
+   p(Laps, Ika, Sos, Tup)
+ "
R> query <- "p(Sepel | do(Laps))"
R> dosearch(data,query,graph, transportability = "T_1, T_2, T_3, T_4, T_5")
\sum_{Tup,Sos,Ika}\left(p(Tup,Sos,Ika)
p(Sepel|Laps,Tup,Sos,Ika,T_1,T_2,T_3,T_4,T_5)\right)
```

```
#==============================================================================
# Tässä vaiheessa SWAN-aineistossa omissa sarakkeissaan jokaisen käynnin tiedot
# kuluneesta ajasta seurannan alusta, iästä ja siitä, onko sairastunut.
# Tallennetaan omaan muuttujaan käynti, jolloin tauti ilmoitettu 1. kerran
# tai jos tautia ei ole, viimeisin käynti.
#==============================================================================
{
  data <- swan
  data$delta <- NA
  data$time <- NA
```

```
  cad <- c("CAD01","CAD02","CAD03","CAD04","CAD05",
           "CAD06","CAD07","CAD08","CAD09","CAD10")
  for(i in 1:nrow(data)){
    if(sum(!is.na(data[i,cad]))==0){    # FALSE lkm = 0 eli kaikki NA
      data$time[i] <- data$INTDAY00[i]  # baseline intday timeen
      data$delta[i] <- 0                # delta on 0, koska havainto sensuroitu
    }
    else if(length(which(data[i,cad]==1))==0){  # nollarivi
      # viimeisimmän käynnin intday timeen
      data$time[i] <- data[i,tail(which(data[i,]==0),1)+2]
      data$delta[i] <- 0               # delta on 0, koska havainto sensuroitu
    }
    else {
      # löytyy 1, eli voidaan 1. ykkösen intday timeen
      data$time[i] <- data[i,which(data[i,]==1)[1]+2]
      data$delta[i] <- 1  # delta on 1, koska havainto on aito
    }
  }

  swan <- data
}
#==============================================================================
# Mukaan vain ne, kenestä seuranta
swan <- swan[swan$time!=0,] # eli time jotain muuta kuin 0

#==============================================================================
# Muokataan aineisto siten, että iän muuttuminen ajassa tulee huomioitua
#==============================================================================
{
  # Iästä riippuva aineisto:
  dep <- swan

  # Iästä riippumaton aineisto:
  indep <- dep
  merge.indep <-tmerge(data1=indep,
                       data2=indep,
                       id=ID,
                       delta=event(time, delta))

