# SUHTEELLISEEN ILMAANTUVUUTEEN PERUSTUVAN KOVARIANSSIRAKENTEEN BAYESILÄINEN MCMC-ESTIMOINTI VALIKOITUNEESSA PERHEAINEISTOSSA

Matti Rantanen

Pro gradu -tutkielma
Kesäkuu 2016

MATEMATIIKAN JA TILASTOTIETEEN LAITOS
TURUN YLIOPISTO

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

RANTANEN, MATTI: Suhteelliseen ilmaantuvuuteen perustuvan kovarianssirakenteen bayesiläinen MCMC-estimointi valikoituneessa perheaineistossa
Pro gradu -tutkielma, 41 s.
Tilastotiede
Kesäkuu 2016

---

Tässä tutkielmassa rintasyövän ja ruoansulatuselinten syöpien perheittäistä kertymistä ja perimäosuutta estimoitiin lapsena tai nuorena syövän sairastaneiden suomalaisten perheaineistoissa. Perheet poimittiin siten, että jokaisessa perheessä on vähintään yksi alle 40-vuotiaana diagnosoitu syöpätapaus vuosina 1970-2012. Rintasyöpäaineisto koostui 4921 perheestä, joissa oli kaikkiaan 26 259 henkilöä. Ruoansulatuselinten aineisto puolestaan koostui 3328 perheestä ja 22 441 henkilöstä.

Syövän perimäosuuden suhteellista ilmaantuvuutta mallinnettiin hierarkkisella bayesiläisellä Poisson-regressio sekamallilla, jossa sairastumisalttiuden vaihtelu jaettiin ympäristön, perimän ja ylihajonnan komponentteihin. Parametrien yhteisposteriorijakaumaa arvioitiin MCMC-otannan avulla JAGS-ohjelmalla. Lisäksi syöpien kertymistä tarkasteltiin estimoimalla sukulaisuussuhteiden mukaan ositettuja suhteellisia syöpäilmaantuvuuksia. Simulaatiotutkimuksella arvioitiin tilastollisen mallin satunnaiskomponenttien estimoituminen ja tarkasteltiin harhan korjauksen vaikutusta tutkimusasetelmaan.

Rintasyöpäaineistossa nuorten syöpäpotilaiden perheenjäsenillä havaittiin 739 syöpää ja perheenjäsenten keskimääräinen syöpäriski oli 81% (95%:n todennäköisyysväli 68-94%) suurempi kuin vastaavalla väestöllä. Rintasyövän perimäosuus oli 26% (0-57%). Ruoansulatuselinten syöpiä havaittiin perheenjäsenillä 574 ja perheenjäsenten syöpäriski oli 60% (48-73%) suurempi kuin väestöllä ja sen perimäosuudeksi estimoitiin 63% (37-88%).

Tutkielman tulosten mukaan ympäristötekijöiden merkitys rintasyöpäaineistossa on suuri. Vastaavasti ruoansulatuselinten syövissä ympäristötekijöiden merkitys on pienempi ja perimän osuus selvästi suurempi.

Asiasanat: perimäosuus, syöpäilmaantuvuus, hierarkkinen Bayes-malli, satunnaiskomponentti, esiinkaivuharha, pro-gradu

# Sisältö

## Tutkielmassa käytetyt merkinnät

| Merkintä | Selitys |
|---:|---|
| $i = 1, \ldots, I$ | perheen indeksi |
| $j = 1, \ldots, J_i$ | henkilön indeksi |
| $n$ | havaintojen lukumäärä |
| $k$ | ikäryhmän indeksi |
| $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ | satunnaismuuttujavektori (vaste) |
| $\boldsymbol{y} = (y_1, \ldots, y_n)'$ | satunnaismuuttujavektorin havaittu arvo |
| $\boldsymbol{e} = (e_1, \ldots, e_n)'$ | odotettujen tapausten lukumäärävektori |
| $\boldsymbol{py} = (py_1, \ldots, py_n)'$ | henkilövuosivektori |
| $\lambda$ | ilmaantuvuus |
| $\boldsymbol{\theta}$ | tilastollisen mallin parametrivektori |
| $p()$ | todennäköisyysjakauma |
| $g()$ | linkkifunktio |
| $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ | regressiokertoimien vektori |
| $\boldsymbol{x} = (x_1, \ldots, x_n)'$ | selittävien tekijöiden vektori |
| $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ | mallimatriisi (selittävät tekijät) |
| $\boldsymbol{u} = (u_1, \ldots, u_m)'$ | satunnaistekijävektori |
| $\boldsymbol{Z}$ | mallimatriisi (satunnaistekijät) |
| $\boldsymbol{R} = diag(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_I)$ | sukulaisuusmatriisi |
| $\boldsymbol{\eta} = \boldsymbol{X\beta}$ | lineaarinen prediktori |
| $\boldsymbol{\Sigma}$ | kovarianssimatriisi |
| $\boldsymbol{I}$ | identiteettimatriisi |
| $\boldsymbol{J}$ | yksikkömatriisi |
| $\sigma$ | keskihajonta |

# 1 Johdanto

Tämän tutkielman tavoitteena on arvioida rintasyövän ja ruoansulatuselinten syöpien perheittäistä kertymistä ja perimäosuutta perheissä, joissa ainakin yhdellä perheenjäsenellä on havaittu nuoruusiän syöpä. Aineisto perustuu kaikkiin Suomessa aikavälillä 1970-2012 havaittuihin nuoriin syöpäpotilaisiin ja heidän lähisukulaisiinsa. Nuorella syöpäpotilaalla tarkoitetaan tässä tutkielmassa alle 40-vuotiaana syöpään sairastunutta henkilöä.

Rintasyöpä on suomalaisten naisten yleisin syöpä. Tämän tutkielman rintasyöpäaineistossa nuoruusiän syöpiä havaittiin 4921 perheessä, joissa on yhteensä 26 259 henkilöä. Ruoansulatuselinten syövät puolestaan ovat Suomen yleisimpiä syöpiä sekä naisilla että miehillä, ja niitä todetaan lähes yhtä paljon molemmilla sukupuolilla. Ruoansulatuselinten syöpien aineistossa on 3328 perhettä, joissa on yhteensä 22 441 henkilöä.

Tutkielmassa käytetty perheaineisto on laajuudeltaan ainutlaatuinen ja kuvaa kattavasti syöpien perheittäistä kasautumista. Tämä on ensimmäinen tutkimus Suomessa, jossa estimoidaan syövän perimäosuutta nuorten syöpäpotilaiden laajennetuissa perheissä. Laajennetulla perheellä tarkoitetaan tässä tutkielmssa perhettä, johon on sisällytetty nuoren syöpäpotilaan lisäksi tämän puoliso, lapset, vanhemmat ja sisarukset.

Tutkielman tavoitteena on arvioida sitä, miten perimä ja ympäristö vaikuttavat riskiin sairastua syöpään. Aineistoon sovelletaan suhteelliseen ilmaantuvuuteen perustuvaa yleistettyä lineaarista sekamallia, jonka parametreja estimoidaan MCMC-otannan keinoin.

Syöpätapaukseen perustuvan poiminnan aiheuttamaa harhaa (*ascertainment bias*) ja harhan korjauksen vaikutusta arvioidaan simulaatiotutkimuksella. Lisäksi simulaatiokokeella selvitetään perimäosuuden estimoinnin luotettavuutta tutkimusaineistossa. Tilastollisen mallin parametrien estimoinnin laskennalliset ongelmat ratkaistaan parametrilaajennuksella sekä moniulotteisen normaalijakauman matriisihajotelmaa hyödyntämällä.

# 2 Bayes-päättely ja yleistetty lineaarinen sekamalli

## 2.1 Bayes-päättely

Olkoon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ satunnaismuuttuja ja vektori $\boldsymbol{y} = (y_1, \ldots, y_n)'$ sen havaittu arvo, missä $n$ on havaintojen lukumäärä. Satunnaismuuttujalle $\boldsymbol{Y}$ voidaan antaa parametrinen todennäköisyysjakauma, jonka määrittelee tuntematon parametri tai parametrivektori $\boldsymbol{\theta}$. Kun satunnaismuuttujan $\boldsymbol{Y}$ arvot havaitaan, saadaan parametrille $\boldsymbol{\theta}$ uskottavuusfunktio

$$p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}). \tag{1}$$

Uskottavuusfunktio on havaitulle aineistolle $\boldsymbol{y}$ määritelty yhteistodennäköisyysjakauma, jossa havainnot oletetaan ehdollisesti riippumattomiksi ja samoin jakautuneiksi.

Bayesiläisessä tilastotieteessä sekä satunnaismuuttuja $\boldsymbol{Y}$ että parametrit $\boldsymbol{\theta}$ ovat satunnaismuuttujia, ja niille määritellään omat todennäköisyysjakaumat. Uskottavuusfunktiolla kuvataan $p(\boldsymbol{y}|\boldsymbol{\theta})$ aineiston generoitumista ja priorijakaumalla $p(\boldsymbol{\theta})$ parametrien $\boldsymbol{\theta}$ todennäköisyysjakaumaa. Priorijakauma kuvaa parametreihin $\boldsymbol{\theta}$ liittyvää epävarmuutta ennen aineiston $\boldsymbol{y}$ havaitsemista.

Tilastollinen päättely kohdistuu posteriorijakaumaan $p(\boldsymbol{\theta}|\boldsymbol{y})$, joka on tuntemattomien parametrien $\boldsymbol{\theta}$ todennäköisyysjakauma ehdolla havaittu aineisto $\boldsymbol{y}$. Parametrien $\boldsymbol{\theta}$ posteriorijakauma voidaan johtaa ehdollisen todennäköisyyden kaavalla eli Bayesin kaavalla

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})}{p(\boldsymbol{y})},$$

jossa $p(\boldsymbol{y}) = \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ on havaintojen $\boldsymbol{y}$ marginaalijakauma. Posteriorijakauma voidaan esittää ilman nimittäjää $p(\boldsymbol{y})$, koska se ei riipu kiinnostavista parametreista $\boldsymbol{\theta}$. Tällöin $p(\boldsymbol{y})$ voidaan ajatella vakiona ja posteriorijakauma on *verrannollinen* uskottavuusfunktion ja priorijakauman yhteisto-

dennäköisyysjakaumaan

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}) = p(\boldsymbol{y}, \boldsymbol{\theta}).$$

(Gelman *et al.*, 2005; Koistinen, 2013)

## 2.2 Yleistetty lineaarinen malli ja sekamalli

Tilastotieteessä regressiomalli vastaa kysymykseen siitä, miten satunnaismuuttujan $\boldsymbol{Y}$ arvot vaihtelevat keskimäärin yhden tai useamman satunnaismuuttujan $\boldsymbol{X}$ eri tasoilla. Mielenkiinnon kohteena olevaa satunnaismuuttujaa $\boldsymbol{Y}$ kutsutaan *vasteeksi* ja sen vaihtelua selittäviä satunnaismuuttujia $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ *kovariaateiksi* tai *selittäviksi muuttujiksi*. Tällöin ollaan kiinnostuneita satunnaismuuttujan $\boldsymbol{Y}$ ehdollisesta jakaumasta $p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X})$. Mallin kovariaatit $\boldsymbol{X}$ yhdessä regressiokertoimien $\boldsymbol{\theta} = \boldsymbol{\beta}$ kanssa muodostavat lineaarisen prediktorin $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. Tällöin lineaarinen regressiomalli voidaan esittää tilastollisena mallina satunnaismuuttujan $\boldsymbol{Y}$ odotusarvolle

$$\mathrm{E}(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

Yleistetyssä lineaarisessa mallissa satunnaismuuttujan $\boldsymbol{Y}$ jakauma kuuluu ns. eksponenttiperheen jakaumiin (McCullagh ja Nelder, 1989). Eksponenttiperheen jakaumat voidaan jakaa tekijöihin, jolloin ne voidaan kirjoittaa yhteneväisesti

$$p(y_i|\boldsymbol{\theta}) = f(y_i)g(\boldsymbol{\theta}) \exp\left\{\phi(\boldsymbol{\theta})' t(y_i)\right\}. \tag{2}$$

Eksponettiperheen tekijöistä nähdään todennäköisyysmallin luonnollinen parametri $\phi(\boldsymbol{\theta})$, jossa $\phi$ on ns. skaalaparametri. Tästä voidaan päätellä monotonisesti kasvava, derivoituva ja kaikkialla määritelty linkkifunktio $g()$, joka kuvaa odotusarvon parametrille $\theta$. Nyt yleistetty lineaarinen malli voidaan esittää odotusarvon funktiona

$$g(\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})) = \boldsymbol{X}\boldsymbol{\beta}, \tag{3}$$

jossa linkkifunktio $g()$ määrittelee vasteen ja kovariaattien välille lineaarisen yhteyden.