  # Iästä riippuvan aineiston kokoaminen:
  new <- tmerge(data1 = indep, data2 = dep, id = ID, tstop = time)
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY00, AGE00))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY01, AGE01))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY02, AGE02))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY03, AGE03))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY04, AGE04))
```

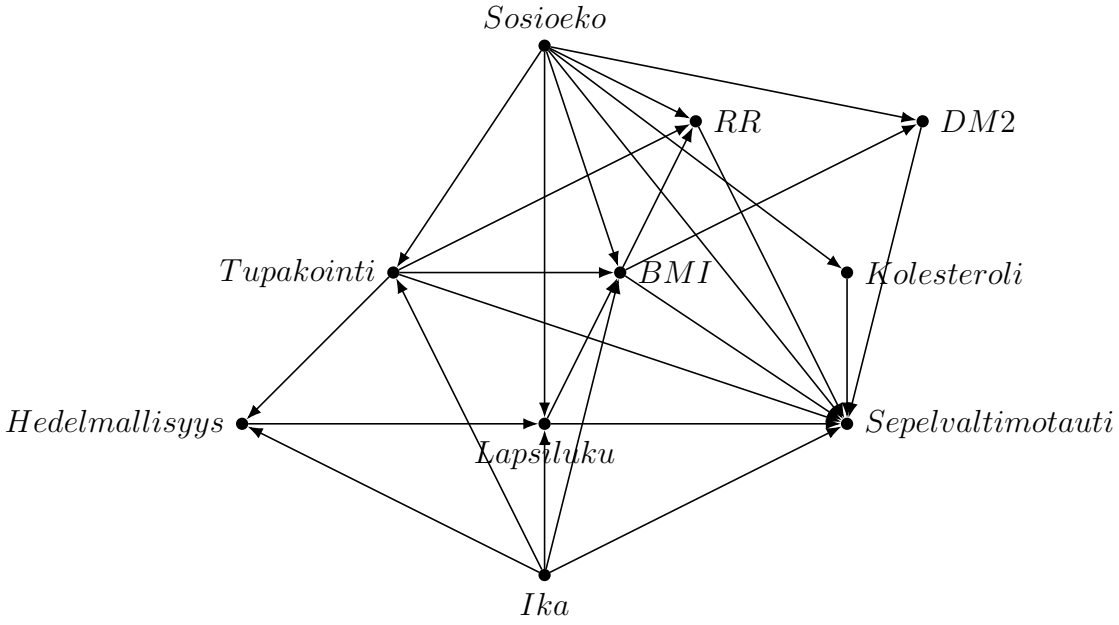
```
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY05, AGE05))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY06, AGE06))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY07, AGE07))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY08, AGE08))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY09, AGE09))
  new <- tmerge(new, dep, id=ID, age = tdc(INTDAY10, AGE10))

  # Iästä riippuvasta aineistosta tarvitaan vain osa tiedoista, jotka liitetään
  # iästä riippumattomaan aineistoon:
  new <- data.frame(ID=new$ID, delta=new$delta, time=new$tstart, age=new$age)

  final <-
    tmerge(data1=merge.indep,
           data2=new,
           id=ID,
           age=tdc(time, age))
}
#########################################################################
# Coxin verrannollisten uhkien malli, jossa aikariippuva ikämuuttuja
#########################################################################
final$Koulutusaste <- relevel(final$Koulutusaste,ref="Korkeakoulututkinto")
final$Etnisyys <- relevel(final$Etnisyys,ref="Kaukasialainen")
final$Siviilisääty <- relevel(final$Siviilisääty,ref ="Avioliitossa/Avoliitossa")

library(survival)
# BMI latentti
fit.cox <- coxph(Surv(tstart, tstop, delta) ~
                   age+Lapsiluku+Tupakointi+Koulutusaste+Siviilisääty+Etnisyys,
                 data = final, id = ID)
# BMI havaittu
fit.bmi <- coxph(Surv(tstart, tstop, delta) ~ age+Lapsiluku+Tupakointi+
                   Koulutusaste+Siviilisääty+Etnisyys+BMI,
                 data = final, id = ID)

#########################################################################
# Ennusteet, Odotettu tapahtumien lkm annetulla aikavälillä
#########################################################################
# predict.coxph()
# expected = "the expected number of events given the covariates and
#             follow-up time"
# "The survival probability for a subject is equal to exp(-expected)."
# tarkemmin kts. ?predict.coxph
#########################################################################

# Tiedot ERMA-datasta:
base <- data.frame(age=erma$Ikä_a,                  # baseline-ikä
                   delta = 1,                       #(*)
```

```
                    tstart=erma$alku,                # 1. käynti
                    tstop=erma$aika,                 # viimeinen käynti
                    Tupakointi=erma$Tupakointi,
                    Koulutusaste=erma$Koulutusaste,
                    Siviilisääty=erma$Siviilisääty,
                    BMI=erma$BMI_itse,               # baseline-bmi
                    Etnisyys=rep("Kaukasialainen", nrow(erma)))

#(*) the newdata argument needs to include both the right
#    and left hand side variables from the formula (doesn't affect the result)

fit <- fit.bmi #tähän malli: bmi latentti / havaittu

{
  # Ei lapsia
  data.0 <- data.frame(base, Lapsiluku="Ei lapsia")
  preds <- predict(fit, data.0, type = "expected", se.fit = T)
  data.0$expected <- preds$fit

  # 1 lapsi
  data.1 <- data.frame(base,Lapsiluku="1 lapsi")
  preds <- predict(fit, data.1, type = "expected", se.fit = T)
  data.1$expected <- preds$fit

  # 2 lasta
  data.2 <- data.frame(base, Lapsiluku="2 lasta")
  preds <- predict(fit, data.2, type = "expected", se.fit = T)
  data.2$expected <- preds$fit

  # 3 lasta
  data.3 <- data.frame(base, Lapsiluku="3 lasta")
  preds <- predict(fit, data.3, type = "expected", se.fit = T)
  data.3$expected <- preds$fit

  # 4 lasta
  data.4 <- data.frame(base, Lapsiluku="4 lasta")
  preds <- predict(fit, data.4, type = "expected", se.fit = T)
  data.4$expected <- preds$fit

  # Yli 4 lasta
  data.yli4 <- data.frame(base, Lapsiluku="Yli 4 lasta")
  preds <- predict(fit, data.yli4, type = "expected", se.fit = T)
  data.yli4$expected <- preds$fit
}

ennusteet <- data.frame(rbind(
```

```
  # 0 lasta
  round(sum(na.omit(data.0$expected)),2),

  # 1 lapsi
  round(sum(na.omit(data.1$expected)),2),

  # 2 lasta
  round(sum(na.omit(data.2$expected)),2),

  # 3 lasta
  round(sum(na.omit(data.3$expected)),2),

  # 4 lasta
  round(sum(na.omit(data.4$expected)),2),

  # yli 4 lasta
  round(sum(na.omit(data.yli4$expected)),2),
))

rownames(ennusteet) <- c("Ei lapsia","1 lapsi","2 lasta",
                         "3 lasta","4 lasta","Yli 4 lasta")


######################
# Uusioluottamusväli
######################
fit <- fit.cox #tähän malli: bmi latentti / havaittu
N <- 1000                                              #1000 otosta
count <- rep(0, N)                                     #ennusteet vektoriin

for (i in 1:N) {
  ind <- sample(589, 589, replace = T)                 #589 palauttaen
  data <- erma[ind,]

  base <- data.frame(age=data$Ikä_a,                    # baseline-ikä
                     delta = 1,                         #(*)
                     tstart=data$alku,                  #1. käynti
                     tstop=data$aika,                   #viimeinen käynti
                     Tupakointi=data$Tupakointi,
                     Koulutusaste=data$Koulutusaste,
                     Siviilisääty=data$Siviilisääty,
                     BMI=data$BMI_itse,                 #baseline-bmi
                     Etnisyys=rep("Kaukasialainen", nrow(data)))

  data.preds <- data.frame(base, Lapsiluku="3 lasta") #vaihda oikea lapsiluku
  preds <- predict(fit, data.preds, type = "expected", se.fit = T)
  data.preds$expected <- preds$fit
```