Reaalimaailmassa vasteen $Y$ arvot eivät useinkaan ole toisistaan riippumattomia ja yksinkertaiseen ehdolliseen riippumattomuusoletukseen perustuvat tilastolliset mallit (kaava 3) eivät kuvaa aineistoa riittävän hyvin. Silloin on luonnollista kuvata vasteen generoitumista monimutkaisemmalla hierarkkisella tilastollisella mallilla, joka kuvaa aineistoa paremmin. Esimerkkinä havaintojen hierarkkisesta generoitumisesta ovat aineistot, joissa havainnot syntyvät useista eri vaihtelun lähteistä tai pitkittäisaineistot, joissa mittaukset ovat toistoja samasta yksilöstä eri aikapisteissä.

Yleistetyssä lineaarisessa sekamallissa tavanomaista yleistettyä lineaarista mallia on laajennettu yhdellä tai useammalla satunnaistekijällä:

$$g(\mathrm{E}(\boldsymbol{y})) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}. \tag{4}$$

$\boldsymbol{Z}$ on $n \times m$ matriisi, joka määrittelee havaintojen välisen riippuvuusrakenteen, missä $m$ on ryhmien lukumäärä. Tässä ryhmällä tarkoitetaan esimerkiksi samasta perheestä poimittuja havaintoja tai saman yksilön mittauksia eri aikapisteissä. Satunnaistekijä $\boldsymbol{u} = (u_1, \ldots, u_m)'$ oletetaan normaalijakautuneeksi odotusarvolla 0 ja hajonnalla $\sigma_u$. Jakauman odotusarvosta ei tehdä päättelyä, mutta sen selittämän vaihtelun $\sigma_u$ suuruus on mielenkiinnon kohteena. (Demidenko, 2005)

## 2.3 Poisson-malli

Tässä tutkielmassa sovelletaan lukumääriä sisältävän aineiston tilastollista mallia eli Poisson-mallia. Oletetaan, että satunnaismuuttuja $\boldsymbol{Y}$ on diskreetti lukumäärää kuvaava satunnaismuuttuja, joka saa arvoja $y = 0, 1, 2, \ldots$. Satunnaismuuttuja $\boldsymbol{Y}$ on Poisson-jakautunut parametrilla $\lambda_i$ eli $Y_i \sim \text{Poisson}(\lambda_i)$. Tällöin havainnon $y_i$ todennäköisyysjakauma on

$$p(y_i|\lambda_i) = \frac{\lambda_i^{y_i}}{y_i!} \exp\{-\lambda_i\}, \tag{5}$$

jossa mallin parametri $\lambda_i > 0$ on keskimääräinen tapauksien lukumäärä. Poisson-mallin erityspiirteenä on, että sen odotusarvo ja varianssi ovat yhtäsuuret, $\mathrm{E}(Y_i) = \mathrm{V}(Y_i) = \lambda_i$.

Havainnon $y_i$ todennäköisyysjakauma voidaan esittää eksponenttiperheen muodossa

$$p(y_i|\lambda_i) \propto \exp\{y_i \log(\lambda_i)\} \exp\{-\lambda_i\},$$

josta on jätetty pois $\lambda$:sta riippumattomat tekijät. Eksponenttiperheen muodosta voidaan päätellä mallin linkkifunktioksi log-linkki, jolloin tilastollinen malli voidaan kirjoittaa

$$\mathrm{E}(Y_i) = \lambda_i = \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}.$$

Poisson-mallin tiheys $\lambda_i$ voidaan suhteuttaa tunnettuun yksilökohtaiseen altisteaikaan $e_i$. Sitä käsitellään tunnettuna vakiona ja sitä voidaan käyttää esimerkiksi henkilövuosien määrittelyyn hasardimalleissa tai odotettujen tapausten määrittelyyn suhteellisen ilmaantuvuuden malleissa. Jatkossa $e_i$ tarkoittaa henkilön $i$ odotettujen tapausten lukumäärää. Nyt Poisson-malli määritellään seuraavasti:

$$Y_i \sim \mathrm{Poisson}(\lambda_i e_i), \tag{6}$$

jossa $\mathrm{E}(Y_i) = e_i \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}$. Uskottavuusfunktio riippumattomille havaintopareille $\{(y_i, e_i);\ i = 1, \ldots, n\}$ on tulo

$$p(\boldsymbol{y}|\lambda_i; \boldsymbol{e}) = \prod_{i=1}^{n} \frac{(\lambda_i e_i)^{y_i}}{y_i!} \exp\{-\lambda_i e_i\}. \tag{7}$$

(Dobson ja Barnett, 2008; Gelman *et al.*, 2005)

## 2.4 Bayes-päättely yleistetyssä lineaarisessa sekamallissa

Vakioitu ilmaantuvuussuhde (SIR – *Standardised Incidence Ratio*) on erityisesti syöpäepidemiologisissa tutkimuksissa usein käytetty tunnusluku, jossa tutkimuskohortin ilmaantuvuutta verrataan taustaväestön ilmaantuvuuteen. SIR:n estimaatti on havaittujen $\sum_{i=1}^{n} y_i$ ja odotettujen $\sum_{i=1}^{n} e_i$ tapauksien lukumäärien suhde. Odotetut tapaukset lasketaan taulukoimalla aineiston henkilövuodet sukupuolen, ikäryhmän ja kalenteriperiodin mukaisiin osiitteisiin, jotka kerrotaan taustaväestön vastaavien osiitteiden ilmaantuvuudella (ks. kappale 5.3). (Breslow *et al.*, 1987)

Suhteellisen ilmaantuvuuden tilastollinen malli havainnoille $\boldsymbol{y}$ voidaan esittää Poisson-regressiona seuraavasti:

$$\log(\mathrm{E}(y_i|\beta, e_i)) = \log(e_i) + \beta, \tag{8}$$

jossa lineaarinen prediktori on vakio $\beta = \boldsymbol{x}_i'\boldsymbol{\beta}$ ja $Y_i \sim \mathrm{Poisson}(e_i \exp\{\beta\})$. Estimaatti suhteelliselle ilmaantuvuudelle on nyt $\widehat{\mathrm{SIR}} = \exp\{\beta\}$. Kun $\widehat{\mathrm{SIR}}$ on suurempi kuin 1, on ilmaantuvuus korkeampaa kuin vastaavalla väestöllä, joka on samaa sukupuolta, samanikäistä ja vastaavassa kalenteriperiodissa kuin tutkimuskohortti. (Breslow *et al.*, 1987)

Bayes-päättelyssä jokaiselle eksponenttiperheen jakaumalle (kaava 2) voidaan johtaa sellainen priorijakauma, joka on samaa muotoa kuin mallin posteriorijakauma. Tällaista jakaumaa kutsutaan *konjugaattijakaumaksi*. Poisson-jakautuneen satunnaismuuttujan konjugaattijakauma on Gamma-jakauma. Konjugaattijakauman käyttö yksinkertaistaa tilastollista laskentaa ja päättelyä, jos sen käyttö on mielekästä. (Gelman *et al.*, 2005)

Mallin 8 suhteellinen ilmaantuvuus voidaan estimoida posteriorijakaumasta $p(\beta|\boldsymbol{y}; \boldsymbol{e})$. Kun suhteelliselle ilmaantuvuudelle $\beta$ määritellään priorijakaumaksi Gamma-jakauma, voidaan posteriorijakauma ratkaista suoravii-

vaisesti:

$$
\begin{aligned}
p(\beta|\boldsymbol{y};\boldsymbol{e}) &\propto p(\boldsymbol{y}|\beta;\boldsymbol{e})p(\beta) \\
&= \prod_{i=1}^{n}\left(\frac{(\beta e_i)^{y_i}}{y_i!}\exp\{-\beta e_i\}\right)\frac{b^a}{\Gamma(a)}\beta^{a-1}\exp\{-b\beta\} \\
&\propto \exp\left\{-\beta\left(\sum_{i=1}^{n}e_i+b\right)\right\}\beta^{-\left(\sum_{i=1}^{n}y_i+a\right)-1},
\end{aligned}
\tag{9}
$$

joka on vakiota vaille Gamma-jakauman tiheys. Tästä seuraa, että $\beta$:n jakaumaksi saadaan

$$
\beta|\boldsymbol{y};\boldsymbol{e} \sim \text{Gamma}\left(a+\sum_{i=1}^{n}y_i,\ b+\sum_{i=1}^{n}e_i\right). \tag{10}
$$

Priorijakauman *hyperparametrien* $a$ ja $b$ avulla kuvataan parametriin $\beta$ liittyvä ennakkokäsitys. Posteriorijakauma $\beta$:lle saadaan sijoittamalla havaitut arvot $\boldsymbol{y}$ ja $\boldsymbol{e}$ kaavaan 10, jolloin sen odotusarvo voidaan ratkaista joko analyyttisesti tai siitä voidaan poimia otoksia tilasto-ohjelmien todennäköisyysjakaumafunktioilla.

Poisson-regressiomallin oletus varianssin ja odotusarvon yhtäsuuruudesta pätee harvoin reaalimaailman hierarkkisessa aineistossa, jossa varianssi on lähes poikkeuksetta odotusarvoa suurempi. Tätä ilmiötä kutsutaan Poisson-ylihajonnaksi. Yksinkertaisin tapa käsitellä aineiston ylihajontaa tilastollisessa mallissa on lisätä jokaiselle havainnolle yksilötason satunnaistekijä eli ylihajontatermi $OD_i$, jolla sallitaan poikkeama keskimääräisestä odotusarvosta. (Dobson ja Barnett, 2008; Gelman ja Hill, 2006)

Yksilötason ylihajontatermi $OD$ voidaan lisätä mallin 6 lineaariseen prediktoriin

$$
\begin{aligned}
y_i &\sim \text{Poisson}(e_i \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + OD_i\}) \\
OD_i &\sim \text{Normal}(0, \sigma_{od}).
\end{aligned}
\tag{11}
$$

Mallin posteriorijakauma voidaan esittää muodossa

$$p(\beta, \sigma_{od}|\boldsymbol{y}; \boldsymbol{e}) \propto \prod_{i=1}^{n} \Big( p(y_i|\beta, OD_i; e_i) p(OD_i|\sigma_{od}) \Big) p(\sigma_{od}) p(\beta).$$

Monimutkaisemmissa malleissa otosjakauman ja priorijakauman muodostama posteriorijakauma $p(\boldsymbol{\theta}|\boldsymbol{y})$ on harvoin suljetussa muodossa tunnettu todennäköisyysjakauma. Tällöin posteriorijakaumaa voidaan arvioida poimimalla siitä riittävä määrä otoksia laskennallisilla menetelmillä. Tämä mahdollistaa tilastollisen päättelyn tekemisen monimutkaisistakin posteriorijakaumista (Gelman *et al.*, 2005). Jatkossa tässä tutkielmassa termillä posteriorijakauma tarkoitetaan otoksista muodostettua jakaumaa.

# 3 Perheaineistojen tilastollinen malli

## 3.1 Perinnöllisyyden käsitteet ja kovarianssirakenteet

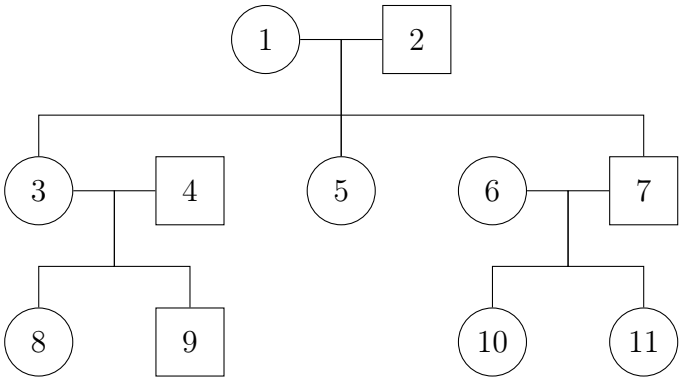
Ihmisen DNA sijaitsee 22:ssa autosomikromosomissa sekä kahdessa sukupuolikromosomissa X ja Y. Jokaista autosomia on kaksi kappaletta, yksi kummaltakin vanhemmalta. Kromosomit sisältävät geenejä, jotka sisältävät perimän informaation. Alleelit ovat kromosomissa tietyn geenin vaihtoehtoisia muotoja, ja ne muodostavat autosomissa alleeliparin. Alleeli on periytyvyyden perusyksikkö, jonka lähtökohtana on, että erilaiset ominaisuudet periytyvät jälkeläisille niiden kautta vaikuttaen siten henkilön syöpäriskiin.

Gregor Mendel kehitti vuonna 1865 periytyvyyden käsitteen, joka tunnetaan *mendeliaanisena periytyvyytenä*. Hän ehdotti, että jokin havaitsematon tekijä (alleeli) aiheuttaa sen, että ominaisuus periytyy muuttumattomana vanhemmalta jälkeläiselle. Mendeliaanista periytyvyyttä on sovellettu onnistuneesti sairauksien periytyvyyden kuvaamisessa perheaineistoissa. Kuten usein laajaan väestödataan perustuvissa tutkimuksissa, ei tässä tutkielmassa ole käytössä genomidataa, joten periytyvyyttä estimoidaan Mendelin periaatteiden mukaisesti. (Khoury *et al.*, 1993)

Perheaineistoissa syövän geneettistä alttiutta tarkasteltaessa on syytä ensin tutkia, havaitaanko perheissä enemmän syöpää kuin olisi odotettavissa,

mikäli perheen yksilöt olisivat täysin riippumattomia. Erään perheen riippuvuusrakennetta on havainnollistettu kuvassa 1. Jos perheittäistä kertymistä havaitaan, on mielekästä ajatella kertymisen johtuvan perimästä tai ympäristötekijöistä. Syövän perheittäistä kertymistä aineistossa voidaan arvioida vertaamalla perheenjäsenten syöpäilmaantuvuutta väestöön esimerkiksi edellä kuvatulla vakioidulla ilmaantuvuussuhteella (SIR).