```
  count[i] <- sum(na.omit(data.preds$expected))       #summataan otoksen ennusteet
}
count <- sort(count)
round(count[round((0.025)*(N + 1),0)],2) #alaraja
round(count[round((0.975)*(N + 1),0)],2) #yläraja
```

LIITE B



KUVA B.1. Eräs vaihtoehtoinen kausaalimalli lapsiluvun vaikutukselle sepelvaltimotaudin riskiin, RR tarkoittaa verenpainetta ja DM2 tyypin 2 diabetesta

Kirjallisuus

Backholer, Kathryn et al. (2016). "Sex differences in the relationship between socioeconomic status and cardiovascular disease: a systematic review and meta-analysis". *Journal of Epidemiology Community Health* 71.6. URL: http://dx.doi.org/10.1136/jech-2016-207890.

CDC (2019). *Racial and ethnic disparities in heart disease*. URL: https://www.cdc.gov/nchs/hus/spotlight/HeartDiseaseSpotlight_2019_0404.pdf.

Cox, D. R. (1972). "Regression models and life-tables". *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, s. 187–202. URL: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

— (1975). "Partial likelihood". *Biometrika* 62.2, s. 269–276. URL: https://doi.org/10.1093/biomet/62.2.269.

ERMA (2019). *Tutkimuksen kuvaus*. URL: https://www.jyu.fi/sport/fi/tutkimus/hankkeet/erma/tutkimuksen_kuvaus.

EsmiRs (2021). *Estrogeeni, mikro-RNA:t ja metabolisten toimintahäiriöiden riski (EsmiRs)*. URL: https://www.jyu.fi/sport/fi/tutkimus/hankkeet/esmirs.

Haapalahti, Petri ja Tomi Mikkola (2015). "Ikääntyvän naisen verisuonten terveys". *Lääketieteellinen Aikakauskirja Duodecim* 131.16. URL: https://www.duodecimlehti.fi/duo12379.

Haukkamaa, Leena et al. (2020). "Risk for subsequent coronary artery disease after preeclampsia". *The American Journal of Cardiology* 93.6. URL: https://doi.org/10.1016/j.amjcard.2003.11.065.

Ho, Frederick K. et al. (2022). "Ethnic differences in cardiovascular risk: examining differential exposure and susceptibility to risk factors". *BioMed Central Med 20* 149.2022. URL: https://doi.org/10.1186/s12916-022-02337-w.

Iversen, Ditte S., Ulrik S. Kesmodel ja Per G. Ovesen (2018). "Associations between parity and maternal BMI in a population-based cohort study". *Acta Obstetricia et Gynecologica Scandinavica* 97.6, s. 694–700. DOI: https://doi.org/10.1111/aogs.13321.

Kettunen, Raimo (2020). *Sydänifarkti ja sydänkohtaus*. URL: https://www.terveyskirjasto.fi/dlk00086/sydaninfarkti-ja-sydankohtaus.

— (2021). *Sepelvaltimotauti*. URL: https://www.terveyskirjasto.fi/dlk00077.

Kiuru, Sirkka, Mika Gissler ja Anna Heino (2019). *Perinataalitilasto – synnyttäjät, synnytykset ja vastasyntyneet 2019*. URL: https://www.julkari.fi/bitstream/handle/10024/140702/Tr48_20.pdf?sequence=5.

Li, Wenchen et al. (2019). "Parity and risk of maternal cardiovascular disease: A dose–response meta-analysis of cohort studies". *European Journal of Preventive Cardiology* 26.6, s. 592–602. URL: https://doi.org/10.1177/2047487318818265.

Lin, D. Y. (2007). "On the Breslow estimator". *Lifetime data analysis* 13.4, s. 471–480. URL: https://doi.org/10.1007/s10985-007-9048-y.

Meadows, Telly A. et al. (2011). "Ethnic differences in cardiovascular risks and mortality in atherothrombotic disease: Insights from the reduction of atherothrombosis for continued health (REACH) Registry". *Mayo Clinic Proceedings* 86.10. URL: doi:%2010.4065/mcp.2011.0010.

Oliver-Williams, Clare et al. (2019). "The association between parity and subsequent cardiovascular disease in women: The atherosclerosis risk in communities study". *Journal of Women's Health* 28.5. URL: https://doi.org/10.1089/jwh.2018.7161.