Perheaineistoissa perimän ja ympäristön osuutta syöpäilmaantuvuuden vaihtelussa voidaan eritellä tarkemman tilastollisten mallin avulla, vaikka varsinaista genomitietoa ei olisi saatavilla. Tilastollisessa mallissa perimä määräytyy havaitun sukulaisuussuhteen perusteella ja kaikkien perheenjäsenten oletetaan jakavan yhteiset ympäristötekijät. Perimän ja ympäristötekijöiden lisäksi malliin voidaan lisätä yksilökohtainen Poisson-ylihajonta (kaava 11), joka selittää aineistossa esiintyvää heterogeenisyyttä. (Thomas *et al.*, 2004; Khoury *et al.*, 1993)




Kuva 1: Simuloidun aineiston perherakenne. Ympyrä tarkoittaa naista ja neliö miestä. Perheen perustajajäsenet 1 ja 2 ovat sukupuun ylimpänä. Heidän jälkeläisään ovat 3, 5 ja 7 ja lasten lapsiaan numerot 8-11. Jälkeläisten (3 ja 7) puolisot 4 ja 6 ovat myös perheen perustajia, koska heille ei ole määritelty vanhempia.

### Sukulaisuusmatriisi $\boldsymbol{R}$

Sukulaisuussuhteiden määrittely on tärkeää toimivan tilastollisen mallin muodostamiseksi. Kuvassa 1 on esitetty sukupuuna esimerkki havaitusta perhe-

rakenteesta. Mendeliaaninen periytyvyys määritellään todennäköisyydeksi, jossa kaksi satunnaista yksilöä jakavat saman alleelin. Todennäköisyys on sitä suurempi mitä läheisempää sukua kaksi yksilöä ovat, kun taas puolisoiden välillä todennäköisyys on nolla. Kun perheen sukulaisuussuhteet on havaittu, perheen $i$ perimän jakamisen todennäköisyyttä voidaan kuvata sukulaisuusmatriisilla $\boldsymbol{R}_i$. Se on symmetrinen kovarianssimatriisi ja voidaan esittää koko aineistolle lohkodiagonaalimatriisina

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & & & \\ & \boldsymbol{R}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{R}_I \end{bmatrix},$$

jossa perheet oletetaan toisistaan riippumattomiksi. Perheenjäsenten hierarkkia määritellään vanhempien perusteella. Jokaisella henkilöllä pitää olla määriteltynä joko molemmat tai ei kumpikaan vanhemmista. Henkilöitä, joiden vanhempia ei ole määritelty, kutsutaan *perustajajäseniksi*.

Matriisin $\boldsymbol{R}_i$ alkio $r_{mn}$ on henkilöiden $s = 1, \ldots, J_i$ ja $d = 1, \ldots, J_i$ välinen geneettinen etäisyys eli todennäköisyys, että he jakavat saman alleelin. Oletetaan myös, että perhe on järjestetty vanhemmista jälkeläisiin. Yhden perheen sukulaisuusmatriisi $\boldsymbol{R}_i$ voidaan muodostaa seuraavalla rekursiivisella algoritmilla (Lange, 2003):

$$r_{sd} = r_{ds} = \begin{cases} 0 & s \text{ ja } d \text{ ovat perustajajäseniä} \\ 1/2 & s = d, s \text{ on perustajajäsen} \\ 1/2 \times (r_{\text{äiti}(s)d} + r_{\text{isä}(s)d}) & s \neq d \\ 1/2 \times (1 + r_{\text{äiti}(s)\text{isä}(s))} & s = d. \end{cases}$$

Lopuksi matriisi skaalataan kertomalla alkiot $r_{sd}$ kahdella, jotta matriisin diagonaalille saadaan luku 1.

Matriisi $\boldsymbol{R}_i$ voidaan ratkaista jokaiselle perheelle erikseen tilastollisella ohjelmistolla R käyttämällä perheaineostojen käsittelyyn tarkoitettua pakettia `kinship2` (Therneau *et al.*, 2012). Kuvan 1 perheen sukulaisuusmatriisiksi

$\boldsymbol{R}_i$ saadaan

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & & & & & & & & & & \\ 0 & 1 & & & & & & & & & \\ 1/2 & 1/2 & 1 & & & & & & & & \\ 1/2 & 1/2 & 1/2 & 1 & & & & & & & \\ 1/2 & 1/2 & 1/2 & 1/2 & 1 & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & & & & \\ 1/4 & 1/4 & 1/2 & 1/4 & 1/4 & 1/2 & 0 & 1 & & & \\ 1/4 & 1/4 & 1/2 & 1/4 & 1/4 & 1/2 & 0 & 1/2 & 1 & & \\ 1/4 & 1/4 & 1/4 & 1/4 & 1/2 & 0 & 1/2 & 1/8 & 1/8 & 1 & \\ 1/4 & 1/4 & 1/4 & 1/4 & 1/2 & 0 & 1/2 & 1/8 & 1/8 & 1/2 & 1 \end{bmatrix}.$$

## 3.2 Perheaineiston hierarkkinen Bayes-malli

Perheaineisto voidaan nyt esittää havaittujen syöpätapausten ja odotettujen tapausten havaintopareina $\{(y_{ij}, e_{ij});\ i = 1, \ldots, I;\ j = 1, \ldots, J_i\}$, jossa $j$ on henkilön indeksi perheessä $i$. Mallin estimointiin sovelletaan yleistettyä lineaarista sekamallia (kaava 4), jossa kokonaisvaihtelu on jaettu kolmeen additiiviseen satunnaiseen komponenttiin: perimään $G_{ij}$, ympäristöön $F_i$, ja Poisson-ylihajontaan $OD_{ij}$. Nyt tilastollinen malli perheaineistolle voidaan kirjoittaa seuraavasti:

$$\begin{aligned} y_{ij} &\sim \text{Poisson}(\lambda_{ij} e_{ij}) \\ \log(\lambda_{ij}) = \boldsymbol{\eta}_{ij} &= \log(e_{ij}) + \beta + G_{ij} + F_i + OD_{ij} \\ G_{ij} &\sim \text{MVNormal}(\mathbf{0}, \sigma_g \boldsymbol{R}) \\ F_i &\sim \text{Normal}(0, \sigma_f) \\ OD_{ij} &\sim \text{Normal}(0, \sigma_{od}). \end{aligned} \tag{12}$$

Lineaarisessa prediktorissa on vain yksi kiinteä tekijä $\beta$, jonka eksponentti

$\exp\{\beta\}$ kuvaa jäljelle jäävää suhteellista ilmaantuvuutta (SIR), kun aineistosta on selitetty ympäristötekijöiden ja perimän aiheuttama kasautuminen. Koska satunnaistekijät oletetaan additiivisiksi ja toisistaan riippumattomiksi, vasteen kokonaisvaihteluksi saadaan

$$\log(\mathrm{V}(y_{ij})) = \sigma_g^2 + \sigma_f^2 + \sigma_{od}^2.$$

Mallin 12 parametreille määritellään hierarkkisessa Bayes-mallissa epäinformatiiviset priorijakaumat $p(\beta)$, $p(\sigma_f)$, $p(\sigma_g)$ ja $p(\sigma_{od})$

$$\begin{aligned} \beta &\sim \mathrm{Normal}(0, 1000) \\ \sigma_g,\ \sigma_f,\ \sigma_{od} &\sim \mathrm{Gamma}(0.001,\ 0.001). \end{aligned} \tag{13}$$

Perheen $i$ havaintojen $\boldsymbol{y}_i$ kovarianssirakenne voidaan kirjoittaa muotoon

$$\mathrm{COV}(\boldsymbol{y}_i, \boldsymbol{y}_i) = \sigma_g \boldsymbol{R}_i + \sigma_f \boldsymbol{J}_{J_i} + \sigma_{od} \boldsymbol{I}_{J_i},$$

jossa $\boldsymbol{J}_{J_i}$ on $J_i \times J_i$ kokoinen matriisi, jonka kaikki alkiot ovat ykkösiä, ja matriisi $\boldsymbol{I}_{J_i}$ on $J_i$ kokoinen identiteettimatriisi. Perimän $\sigma_g$ kovarinassirakenne kiinnitetään geneettisellä etäisyydellä $\boldsymbol{R}_i$. Ympäristöllä $\sigma_f$ on tasakorrelaatiorakenne, jossa jokainen perheenjäsen jakaa yhtä suuren ympäristön vaikutuksen. Lopuksi $\sigma_{od}$ selittää yksilökohtaista Poisson-ylihajontaa.

Perimän ja ympäristön satunnaiskomponentit eivät identifioidu, jos perheenjäsenten geneettiset etäisyydet ovat yhtä suuria eli sukulaisuusmatriisi $\boldsymbol{R}_i$ on vakiomatriisi $\boldsymbol{J}_{J_i}$. Näin tapahtuu esimerkiksi silloin, kun perhe koostuu vain yhdestä sukupolvesta. Usein satunnaiskomponenttien identifioituminen vaatiikin aineiston, jossa on vähintään kaksi sukupolvea. Poikkeuksena ovat kaksostutkimukset ja adoptiotutkimukset.

Perimäosuus $h^2$ syövän suhteellisen ilmaantuvuuden kokonaisvaihtelusta saadaan kaavalla

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_f^2 + \sigma_{od}^2}. \tag{14}$$

Perimäosuus on suhde, jonka arvoalue sijoittuu välille $[0, 1]$. Lähelle lukua 0 estimoitu arvo tarkoittaa, että perimän selittämä osuus syövän suhteellisen

ilmaantuvuuden vaihtelusta on pieni. Jos arvo estimoidaan lähelle lukua 1, on vaihtelu lähes kokonaan geneettistä.

Nyt estimaatit mallin 12 kiinnostaville parametreille $\beta$, $\sigma_f$, $\sigma_g$ ja $\sigma_{od}$ saadaan poimimalla otoksia niiden yhteisposteriorijakaumasta

$$\begin{aligned} p(\beta, \sigma_f, \sigma_g, \sigma_{od}|\boldsymbol{y}; \boldsymbol{e}) \propto \prod_{i=1}^{n} \prod_{j=1}^{J_i} \big(p(y_{ij}|\beta, G_{ij}, F_i, OD_{ij}; e_{ij}) \times \\ p(G_{ij}|\sigma_g) p(F_i|\sigma_f) p(OD_{ij}|\sigma_{od})\big) \times \\ p(\beta) p(\sigma_g) p(\sigma_f) p(\sigma_{od}). \end{aligned} \tag{15}$$

# 4 Hierarkkisen Bayes-mallin estimointi

Hierarkkisten sekamallien haasteena on regressio- ja satunnaiskomponenttien parametrien estimointi. Tämä voidaan kuitenkin toteuttaa Bayes-estimoinnilla, jossa hyödynnetään prioritietoa kaikkien parametrien estimoinnissa. Tässä tutkielmassa käytetään simulointiin perustuvaa MCMC-otantaa, sillä mallin posteriorijakaumaa ei voida ratkaista suljetussa muodossa. Estimointi tehdään JAGS-ohjelmalla (Plummer *et al.*, 2003). Simulointiin perustuvassa lähestymistavassa mielekästä on, että jokaista parametrin simulointiketjua voidaan tarkastella yksitellen. Tämän seurauksena myös hajontaparametrien luotettavuutta voidaan mitata. (Gelman ja Hill, 2006)

## 4.1 Markov Chain Monte Carlo -otanta

Bayesiläisen tilastotieteen estimointimenetelmistä suurin osa perustuu otosten poimintaan posteriorijakaumasta. Markov Chain Monte Carlo -otannassa eli MCMC-otannassa poimitaan kiinnostavien parametrien $\boldsymbol{\theta}$:n likimääräisestä jakaumasta arvoja, joita tarkennetaan erilaisin menetelmin. Markovin ketju on satunnaislukujen järjestetty sarja $\theta^1$, $\theta^2$, $\theta^3$, ..., jossa jokaiselle sarjan alkion indeksille $t = 1, 2, 3, \ldots$ pätee, että satunnaisluku $\theta^t$ riippuu vain ja ainoastaan sitä edeltävästä arvosta $\theta^{t-1}$. Jatkossa tällaista Markovin ketjua kutsutaan yksikertaisesti tässä tutkielmassa ketjuksi. (Gelman *et al.*, 2005)

Vaikka yhteisposteriorijakauma ei olisi tunnettu todennäköisyysjakauma,

voidaan sitä arvioida tietyillä parametrien arvoilla. MCMC-simuloinnissa yhteisposteriorijakauman parametriavaruutta ei käydä läpi systemaattisesti vaan ketju poimii arvoja satunnaisesti.

Sovelluksesta riippuen ketjut konvergoituvat eli tarkentuvat posteriorijakaumaan vasta useiden iteraatioiden jälkeen. Ketjun konvergenssilla tarkoitetaan sitä, kun ketju saavuttaa parametrin odotusarvon ja otokset muodostavat stationäärisen sarjan. Stationäärinen ketju on edellytys estimoitavista parametreista tehtävälle tilastolliselle päättelylle. Posteriorijakaumasta poimitaan riittävä lukumäärä otoksia, jotta ketju kattaa sen koko todennäköisyysalueen. Ketjun alun iteraatiot jätetään pois, jolloin varsinaiseen parametrien estimointiin sisällytetään vain konvergoitunut osa.

Konvergenssia voidaan arvioida usean riippumattoman ja eri alkuarvoilla aloitetun ketjun aikasarjakuvaajasta. Jos ketjut tarkentuvat samaan jakaumaan, voidaan vakuuttua, että yksihuippuinen posteriorijakauma on löytynyt. Tässä tutkielmassa hyödynnetään ketjujen välisen ja sisäisen vaihtelun suhteeseen perustuvaa Gelman-Rubin -tunnuslukua $\hat{R}$, joka saa arvoja luvusta 1 ylöspäin. Arvo 1 indikoi erinomaista konvergessia ja ohjesääntönä on, että $\hat{R}$ tulee olla korkeintaan 1.1 kaikille parametreille. (Gelman ja Rubin, 1992)

Usein ketjun peräkkäiset arvot riippuvat toisistaan. Tällöin ketjut sekoittuvat hitaasti ja autokorrelaatio MCMC-otannassa on suuri. Autokorrelaation arvo on ihanteellisissa tapauksissa alle 0.1, jolloin ketjun peräkkäiset otokset ovat lähes riippumattomia. Ketjusta voidaan tehdä tilastollista päättelyä vaikka autokorrelaatio olisi suurta. Sitä voidaan pienentää joko mallin uudelleen parametroinnilla tai ketjua *harventamalla*, jossa siihen sisällytetään vain ketjun joka $l$:s arvo. Käytännössä ketjen harventaminen tehdään talletustilan säästämiseksi. Käytäntönä on valita sellainen kokonaisluku $l$, että ketjussa olisi vähintään 100 riippumatonta otosta, jolloin voidaan tehdä mielekästä päättelyä posteriorijakauman sijainnista ja vaihtelusta. Riippumattomien otoksien määrää arvioidaan efektiivisellä otoskokoolla $n_{\text{eff}}$. (Gelman *et al.*, 2005)

Parametrin posteriorijakaumaa voidaan approksimoida normaalijakaumalla. Sen odotusarvoa arvioidaan keskiarvolla ja sen vaihtelua otoskeskiha-

jonnalla. Poikkeuksellisesti satunnaiskomponenttien hajontaparametrin estimaattorina käytetään mediaania, koska sen posteriorijakauma voi olla vino. Posteriorijakaumien hajonnan tunnuslukuna on 95%:n posterioritodennäköisyysväli. Todennäköisyysvälillä tarkoitetaan jakauman korkeinta posterioritiheyttä (*Highest Posterior Density*) eli jakauman aluetta, joka sisältää $100(1-\alpha)\%$ posterioritodennäköisyydestä. Todennäköisyysvälit tarkoittavat nimensä mukaisesti 95% todennäköisyyttä sille, että parametrin arvo on kyseisen välin sisällä. (Gelman *et al.*, 2005)

### Otanta-algoritmit

Tässä tutkielmassa estimointi tehdään JAGS-ohjelmalla (*Just Another Gibbs Sampler*, Plummer *et al.*, 2003), joka on MCMC-otantaan tarkoitettu avoimen lähdekoodin tilastollinen ohjelma. Sen otanta-algoritmit on kirjoitettu C-kielellä tehokkuuden ja yhteensopivuuden saavuttamiseksi eri alustoilla. Ohjelma vastaa syntaksiltaan tunnettua BUGS-ohjelmaa (Spiegelhalter *et al.*, 1996), mutta tässä tutkielmassa JAGS-ohjelma osoittautui hieman tehokkaammaksi vaihtoehdoksi. JAGS-ohjelma käyttää yhteisposteriorijakauman estimointiin Metropolis-Hastings- ja Gibbs-otantaa. Mallin 12 estimointiin tehty JAGS-ohjelmakoodi on esitetty kappaleessa 9.

Metropolis-Hastings -otannassa ketjulle poimitaan ehdokasarvot parametriavaruudesta käyttämällä sopivaa ehdotusjakaumaa. Ne hyväksytään uusiksi arvoksi, jos ns. Metropolis-suhde on suurempi kuin tasajakaumasta (0,1) poimittu satunnaisluku. Jos uusi arvo hylätään, poimitaan ketjun indeksille uudet parametriarvot. Hyväksymistodennäköisyyttä arvioidaan laskemalla hyväksyttyjen arvojen osuus. Se pyritään optimoimaan lähelle 44% ketjun adaptaatiovaiheessa, jotta ketjun arvot eivät olisi liian korreloituneita. Tuloksena on hyväksyttyjen arvojen sarja, joka konvergoi posteriorijakaumaan.

Gibbs-otanta on Metropolis-Hastings -otannan erikoistapaus, jota hyödynnetään erityisesti satunnaiskomponenttien estimoinnissa. Gibbs-otannassa jokaiselle parametrille $\theta_k$ poimitaan uusi arvo parametrin täysin ehdollistetusta jakaumasta (*full conditional distribution*). Gibbs-otanta on yleensä huomattavasti tehokkaampi kuin Metropolis-Hastings -otanta, mikäli moniu-

lotteiset täysin ehdollistetut jakaumat ovat tunnettuja. (Koistinen, 2013)

Havainnollistetaan seuraavaksi täysin ehdollistettujen jakaumien muodostamista Poisson-uskottavuuden ja Gamma-priorin posteriorijakaumasta (kaava 9). Lisäksi Gamma-jakauman parametrille $b$ määritellään priorijakauma $b \sim \text{Gamma}(\nu, \upsilon)$. Muodostetaan seuraavaksi mallin parametrille $(\lambda_1, \ldots, \lambda_n, b)$ yhteisposteriorijakauma:

$$\begin{aligned} p(\boldsymbol{\lambda}, b|\boldsymbol{y}; \boldsymbol{e}) &\propto p(\boldsymbol{y}|\boldsymbol{\lambda}; \boldsymbol{e}) p(\boldsymbol{\lambda}|b) p(b) \\ &= \prod_{i=1}^{n} \left( \frac{(\lambda_i e_i)^{y_i}}{y_i!} \exp\{-\lambda_i e_i\} \times \frac{b^a}{\Gamma(a)} \lambda_i^{a-1} \exp\{-b\lambda_i\} \right) \\ &\times \frac{\upsilon^\nu}{\Gamma(\nu)} b^{\nu-1} \exp\{-\upsilon b\} \\ &\propto \left( \prod_{i=1}^{n} \exp\{-\lambda_i(e_i + b)\} \lambda_i^{(y_i+a)-1} \right) \times b^{\nu-1} \exp\{-\upsilon b\}. \end{aligned}$$

Tästä voidaan päätellä täysin ehdollistetut jakaumat

$$\begin{aligned} p(\lambda_i|\lambda_{-i}, b, y_i; e_i) &\propto \lambda_i^{(y_i+a)-1} \exp\{-\lambda_i(e_i + b)\} \quad \text{ja} \\ p(b|\lambda_i, \nu, \upsilon) &\propto \lambda_i^{\nu-1} \exp\{-\upsilon\lambda_i\}, \end{aligned}$$

jotka ovat tunnettuja tilastollisia jakaumia

$$\begin{aligned} \lambda_i|\lambda_{-i}, b, y_i; e_i &\sim \text{Gamma}(y_i + a, e_i + b) \\ b|\lambda_i, \nu, \upsilon &\sim \text{Gamma}(\nu, \upsilon). \end{aligned}$$

### Parametrilaajennus

Kun mallin satunnaisvaihtelu on jaettu useaan osaan, voivat satunnaiskomponentin hajontaparametrin arvot olla hyvin pieniä. Silloin satunnaiskomponenttien kertoimet painottuvat lähelle ryhmän keskiarvoa ja hajontaparametrin ketju puolestaan jumiutuu lähelle nollaa. Tällöin Gibbs-otanta konvergoi hitaasti. Satunnaiskertoimien ja hajontaparametrin välistä riippuvuutta voidaan vähentää *parametrilaajennuksella*.

Parametrilaajennus on laskennallinen keino, jolla parannetaan satunnais-

komponetin hajonnanparametrin konvergenssia. Satunnaiskomponentin $\boldsymbol{u} \sim \text{Normal}(0, \sigma_u)$ ja sen hajontaparametrin $\sigma_u$ välistä riippuvuutta pienennetään kertomalla satunnaiskomponentin $u$ arvot vakiolla $\xi$. Nyt satunnaiskomponentti määritellään uudelleen $u^* = \xi u$, jossa $\xi \sim \text{Normal}(0, 100000)$. Hajonta $\sigma_u$ skaalataan takaisin sen alkuperäiselle asteikolle kertomalla se $\xi$:n itseisarvolla, $\sigma_u^* = |\xi|\sigma_u$. (Gelman *et al.*, 2005; Gelman ja Hill, 2006)

## 4.2 Kovarianssimatriisin matriisihajotelma ja sen estimointi

Riippuvien satunnaislukujen poiminta moniulotteisesta normaalijakaumasta voidaan tehdä tehokkaasti kertomalla normaalijakautunut satunnaisvektori kovarianssimatriisin matriisihajotelmalla (Gelman *et al.*, 2005). Menetelmää on sovellettu perheaineistoissa sukulaisuussuhteiden kovarianssirakenteen mallinnuksessa (Bae *et al.*, 2014; Jamsen *et al.*, 2012). Sen etuna on, että matriisihajotelma ratkaistaan vain kerran kunkin perheen sukulaisuusmatriisille $\boldsymbol{R}_i$ ja se voidaan antaa malliin tunnettuna tietona. Aineiston koon kasvaessa kovarianssimatriisi kasvaa toisen potenssin funktiona, minkä seurauksena sen käsittely hidastuu. Kuten Jamsen *et al.* (2012), myös tässä tutkielmassa matriisin koko minimoitiin pinoamalla perheiden sukulaisuusmatriisien hajotelmat päällekkäin lohkodiagonaalimuodon sijaan.

Seuraavaksi johdetaan matriisihajotelman hyödyntäminen perheaineistossa. Kahden riippumattoman standardinormaalijakautuneen muuttujan $\boldsymbol{x}$ ja $\boldsymbol{y}$ kovarianssi voidaan kirjoittaa

$$\text{COV}(\boldsymbol{x}, \boldsymbol{y}) = \text{E}((\boldsymbol{x} - \text{E}(\boldsymbol{x}))(\boldsymbol{y} - \text{E}(\boldsymbol{y}))') = \text{E}(\boldsymbol{x}\boldsymbol{y}') = \boldsymbol{\Sigma}.$$

Kovarianssimatriisi $\boldsymbol{\Sigma}$ on siis $2 \times 2$ matriisi, jonka diagonaalit ovat $\Sigma_{11} = \text{V}(\boldsymbol{x})$ ja $\Sigma_{22} = \text{V}(\boldsymbol{y})$. Koska satunnaismuuttujat $\boldsymbol{x}$ ja $\boldsymbol{y}$ ovat riippumattomia, diagonaalin ulkopuoliset alkiot ovat $\Sigma_{12} = \Sigma_{21} = 0$, ja tällöin $\boldsymbol{\Sigma} = \boldsymbol{I}$.

Oletetaan, että riippuvien standardinormaalien satunnaisvektoreiden muodostama matriisi on $\boldsymbol{Z}$. Matriisi $\boldsymbol{Z}$ voidaan lausua matriisien tulona $\boldsymbol{G}\boldsymbol{X}$, jossa $\boldsymbol{X}$:lle pätee $\text{E}(\boldsymbol{X}\boldsymbol{X}') = \boldsymbol{I}$ ja $\boldsymbol{G}$ on kerroinmatriisi, joka määrittelee

havaintojen välisen riippuvuuden. Esitetään $\boldsymbol{Z}$:n kovarianssi

$$\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}') = \mathrm{E}(\boldsymbol{G}\boldsymbol{X}(\boldsymbol{G}\boldsymbol{X})') = \boldsymbol{G}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}')\boldsymbol{G}' = \boldsymbol{G}\boldsymbol{G}' = \boldsymbol{\Sigma},$$

jossa $\boldsymbol{\Sigma}$ on kovarianssimatriisi ja $\boldsymbol{G}\boldsymbol{G}'$ sen eräs matriisihajotelma. Koska kovarianssimatriisi $\boldsymbol{\Sigma}$ on symmetrinen ja sille on määritelty käänteismatriisi, voidaan se esittää spektraalihajotelmana eli ominaisarvo-ominaisvektori -hajotelmana

$$\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}'.$$

Matriisi $\boldsymbol{T}$ on ortogonaalinen ominaisvektorimatriisi, jolle pätee $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{I}$. Diagonaalimatriisi $\boldsymbol{\Lambda} = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_n),\ \Lambda_1 \geq \cdots \geq \Lambda_n$ on järjestetty ominaisarvomatriisi, jolle pätee $\sqrt{\boldsymbol{\Lambda}} = \sqrt{\boldsymbol{\Lambda}}'$.