Parker, Bethanny (2022). *Do Trade Schools Require High School Diplomas?* URL: https://www.bestcolleges.com/trades/do-trade-schools-require-high-school-diplomas/.

Patricia M. Grambsch, Terry M. Therneau (1994). "Proportional hazards tests and diagnostics based on weighted residuals". *Biometrika* 81.3, s. 515–526. URL: https://doi.org/10.1093/biomet/81.3.515.

Pearl, Judea (1995). "Causal Diagrams for Empirical Research". *Biometrika* 82.4, s. 669–688. URL: https://www.jstor.org/stable/2337329.

— (2009). *Causality: Models, Reasoning and Inference*. 2nd edition. New York: Cambridge University Press. URL: https://yzhu.io/courses/core/reading/04.causality.pdf.

Pearl, Judea ja Elias Bareinboim (2014). "External Validity: From Do-Calculus to Transportability Across Populations". *Statistical Science* 29.4, s. 579–595. URL: https://doi.org/10.1214/14-STS486.

Psaltopoulou, Theodora et al. (2017). "Socioeconomic status and risk factors for cardiovascular disease: Impact of dietary mediators". *Hellenic Journal of Cardiology* 58.1. URL: https://doi.org/10.1016/j.hjc.2017.01.022.

Ramachandran, Kandethody M. ja Chris P. Tsokos (2021). *Mathematical Statistics with Applications in R*. 3rd edition. Academic Press. URL: https://doi.org/10.1016/B978-0-12-817815-7.00013-0.

Rössner, Stephan ja Agneta Öhlin (1995). "Pregnancy as a risk factor for obesity: Lessons from the Stockholm Pregnancy and Weight Development Study". *Obesity Research* 3, s. 267–275. DOI: https://doi.org/10.1002/j.1550-8528.1995.tb00473.x.

Study of Women's Health Across the Nation (SWAN) Series (2019). URL: https://www.icpsr.umich.edu/web/ICPSR/series/00253.

Tarkoma, Jari (2018). *Syntyvyys pienentynyt kaikissa koulutusryhmissä*. URL: https://www.stat.fi/til/synt/2017/02/synt_2017_02_2018-12-04_tie_001_fi.html.

Terry M. Therneau Patricia M. Grambsch, Thomas R. Fleming (1990). "Martingale-based residuals for survival models". *Biometrika* 77.1, s. 147–160. URL: https://doi.org/10.1093/biomet/77.1.147.

Therneau, Terry, Cynthia Crowson ja Elizabeth Atkinson (2023). *Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model*. URL: https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf.

Therneau, Terry M et al. (2023). *survival: Survival Analysis*. URL: https://cran.r-project.org/web/packages/survival/index.html.

THL (2021). *Perinataalitilasto 2021 liitetaulukko*. URL: https://www.thl.fi/tilastoliite/tilastoraportit/2022/Liitetaulukot/Perinataalitilasto_2021_ENNAKKO_Liitetaulukot.pdf.

Tiitinen, Aila (2022). *Lapsettomuus*. URL: https://www.terveyskirjasto.fi/dlk00151.

Tikka, Santtu, Antti Hyttinen ja Juha Karvanen (2021a). "Causal Effect Identification from Multiple Incomplete Data Sources: A General Search-Based Approach". *Journal of Statistical Software* 99.5. URL: https://doi.org/10.18637/jss.v099.i05.

— (2021b). *dosearch: Causal Effect Identification from Multiple Incomplete Data Sources*. URL: https://cran.r-project.org/web/packages/dosearch/index.html.

Verma, T.S. (1992). *Graphical Aspects of Causal Models. Technical Report R-191,. UCLA*. URL: http://ftp.cs.ucla.edu/pub/stat_ser/r191.pdf.

WHO (2021). *Obesity and overweight*. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

Wong, Chun Wai et al. (2018). "Marital status and risk of cardiovascular diseases: a systematic review and meta-analysis". *Heart* 104.23. URL: http://dx.doi.org/10.1136/heartjnl-2018-313005.

Yasukawa, Sumiyo et al. (2022). "Super-additive associations between parity and education level on mortality from cardiovascular disease and other causes: the Japan Collaborative Cohort Study". *BMC Women's Health* 22.278. URL: https://doi.org/10.1186/s12905-022-01805-y.