Matriisihajotelman avulla ratkaistaan kerroinmatriisi $\boldsymbol{G}$

$$\boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}' = \boldsymbol{T}\sqrt{\boldsymbol{\Lambda}}\sqrt{\boldsymbol{\Lambda}}'\boldsymbol{T}' = (\boldsymbol{T}\sqrt{\boldsymbol{\Lambda}})(\boldsymbol{T}\sqrt{\boldsymbol{\Lambda}})' = \boldsymbol{G}\boldsymbol{G}'.$$

Nyt standardinormaalien satunnaismuuttujien $\boldsymbol{Z}$ havaintojen välinen riippuvuus on määritelty kerroinmatriisin ja riippumattomien standardinormaalien satunnaisvektoreiden tulona $\boldsymbol{G}\boldsymbol{X}$, jossa $\boldsymbol{G} = \boldsymbol{T}\sqrt{\boldsymbol{\Lambda}}$.

### Sovellus perheaineistossa

Edellä kuvattua menetelmää voidaan soveltaa perherakenteiden mallinnuksessa (Bae *et al.*, 2014). Perheenjäsenten välinen riippuvuusrakenne voidaan kuvata tunnettuna matriisina $\boldsymbol{R}_i$ (ks. kappale 3), jonka matriisihajotelma on $\boldsymbol{G}_i\boldsymbol{G}_i' = \boldsymbol{R}_i$. Nyt perimän satunnaistekijä $F_i$ noudattaa normaalijakaumaa odotusarvolla 0 ja hajonnalla $\sigma_f$, johon voidaan lisätä riippuvuusrakenne matriisitulolla $\boldsymbol{G}_i F_i$.

Samaa menetelmää voidaan käyttää perheaineiston simuloinnissa, jossa syöpäriskin halutaan korreloivan sukulaisuussuhteen mukaan. Perheenjäsenten $1, \ldots, J_i$ syöpäriskin riippuvuus saadaan simuloimalla $J_i$ pituinen riippumaton normaalijakautunut satunnaisvektori $F_i$ hajonnalla $\sigma_f$. Perheenjäsenten riippuvuusrakenne satunnaisvektoriin saadaan matriisitulosta $F_i^* = \boldsymbol{G}_i F_i$.

# 5 Nuoruusiän syövän väestöpohjainen perheaineisto

## 5.1 Nuoruusiän syöpäaineisto

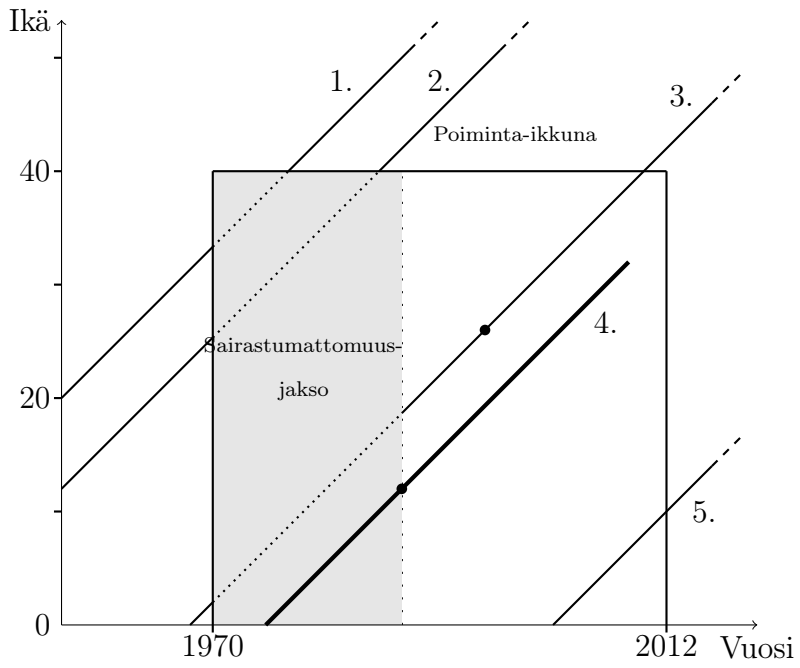
Lähtöaineistona on Suomen Syöpärekisterin 1.1.1970 ja 31.12.2012 välillä sekä alle 40-vuotiaana diagnosoidut nuoret syöpäpotilaat. Aineistoa on laajennettu hakemalla Suomen Väestörekisterikeskuksesta syöpäpotilaille vanhemmat, sisarukset, jälkeläiset sekä jälkeläisten toinen vanhempi. Tutkimuskohortista on poistettu adoptoidut ja adoptioon annetut henkilöt. Aineisto sisältää sukulaisuussuhteen lisäksi tiedon sukupuolesta sekä seuranta-ajan määrittelyyn tarvittavat päivämäärät. Perheenjäsenille on haettu syöpätiedot Suomen Syöpärekisteristä 1.1.1953 alkaen. Päivämäärä on sama kuin se, josta perheenjäsenten seuranta on aloitettu.

Perheen poimintaikkunan sisällä todettu ensimmäinen syöpätapaus määritellään perheen indeksihenkilöksi. Jokaisessa perheessä on vain yksi indeksihenkilö, jonka ympärille perhettä laajennetaan.

Poiminta-asetelma ja seuranta on havainnollistettu Lexis-kuvaajalla kuvassa 2 (Keiding, 1990). Jokaiselle perheelle määritellään kuvassa esitetty *sairastumattomuusjakso*, joka alkaa vuodesta 1970, päättyy perheen indeksihenkilön syöpädiagnoosiin ja koskee seurantaa vain alle 40 vuoden iässä. Perheenjäsen on syöpäriskissä ainoastaan sairastumattomuusjakson ulkopuolella lukuunottamatta indeksihenkilön jälkeläisen toista vanhempaa, jolle seuranta määritellään syntymästä alkaen. Tällä tavoin poistetaan indeksihenkilön valinnasta syntyvä harha (ks. kappale 5.2). Seuranta lopetataan joko kuolemaan, maastamuuttoon tai tutkimuksen päättymiseen 31.12.2012.

## 5.2 Harhan lähteet

Esiinkaivuharha on valikoituneisuutta ja sillä tarkoitetaan havaintojen poimimista siten, että poiminnan todennäköisyys riippuu kiinnostuksen kohteena olevasta tapahtumasta. Valikoituneisuutta voidaan korjata poistamalla perheen indeksihenkilö, jonka perusteella perhe on alunperin tullut poimituk-




Kuva 2: Lexis-kuvaaja esimerkkiperheestä, jossa indeksihenkilö (4.) on perheen kalenteriajassa mitattuna ensimmäinen syöpädiagnoosi poimintaikkunan sisällä. Vanhemmat (1. ja 2.) ovat riskissä poimintaikkunan ulkopuolella, sisarus (4.) ja lapsi (5.) vasta indeksihenkilön syövän jälkeen. Poiminnasta johtuva sairastumattomuusjakso on merkitty harmaalla alueella ja henkilön todellinen seuranta-aika yhtenäisellä viivalla.

si tutkimukseen (Fisher, 1934). Tällaisen korjauksen toimivuutta arvoidaan kappaleessa 6.2.

Sairastumattomuusharha syntyy väärin määritellystä seuranta-ajasta, jolloin henkilö ei ole esimerkiksi poimintatavasta johtuen voinut saada syöpädiagnoosia. Harha voidaan eliminoida määrittelemällä riskissäoloaika asetelman vaatimalla tavalla. (Lévesque *et al.*, 2010)

## 5.3 Odotettujen tapausten laskenta

Perheenjäsenten riskissäoloaika muutetaan odotettujen tapausten lukumääräksi $e_{ij}$ suhteuttamalla havaittu riskissäoloaika Suomen väestön syöpäriskiin. Odotettujen tapausten laskentaa varten $i$:n perheen henkilön $j$ henkilövuo-

det $py_{ij}$ pilkotaan 5-vuotisikäryhmän, kalenterivuoden ja sukupuolen mukaisiin ositteisiin $k$. Nyt yksilön henkilövuodet $py_{ijk}$ kerrotaan Suomen väestön vastaavan ositteen $k$ syöpäriskillä $\lambda_k$. Näin perheen $i$ henkilölle $j$ saadaan odotettujen syöpätapauksien lukumäärä summaamalla yli $k$:n

$$e_{ij} = \sum_{k=1}^{K} py_{ijk}\lambda_k.$$

Odotettujen tapausten lukumäärän määrittely yksilötasolla mahdollistaa mallin 12 mukaisen yksilötason kovarianssirakenteiden hyödyntämisen.

# 6 Simulaatiotutkimus

Simuloimalla arvioidaan paitsi esiinkaivuharhan korjauksen vaikutusta mallin 12 parametrien estimaatteihin myös sitä, miten suurta ympäristön ja perimän vaihtelun pitää olla, jotta satunnaiskomponenttien hajontaparametrit estimoituvat luotettavasti. Edellämainittuja arvioidaan estimoimalla simuloiduista aineistoista sekä parametrien estimaattien harhaa että niiden MCMC-ketjujen konvergenssia.

Keskimääräinen harha saadaan, kun lasketaan paljonko estimoitu arvo $\hat{\theta}$ poikkeaa sen todellisesta lähtöarvosta $\theta$. Se voidaan esittää erotuksien keskiarvona

$$\widehat{\text{harha}(\theta)} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta),$$

jossa $i = 1, \ldots, N$ on simulaatiokierroksen indeksi ja $\hat{\theta}_i$ on $i$:nnen simulaatiokierroksen estimaatti sen todellisesta arvosta $\theta$. Vastaavasti lasketaan simulaatiokierrosten varianssi $\mathrm{V}(\theta)$ estimaatille. Peitto on todennäköisyys sille, että todellinen arvo $\theta$ sisältyy estimoidun parametrin $\hat{\theta}$ 95%:n todennäköisyysvälin sisälle $(\hat{\theta}_L, \hat{\theta}_U)$ eli $P(\theta \in (\hat{\theta}_L, \hat{\theta}_U))$.

Lisäksi simulaatiotutkimuksessa arvioidaan posteriorijakaumasta simuloitujen Markovin ketjujen konvergenssia. Tämä tapahtuu laskemalla todennäköisyys sille, että Gelman-Rubinin -tunnusluku $\hat{R}$ on alle 1.1 eli $P(\hat{R} < 1.1)$. Ketjujen autokorrelaatiota ja sekoittumista arvioidaan puolestaan efektiivi-

sen otoskoon keskiarvolla $\bar{n}_{\text{eff}}$ (ks. kappale 4).

## 6.1 Perheaineiston generointi

Sairastumisikä on syövän oleellisin etiologinen tekijä, jonka vuoksi syöpätapaukset generoidaan paloittain ikäryhmäkohtaisista eksponenttijakaumista käyttäen ikäryhmäkohtaista riskiä $\lambda_k$, $k = 1, \ldots, 18$. Syöpätapauksien generointi yhdelle perheelle on kuvattu algoritmissa 1.

Perheenjäsenten ikäryhmäkohtaiseen riskiin $\lambda_k$ lisätään ympäristön $f$, perimän $g$ ja ylihajonnan *od* lisäriski (algoritmi 1, rivit 13). Lisäriskit simuloidaan annetuilla parametreilla $\sigma_f$ $\sigma_g$ ja $\sigma_{od}$ (algoritmi 1, rivit 6-8). Oletetaan lisäksi, että perheen sukulaisuusmatriisi $R_i$ ja sen hajotelma $G_i$ on ratkaistu valmiiksi.

**Algoritmi 1** Riippuvien syöpätapauksien generointi

1: **Proseduuri** GENEROINTI
2: *Tunneutut parametrit*:
3: $\lambda_k \leftarrow$ Ikäryhmäkohtainen riski, $k = 1, \ldots, 18$
4: $py_{jk} \leftarrow$ Perheenjäsenen $j$ henkilövuodet ikäryhmässä $k$
5: *Simuloidut parametrit*:
6: $f \sim \text{Normal}(0, \sigma_f)$ ▷ perheen lisäriski
7: $\boldsymbol{u} \sim \text{Normal}(\mathbf{0}_{J_i}, \sigma_g \boldsymbol{I}_N)$ ▷ geneettinen riippuvuus
8: $\boldsymbol{od} \sim \text{Normal}(\mathbf{0}_{J_i}, \sigma_{od} \boldsymbol{I}_N)$ ▷ Poisson ylihajonta
9: $\boldsymbol{g} \leftarrow \boldsymbol{G}\boldsymbol{u}$ ▷ $\boldsymbol{G}$ on tunnettu matriisi
10: *Syöpien generointi ikäryhmittäin*:
11: **for** $j \leftarrow 1 : J_i$ **do** ▷ perheenjäsenten lukumäärä
12: **for** $k \leftarrow 1 : 18$ **do**
13: $\lambda_k^* \leftarrow \exp\{\log(\lambda_k) + f + g_j + od_j\}$ ▷ malli 12
14: $t_k \sim \text{Exp}(\lambda_k^*)$
15: **if** $t_k < py_k$ **then**
16: $py_k \leftarrow t_k$
17: $py_l \leftarrow 0$, jossa $l > k$ ▷ seuranta päättyy syöpähavaintoon
18: $y_k \leftarrow 1$
19: **Break**
20: **else**
21: $y_k \leftarrow 0$
22: $e_k \leftarrow py_k \times \lambda_k$ ▷ odotettujen tapausten määrittely

Nyt eksponenttijakaumasta simuloidaan ikäryhmän $k$ ilmaantuvuuteen $\lambda_k^*$ perustuva tapahtumahetki $t_k$, jota verrataan henkilön seuranta-ajan pituuteen $py_k$. Jos tapahtuma-aika $t_k$ on seuranta-ajan $py_k$ sisällä, merkitään henkilölle syöpähavainto ikäryhmässä $k$ ja henkilön seuranta lopetetaan hetkeen $t_k$ (rivit 14-18). Jos tapahtuma-aika $t_k$ on kyseisen ikäryhmän seuranta-ajan $py_k$ jälkeen, niin syöpää ei havaita ja siirrytään seuraavaan ikäryhmään. Lisäksi kullekin henkilölle määritellään henkilövuosien $py_k$ ja riskin $\lambda_k$ tulosta ikäryhmäkohtainen odotettujen tapausten lukumäärä $e_k = py_k\lambda_k$.

Tapauksien generointi eksponenttijakaumasta perustuu sen ja Poisson-jakaumaan väliseen yhteyteen, jossa Poisson-prosessin insidenssien aikapisteiden väli on eksponenttijakautunut. (Holford, 1980)

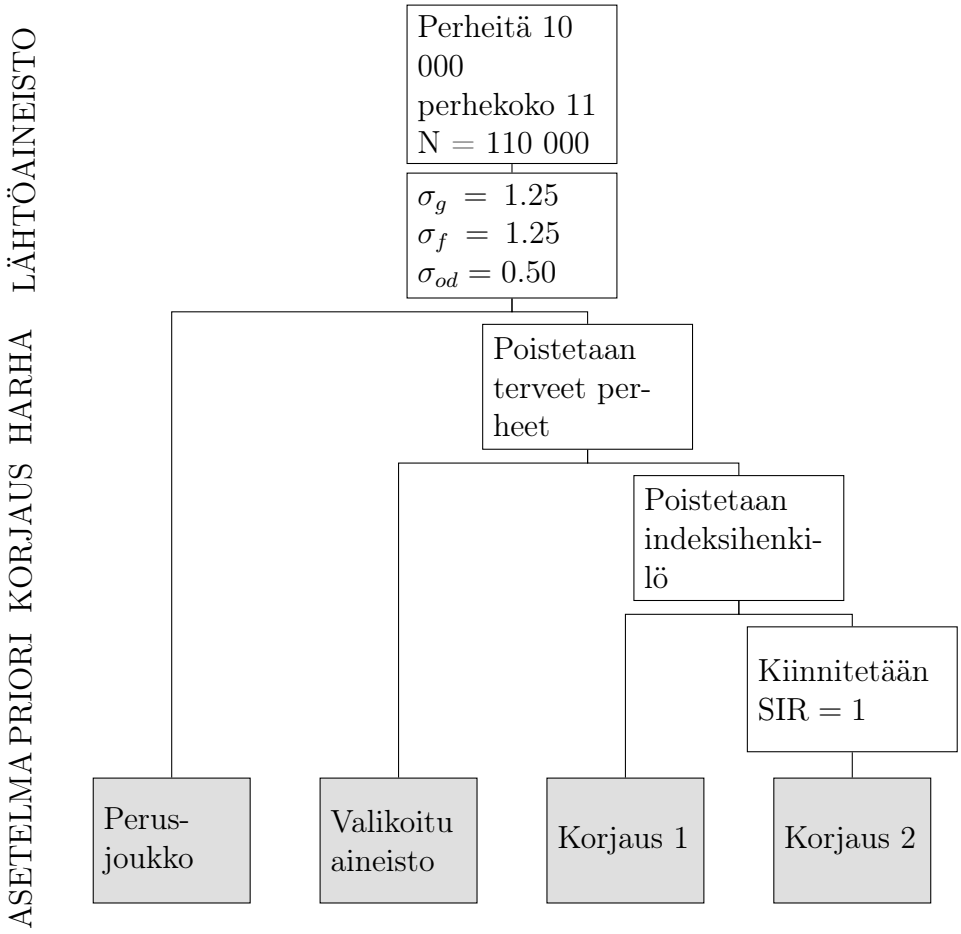
## 6.2 Estimaattorin ominaisuudet

Ensimmäisessä simulaatiotutkimuksessa arvioidaan esiinkaivuharhan vaikutusta ja sen korjausta. Sen tavoitteena on jäljitellä kappaleessa 5.1 kuvattua perheiden poimintaa perusjoukosta, jossa poimintakriteerit täyttävä indeksihenkilö sisällyttää perheensä aineistoon.

Simuloidussa aineistossa on 10 000 sukulaisuussuhteiltaan ja seuranta-ajoiltaan samanlaista 11 henkilön perhettä. Perherakenne on esitetty kuvassa 1. Perhe koostuu perustajajäsenistä (1 ja 2), heidän kolmesta lapsestaan ja lasten puolisoista (3, 4, 5, 6 ja 7), joista kahdella parilla on yhteensä neljä lasta (8, 9, 10 ja 11). Ikäryhmäkohtaisen syöpäriskin profiili valittiin simulaatiotutkimuksessa vastaamaan suunnilleen ruoansulatuselinten keskimääräistä ikäryhmäkohtaista syöpäriskiä suomalaisessa väestössä.

Mallin 12 yhteisposteriorista (kaava 15) simuloitiin kaksi 25 000 otoksen ketjua, joista molemmista poistettiin ensimmäiset 5000 iteraatiota. Talletustilan säästämiseksi tilastolliseen päättelyyn sisällytettiin lopuista 20 000:sta iteraatiosta joka 20:s. Simulaatio toistettiin jokaisessa asetelmassa 50 kertaa. Yksi simulaatiokokeen kierros kesti noin 10 tuntia.

Lukuunottamatta perusjoukkoa aineistoon luotiin poimintaharha poistamalla perheet, joihin ei generoitunut yhtään syöpäpotilasta. Harhaa korjattiin asetelmissa poistamalla indeksihenkilö (korjaukset 1 ja 2) ja lisäksi kiin-




Kuva 3: Asetelmaharhan suuruutta arvioidaan *valikoidulla aineistolla*, jossa *perusjoukosta* on valittu sairaat perheet. Valikointia korjataan poistamalla yksi syöpätapaus *(korjaus 1)* sekä lisäksi kiinnittämällä SIR *(korjaus 2.)*

nittämällä mallin vakiotermi $\beta$ lukuun nolla (korjaus 2). Perheen indeksihenkilöksi valittiin syöpätapauksista satunnaisesti yksi henkilö. Indeksihenkilön syöpähavainnot ja odotettujen tapausten lukumäärä asetettiin nollaksi, jotta perheen geneettiset etäisyydet pysyisivät määriteltyinä. Aineiston generointi ja harhan korjaukset eri asetelmissa on esitetty kuvassa 3.

Simulaatiokokeen tulokset on raportoitu taulukkoon 1. Perusjoukko on kokonaisotos ja estimaatit ovat odotetusti lähes harhattomia. Mallin vakiotermi ($\exp\{\beta\} = \text{SIR}$) on yksi, koska kaikki ilmaantuvuuden vaihtelu selit-

Taulukko 1: Parametrien estimaatit (95%:n todennäköisyysvälit), estimaattoreiden harha (suhteellinen harha), peitto 50 simulaatiokierroksen perusteella.

| | Todellinen | Perusjoukko[a] | Valikoitu aineisto[b] | Korjaus 1[c] | Korjaus 2[d] |
|---|---|---|---|---|---|
| Malli 8 | | | | | |
| SIR | | 5.06 (4.94-5.16) | 13.8 (13.5-14.1) | 7.16 (6.93-7.38) | 7.16 (6.93-7.38) |
| Malli 12 | | | | | |
| $\widehat{\mathrm{E}(\mathrm{SIR})}$ | 1 | 1.01 (0.88-1.15) | 7.29 (6.86-7.73) | 1.37 (1.18-1.55) | 1.01 (0.99-1.03) |
| $\widehat{\mathrm{E}(\sigma_f)}$ | 1.5 | 1.51 (1.42-1.59) | 0.05 (0.00-0.13) | 1.53 (1.44-1.63) | 1.65 (1.56-1.74) |
| $\widehat{\mathrm{E}(\sigma_g)}$ | 1.5 | 1.48 (1.37-1.60) | 1.38 (1.32-1.44) | 1.58 (1.45-1.71) | 1.67 (1.54-1.80) |
| $\widehat{\mathrm{E}(\sigma_{od})}$ | 0.5 | 0.47 (0.16-0.72) | 0.62 (0.45-0.76) | 0.72 (0.46-0.94) | 0.91 (0.71-1.09) |
| $\widehat{\mathrm{E}(h^2)}$ | 0.43 | 0.47 (0.40-0.53) | 0.83 (0.76-0.90) | 0.46 (0.39-0.54) | 0.44 (0.37-0.51) |
| V(SIR) | | 0.0050 | 0.0509 | 0.0093 | 0.0001 |
| $\mathrm{V}(\sigma_f)$ | | 0.0020 | 0.0016 | 0.0025 | 0.0021 |
| $\mathrm{V}(\sigma_g)$ | | 0.0036 | 0.0010 | 0.0048 | 0.0047 |
| $\mathrm{V}(\sigma_{od})$ | | 0.0235 | 0.0066 | 0.0169 | 0.0101 |
| harha(SIR) | | 0.01 (1%) | 6.29 (629%) | 0.37 (37%) | 0.01 (1%) |
| harha($\sigma_f$) | | 0.01 (1%) | -1.45(-116%) | 0.03 (2%) | 0.15 (12%) |
| harha($\sigma_g$) | | -0.02 (-2%) | -0.12 (-10%) | 0.08 (6%) | 0.17 (14%) |
| harha($\sigma_{od}$) | | -0.03 (-6%) | 0.12 (24%) | 0.22 (44%) | 0.41 (82%) |
| harha($h^2$) | | 0 (0%) | 0.24 (56%) | -0.02 (-5%) | -0.04 (-9%) |
| peitto(SIR) | | 98 | 0 | 0 | 100 |
| peitto($\sigma_f$) | | 86 | 0 | 86 | 5 |
| peitto($\sigma_g$) | | 94 | 2 | 80 | 35 |
| peitto($\sigma_{od}$) | | 90 | 66 | 56 | 5 |
| peitto($h^2$) | | 76 | 0 | 84 | 90 |

[a] Perusjoukko on koko lähtöaineisto.
[b] Valikoitu aineisto on ne perusjoukon perheet, joissa on vähintään yksi syöpätapaus.
[c] Korjaus 1 on valikoitu aineisto, joista on poistettu yksi syöpätapaus.
[d] Korjaus 2 on myös valikoitu aineisto, jossa poistetun syöpätapauksen lisäksi suhteellisen ilmaantuvuuden estimaatti on kiinnietty lukuun 1.

tyy satunnaiskomponenteilla. Vakiomallin (malli 8) perusteella suhteellisen ilmaantuvuuden estimaatti on viisinkertainen (5.06) satunnaiskomponenttimalliin (malli 12) verrattuna (1.01).

Valikoitu aineisto sisältää vain sellaiset perheet, joissa on vähintään yksi syöpätapaus ilman mitään harhan korjausta. Suhteellisen ilmaantuvuuden

estimaatit mallien 8 ja 12 perusteella ovat erittäin harhaisia. Ympäristön satunnaiskomponentin harha on erittäin suuri. Peitto on kaikissa parametreissa välttävä tai olematon.

Korjaus 1 arvioi miten harhan korjaus toimii, kun perheistä poistetaan indeksihenkilö. Nyt suhteellinen ilmaantuvuus on yliestimoitunut ja sen peitto on nolla. Myös satunnaistermien hajonnat ovat järjestäen hieman yliestimoituneita, mutta perimäosuuden estimaatti $\hat{h}^2$ on tästä huolimatta lähellä todellista arvoaan. Myös satunnaistermien ja perimäosuuden peittotodennäköisyydet ovat kohtalaisen hyvät.

Korjaus 2 vastaa korjausta 1 sillä lisäyksellä, että suhteellisen ilmaantuvuuden estimaatti on kiinnitetty lukuun 1. Parametrin kiinnitys mallissa olettaa, että kaikki lisäriski johtuu perimästä, ympäristöstä tai ylihajonnasta, mikä tässä aineistossa on totta. Satunnaiskomponenttien hajonnat ovat yliestimoituneita, mutta perimäosuuden peitto on hyvä. Vaikka olisikin realistista olettaa, että mallin satunnaiskomponentit selittävät kaiken vaihtelun, aiheuttaa poiminta harhaa satunnaiskomponentteihin. Molempien korjauksien 2 ja 3 jälkeen myös ilman satunnaiskomponetteja mallinnettu SIR (malli 8) on käyttökelpoinen, sillä se kuvaa indeksihenkilön perheenjäsenten keskimääräistä SIR:iä ja sitä voidaan soveltaa aineistoanalyysissa.

Jatkossa tässä tutkielmassa sovelletaan korjausta 1, koska se antaa parhaat peittotodennäköisyydet satunnaiskomponenttien hajontaparametreille.

## 6.3 Informaation riittävyyden arviointi syöpäaineistossa

Toisessa simulaatiokokeessa arvioidaan sitä, paljonko perheisiin tarvitaan syöpätapauksia, jotta estimointia voidaan pitää luotettavana. Aineistona on nuoruusiällä ruoansulatuselinten syöpään sairastuneiden perheet, joihin generoidaan havaintoja todellisilla ikäryhmäkohtaisilla ilmaantuvuuksilla $\lambda_k$. Ikäryhmäkohtaisten ilmaantuvuuksien laskennassa on hyödynnetty väestön sukupuolen ja kalenteriperiodin mukaan laskettua syöpäilmaantuvuutta.

Ympäristön ja perimän satunnaiskomponenttien suuruusluokka haarukoitiin taulukoimalla generoitujen syöpätapauksien kasautumista eri $\sigma_g$:n ja

Taulukko 2: Havaintojen kasautumista perheisiin arvioitiin perheen syöpäfrekvenssillä (Syöpien lkm) ruoansulatuselinten syöpäaineistossa. Todelliset syöpätapaukset (Todellinen) on korjattu poistamalla perheen indeksihenkilö. Kasautumista eri hajontaparametrien arvoilla (Generointi) arvioitiin generoimalla tapaukset algoritmilla 1, jossa $\sigma = \sigma_g = \sigma_f$ ja $\sigma_{od} = 0.5$.

| | | Generoitu | | |
|---|---|---|---|---|
| Syöpien lkm | Todellinen | $\sigma = 1.25$ | $\sigma = 0.75$ | $\sigma = 0.50$ |
| 0 | 2802 | 2609 | 2844 | 2896 |
| 1 | 444 | 543 | 424 | 388 |
| 2 | 62 | 139 | 51 | 40 |
| 3 | 12 | 24 | 5 | 4 |
| 4 | 3 | 9 | 4 | 0 |
| 5 | 2 | 2 | 0 | 0 |
| 6 | 1 | 2 | 0 | 0 |

$\sigma_f$:n arvoilla, ja vertailemalla niitä todellisiin syöpätapauksiin (taulukko 2).

Todelliset syöpätapaukset on korjattu poistamalla perheen indeksihenkilö, kuten kappaleessa 6.2 on esitetty. Kun hajontaparametrien arvot ovat välillä $0.5 \leq \sigma_f, \sigma_g \leq 1.25$, onnistutaan generoimaan todellista aineistoa vastaava syöpien perheittäinen kasautuminen. Simulaatiokoe toteutettiin kaikilla hajontaparametrien kombinaatioilla $\sigma_f = \{1.25, 0.75, 0.50\}$ ja $\sigma_g = \{1.25, 0.75, 0.50\}$. Ylihajonta on sama jokaisessa asetelmassa $\sigma_{od} = 0.5$. Eli yhteensä yhdeksän eri asetelman lähtöparametrit, parametrien estimaatit sekä konvergenssin arviointi on esitetty taulukossa 3.

Kaikissa yhdeksässä eri asetelmassa simulaatiokoe toistettiin 50 kertaa, jossa yhden iteraation laskenta-aika kesti noin 10 tuntia. Parametrit estimoitiin simuloimalla kaksi ketjua eri alkuarvoilla. Posteriorijakaumasta poimittiin 25 000 otosta, joista ensimmäiset 5000 jätettiin käyttämättä. Ketjuja ohennettiin valitsemalla vain joka 20:s, jolloin ketjun pituus oli kokonaisuudessaan 1000.

Parametrit estimoituivat pääosin lähelle todellisia lähtöarvojaan ja niiden peitto oli 80-100% välillä. Suurimmilla lähtöarvoilla harha oli pienempää ja todennäköisyysvälit kapeampia.

Ketjut konvergoituivat hyvin ja $\hat{R}$ pysyi alle viitearvon 1.1 vaihdellen välillä 68-100% kaikissa eri asetelmissa. Ketjujen sekoittumista kuvaava efektiivisen otoskoon keskiarvo $\bar{n}_{\text{eff}}$ oli yli 100 kaikissa asetelmissa. SIR:n ketjut sekoittuvat erityisen hyvin, mutta keskimääräinen efektiivinen otoskoko madaltui, kun hajontaparametrien lähtöarvot pienenivät. Tuloksista nähdään, että malli identifioi hyvin jopa kolme vaihtelun lähdettä. On siis varsin uskottavaa, että ruoansulatuselinten syöpien perimäosuus estimoituu oikein myös todellisessa aineistossa.

Taulukko 3: Informaation riittävyyttä arvioitiin simuloimalla yhdeksän eri lähtöarvokompinaatiota (vasen sarake) 50 kertaa. Tulosten tarkkuutta arvioitiin estimaatin odotusarvolla $\hat{\mathrm{E}}$ ja sen 95%:n todennäköisyysväleillä sekä parametrin peiton todennäköisyydellä. Ketjun konvergenssia tarkasteltiin efektiivisellä otoskoolla ($n_{\mathrm{eff}}$) ja osuudella, että Gelman-Rubinin tunnusluku $R$ on alle 1.1, $\mathrm{P}(R < 1.1)$.

| Lähtöarvo | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SIR | $\sigma_f$ | $\sigma_g$ | $\sigma_{od}$ | $h^2$ | | SIR | $\sigma_f$ | $\sigma_g$ | $\sigma_{od}$ | $h^2$ | | SIR | $\sigma_f$ | $\sigma_g$ | $\sigma_{od}$ |
| 1 | 1.25 | 1.25 | 0.50 | 0.46 | $\hat{\mathrm{E}}$ | 1.02 (0.81-1.24) | 1.24 (1.07-1.41) | 1.17 (0.79-1.48) | 0.51 (0.08-0.95) | 0.43 (0.21-0.61) | $n_{\mathrm{eff}}$ | 713 | 592 | 209 | 132 |
| | | | | | Peitto | 94% | 91% | 87% | 96% | 88% | $\mathrm{P}(\hat{R} < 1.1)$ | 100 | 100 | 98 | 90 |
| 1 | 0.75 | 1.25 | 0.50 | 0.66 | $\hat{\mathrm{E}}$ | 0.96 (0.73-1.21) | 0.74 (0.40-1.01) | 1.22 (0.72-1.58) | 0.57 (0.06-1.06) | 0.61 (0.27-0.88) | $n_{\mathrm{eff}}$ | 555 | 208 | 156 | 136 |
| | | | | | Peitto | 86% | 95% | 92% | 98% | 95% | $\mathrm{P}(\hat{R} < 1.1)$ | 100 | 94 | 90 | 84 |
| 1 | 0.50 | 1.25 | 0.50 | 0.76 | $\hat{\mathrm{E}}$ | 1.02 (0.76-1.29) | 0.44 (0.08-0.75) | 1.21 (0.76-1.55) | 0.48 (0.02-0.98) | 0.74 (0.36-0.98) | $n_{\mathrm{eff}}$ | 513 | 152 | 175 | 146 |
| | | | | | Peitto | 92% | 96% | 93% | 100% | 99% | $\mathrm{P}(\hat{R} < 1.1)$ | 100 | 78 | 88 | 86 |
| 1 | 1.25 | 0.75 | 0.50 | 0.24 | $\hat{\mathrm{E}}$ | 0.98 (0.75-1.22) | 1.26 (1.10-1.42) | 0.64 (0.14-1.05) | 0.56 (0.08-0.97) | 0.19 (0.02-0.40) | $n_{\mathrm{eff}}$ | 508 | 748 | 171 | 177 |
| | | | | | Peitto | 91% | 94% | 93% | 96% | 91% | $\mathrm{P}(\hat{R} < 1.1)$ | 100 | 100 | 96 | 96 |
| 1 | 0.75 | 0.75 | 0.50 | 0.41 | $\hat{\mathrm{E}}$ | 0.99 (0.69-1.30) | 0.71 (0.39-0.96) | 0.63 (0.05-1.09) | 0.59 (0.03-1.09) | 0.30 (0.02-0.72) | $n_{\mathrm{eff}}$ | 319 | 301 | 152 | 167 |
| | | | | | Peitto | 94% | 90% | 96% | 100% | 96% | $\mathrm{P}(\hat{R} < 1.1)$ | 100 | 82 | 86 | 90 |
| 1 | 0.50 | 0.75 | 0.50 | 0.53 | $\hat{\mathrm{E}}$ | 0.93 (0.61-1.26) | 0.43 (0.08-0.73) | 0.65 (0.10-1.13) | 0.62 (0.05-1.17) | 0.39 (0.04-0.86) | $n_{\mathrm{eff}}$ | 272 | 216 | 169 | 152 |
| | | | | | Peitto | 94% | 93% | 94% | 97% | 96% | $\mathrm{P}(\hat{R} < 1.1)$ | 98 | 84 | 88 | 88 |
| 1 | 1.25 | 0.50 | 0.50 | 0.12 | $\hat{\mathrm{E}}$ | 0.99 (0.75-1.23) | 1.23 (1.08-1.39) | 0.46 (0.04-0.86) | 0.47 (0.04-0.88) | 0.11 (0.00-0.32) | $n_{\mathrm{eff}}$ | 461 | 903 | 226 | 216 |
| | | | | | Peitto | 94% | 98% | 98% | 98% | 98% | $\mathrm{P}(\hat{R} < 1.1)$ | 100 | 100 | 96 | 96 |
| 1 | 0.75 | 0.50 | 0.50 | 0.24 | $\hat{\mathrm{E}}$ | 0.95 (0.66-1.25) | 0.67 (0.35-0.94) | 0.53 (0.04-1.02) | 0.52 (0.02-1.05) | 0.26 (0.01-0.72) | $n_{\mathrm{eff}}$ | 300 | 339 | 184 | 169 |
| | | | | | Peitto | 97% | 85% | 99% | 100% | 99% | $\mathrm{P}(\hat{R} < 1.1)$ | 100 | 88 | 94 | 98 |
| 1 | 0.50 | 0.50 | 0.50 | 0.33 | $\hat{\mathrm{E}}$ | 0.90 (0.59-1.23) | 0.37 (0.03-0.70) | 0.54 (0.03-1.04) | 0.57 (0.02-1.12) | 0.34 (0.00-0.88) | $n_{\mathrm{eff}}$ | 268 | 240 | 200 | 183 |
| | | | | | Peitto | 96% | 96% | 100% | 100% | 99% | $\mathrm{P}(\hat{R} < 1.1)$ | 98 | 76 | 90 | 86 |

# 7 Nuoruusiän syövän perheaineiston analyysi

Tutkielmassa estimoitiin rintasyövän ja ruoansulatuselinten syöpien perheittäiset SIR:t (malli 8) sekä syöpien perimäosuudet (malli 12). Lisäksi aineistoihin tehtiin asetelmakorjaus poistamalla indeksihenkilö (kappaleen 6.2 korjaus 1). Yhteisposteriorijakaumasta simuloitiin kaksi 200 000 otosta, joista poistettiin ensimmäiset 100 000.

Syövän kasautumista perheisiin arvioitiin indeksihenkilöiden perheenjäsenten SIR:llä (taulukko 4). SIR estimoitiin sovittamalla malli 8 sekä kaikille perheenjäsenille yhdessä että ositettuna sukulaisuussuhteen mukaan. Lisäksi indeksihenkilön sisaruksen SIR estimoitiin erikseen riippuen onko ainakin toisella vanhemmista syöpä (sairas) vai ei (terve).

Taulukko 4: Mallin 8 mukaiset vakioidun ilmaantuvuussuhteen (SIR) posteriorikeskiarvot ja 95% todennäköisyysvälit ruoansulatuselinten syöpien ja rintasyövän indeksihenkilöiden perheille sukulaissuusuhteen mukaan.

| Syöpä | Sukulaisuussuhde | N | Henkilövuodet | Havaitut tapaukset | Odotetut tapaukset | SIR |
|---|---|---|---|---|---|---|
| Ruoansulatus | Äiti | 2697 | 120016 | 184 | 123 | 1.50 (1.27-1.71) |
| | Isä | 2544 | 104411 | 223 | 146 | 1.52 (1.33-1.74) |
| | Sisarus | 6298 | 170312 | 95 | 47 | 2.03 (1.65-2.44) |
| | – sairas[a] | 907 | 27466 | 33 | 8 | 4.36 (3.07-5.95) |
| | – terve[b] | 5391 | 142846 | 62 | 39 | 1.57 (1.17-1.97) |
| | Lapsi | 5023 | 108825 | 29 | 5 | 5.95 (3.96-8.16) |
| | Puoliso | 2551 | 80518 | 43 | 38 | 1.15 (0.84-1.50) |
| | Yhteensä | 19113 | 584082 | 574 | 359 | 1.60 (1.48-1.73) |
| Rinta | Äiti | 3913 | 184946 | 449 | 272 | 1.65 (1.50-1.81) |
| | Sisarus | 8968 | 245186 | 247 | 123 | 2.02 (1.78-2.27) |
| | – sairas[a] | 855 | 22514 | 36 | 11 | 3.20 (2.25-4.32) |
| | – terve[b] | 8113 | 222672 | 211 | 112 | 1.89 (1.65-2.16) |
| | Lapsi | 8257 | 171115 | 43 | 14 | 3.20 (2.23-4.16) |
| | Yhteensä | 21138 | 601247 | 739 | 408 | 1.81 (1.68-1.94) |

[a] Indeksihenkilön sisarus, jonka vähintään toiselle vanhemmista todettu syöpä
[b] Indeksihenkilön sisarus, jonka molemmat vanhemmista terveitä

Ruoansulatuselinten syöpien keskimääräinen SIR indeksihenkilöiden perheenjäsenille on 1.60 (95%:n todennäköisyysvälit 1.48-1.73) (taulukko 4) eli

perheenjäsenten syöpään sairastumisen riski on 60% suurempi normaaliväestöön verrattuna. Suhteellinen ilmaantuvuus on tilastollisesti merkitsevästi koholla kaikilla muilla sukulaisilla lukuunottamatta puolisoita, joilla suhteellinen ilmaantuvuus on 1.15 (0.84-1.50). Taulukosta havaitaan myös, että indeksihenkilön vanhempien SIR:t 1.50 (1.27-1.71) ja 1.52 (1.33-1.74) ovat pienemmät kuin sisaruksien 2.03 (1.65-2.44). Vastaavasti sisaruksen SIR on pienempi kuin indeksihenkilön lapsen 5.95 (3.96-8.16). Jos molemmat vanhemmista ovat terveitä, on sisaruksen suhteellinen syöpäilmaantuvuus huomattavasti pienempi 1.57 (1.17-1.97) kuin niissä perheissä, joissa ainakin toinen vanhemmista on sairas 4.36 (3.07-5.95). Mallin 12 perusteella estimoitu ruoansulatuselinten syöpien SIR on 0.74 (0.55-0.92) ja perimäosuuden estimaatiksi saadaan 63% (37-88%).

Rintasyöpäaineiston perheenjäsenten keskimääräinen SIR on 1.81 (1.68-1.94), joka tarkoittaa 81% kohonnutta riskiä vastaavanlaiseen väestöön verrattuna (taulukko 4). Kaikilla perheenjäsenillä riski sairastua syöpään on tilastollisesti merkitsevä. Indeksihenkilöiden äitien estimoitu SIR 1.65 (1.50-1.81) on pienempi kuin sisaruksien 2.02 (1.78-2.27). Korkein riski on indeksihenkilön lapsilla, joilla se on yli 300% suurempi kuin normaaliväestössä. Sisaruksien SIR oli suurempi, jos vanhemmalla on diagnosoitu syöpä 3.20 (2.25-4.32) verrattuna siihen, että vanhemmat ovat terveet 1.89 (1.65-2.16). Mallin 12 mukaan estimoitu SIR (taulukko 5) on 1.36 (1.17-1.57) ja perimäosuus on puolestaan 26% (0-57%).

Taulukko 5: Mallin 12 mukaan estimoidut posteriorikeskiarvot ja 95%:n todennäköisyysvälit suhteelliselle ilmaantuvuudelle (SIR) sekä satunnaiskomponenttien hajontaparametreille rintasyöpä- ja ruoansulatuselinten syöpien aineistossa asetelmakorjauksen (korjaus 1) jälkeen.

| Estimaatti | Ruoansulatuselimet | Rinta |
|---:|---|---|
| $\widehat{\text{SIR}}$ | 0.74 (0.55-0.92) | 1.36 (1.17-1.57) |
| Ympäristö $\hat{\sigma}_f$ | 0.39 (0.00-0.73) | 0.62 (0.31-0.88) |
| Perimä $\hat{\sigma}_g$ | 1.21 (0.85-1.54) | 0.30 (0.00-0.69) |
| Ylihajonta $\hat{\sigma}_{od}$ | 0.36 (0.00-0.80) | 0.19 (0.00-0.45) |
| Perimäosuus $\hat{h}^2$ | 0.63 (0.37-0.88) | 0.26 (0.00-0.57) |

# 8 Päätelmät

Tämän tutkielman tavoitteena oli tarkastella perimän ja ympäristötekijöiden vaikutusta nuoruusiän syöpäsairauksien suhteelliseen ilmaantuvuuteen laajennetuissa perheaineistoissa. Perimäosuudeksi estimoitiin suhteellisen ilmaantuvuuden tilastollisella mallilla ruoansulatuselinten syöville 63% (95%:n todennäköisyysvälit 37-88%) ja rintasyövälle 26% (0-57%).

Tutkielma on ajankohtainen ja kiinnostava, sillä ihmisten genomien mittaaminen on viime aikoina jatkuvasti yleistynyt ja herättää runsaasti keskustelua paitsi tutkijayhteisöissä myös valtakunnan medioissa. Laajojen väestöryhmien genomimittaus on kuitenkin kallista ja vie paljon rerusseja, joten vaihtoehtoisten tutkimusmenetelmien soveltaminen riskiryhmien identifioimiseen isoista populaatioista on tarpeellista, jotta genomimittausta voidaan kohdentaa tehokkaammin. Tällainen lähestymistapa näin laajamittaisen perheaineiston analyysiin on suomalaisessa mittakaavassa uutta.

Tutkielmasssa sovellettiin perherakenteen riippuvuuden kuvaamisessa matriisihajotelmaa, joka mahdollisti perimän vaikutuksen arvioinnin laajennetussa perheaineistossa ilman genomidataa. Tällaiset mallit olettavat, että ympäristö ja perinnöllisyys ovat riippumattomia ja että ne vaikuttavat syöpään samalla tavalla sukupuolesta riippumatta. Lisäksi malleissa oletetaan, että puolisoiden pariutuminen on satunnaista. (Burton *et al.*, 1999)

Aikaisemmin perheaineistoja on mallinnettu parametroimalla sukulaisuussuhteet satunnaiskomponenttien eri kombinaatiolla (ks. mm. Burton *et al.*, 1999; Scurrah *et al.*, 2000). Tällöin mallin määrittely hankaloituu laajennetuissa perheaineistoissa, jossa on mukana myös erikoisempia perherakenteita. Tässä tutkielmassa sovellettu matriisihajotelman hyödyntäminen perheen riippuvuusrakenteen kuvaamisessa on uusi ja tehokas lähestymistapa (Bae *et al.*, 2014), ja osoittautui varsin toimivaksi sekä simuloiduilla aineistoilla että nuoruusiän syövän perheaineistoissa.

MCMC-estimointi mahdollisti sen, että posteriorijakauman avulla saadaan enemmän informaatiota tilastollisen päättelyn tueksi pelkkään uskottavuuspäättelyyn perustuvaan REML-estimoituihin varinssikomponentteihin verrattuna. Tässä tutkielmassa erityisesti satunnaiskomponenttien hajontaparametrien todennäköisyysvälit ovat välttämättömät niiden luotettavuuden arvioinnissa. Perheaineistojen analysointi MCMC-otannalla onkin osoittautunut nopeammaksi kuin REML-estimoiti (Jamsen *et al.*, 2012). Toteutettu parametrilaajennus paransi satunnaiskomponenttien hajontaparametrien estimoitumista merkittävästi.

Satunnaiskomponentin hajonnan estimaatti kuvaa ryhmien välistä *vaihtelua*, joka tarkentuu informaation kasvaessa. Tätä vaihtelua voidaan tulkita joko äärellisen populaation tai superpopulaation varianssina. Äärellisen populaation varianssi estimoidaan ryhmien havaituista keskiarvoista ja se kuvaa vaihtelua havaitussa populaatiossa. Superpopulaation varianssi estimoidaan satunnaiskomponentin hajontaparametrista. Silloin oletetaan, että havaitut ryhmät on poimittu superpopulaatiosta. Usein nämä kaksi estimaattia ovat lähellä toisiaan, mutta ne mittavat eri asioita. Tässä tutkielmssa ryhmien oletettiin luonnollisesti olevan otos superpopulaatiosta eli Suomen väestöstä, jonka väestöparametrit olivat kiinnostuksen kohteena. (Gelman ja Hill, 2006)

Lähtöaineistona tutkielmassa käytettiin Suomen Syöpärekisterin 1.1.1970 ja 31.12.2012 välillä sekä alle 40-vuotiaana diagnosoitujen nuorten syöpäpotilaiden rekisteriä. Aineistoa laajennettiin hakemalla Suomen Väestörekisterikeskuksesta syöpäpotilaille vanhemmat, sisarukset, jälkeläiset sekä jälkeläisten toinen vanhempi. Aineistoon valittiin rintasyöpää ja ruoansulatussyöpiä sairastavien henkilöiden perheet, koska kyseiset syövät ovat Suomen yleisimpiä syöpiä. Huomioon on syytä ottaa, että erityisesti ruoansulatuselinten syöpiin lukeutuu useita eri syöpätyyppejä, kuten esimerkiksi maha, ohutsuoli, paksusuoli ja peräsuoli. Tässä tutkielmassa oletuksena oli, että ympäristötekijöiden ja perimän riippuvuus on samanlaista kaikilla näillä syöpätyypeillä.

Kuvatulla tavalla poimittuun aineistoon muodostuu poimintaharhaa, kos-

ka terveet perheet eivät ole edustettuina. Harhan määrä riippuu siitä, mikä on koko populaation ja asetelmakorjatun aineiston terveiden ja sairaiden perheiden suhde. Asetelmakorjaus on mahdollista tehdä myös korjaamalla uskottavuusfunktiota (Burton, 2003; Burton *et al.*, 2000), mutta tätä ei sovellettu tässä tutkielmassa, sillä se olisi poissulkenut JAGS/BUGS-ohjelmien käytön. Asetelmakorjauksen harhaa arvioidessa tuli huomoida, että simulaatiokokeessa käytetyn aineiston koko ei vastaa täysin Suomen väestöä ja syöpätapauksia generoituu suhteessa eri määrä kuin väestössä. Liian suuri aineisto olisi hidastanut laskentaa ja toisaalta liian pieni syöpäriski hankaloittanut estimointia kohtuuttomasti.

Simulaatiokokeilla varmistettiin asetelmakorjauksen toimivuus ja se, että malli antaa oikeat tulokset. Parametrit estimoituivat simulaatiokokeissa erittäin hyvin, vaikka vaste on erittäin harvinainen ja sen vaihtelu oli jaettu kolmeen eri satunnaiskomponenttiin.

Nuorten naisten rintasyövän on esitetty olevan periytyvää tyyppiä ja oli odotettavissa, että siinä havaittaisiin perheittäistä kasautumista (Dong ja Hemminki, 2001). Myös ruoansulatuselinten syöpien kasautumisessa on todettu olevan geneettinen komponetti (Koskenvuo *et al.*, 2016). Pohjoismaiden laajuisesta kaksostutkimuksesta on raportoitu rintasyövän perimäosuudeksi 31% (95% CI: 11-51%) (Mucci *et al.*, 2016), joka on hieman korkeampi kuin tässä tutkielmassa. Sama tutkimus raportoi vatsan, paksusuolen ja peräsuolen perimäosuudeksi 14%-22%, joka puolestaan on huomattavasti pienempi verrattuna tämän tutkielman tuloksiin.

Perheittäisestä kasautumisesta voidaan tulkita myös geneettisen vaikutuksen luonnetta. Dominantti vaikutus nähdään syöpäsairaiden jälkeläisten kohonneena ilmaantuvuutena. Resessiivinen vaikutus, joka usein siirtyy sukupolven yli, havaitaan sairaan henkilön sisaruksien kohonneena riskinä, kun vanhemmat ovat terveitä (Dong ja Hemminki, 2001). Tulosten analyysin perusteella molemmissa aineistoissa havaittiin selvä ilmiö, jossa jälkeläisen syöpäriski oli väestöä suurempaa, kun ainakin toisella vanhemmista oli havaittu

syöpä. Tutkimushenkilöiden lasten syöpäriski on suuri, kuten myös indeksihenkilöiden sisaruksilla, joilla oli syöpäsairas vanhempi. Tämä viittaa siihen, että syöpä periytyy jälkeläiselle dominantisti.

Tuloksien sensitiivisyyttä arvioitiin kiinnittämällä SIR joko lukuun 1 tai aineiston mallintamattomaan SIR:n estimaattiin informatiivisella priorijakaumalla. Tarkoituksena oli pakottaa suhteellisen ilmaantuvuuden ylimääräinen vaihtelua ympäristön, perimän ja ylihajonnan komponentteihin. Ruoansulatuselinten syövissä havaittiin, että kiinnittäessä SIR:n estimaatti sen vaihtelu siirtyi lähes yksinomaan geneettiseen komponenttiin, jolloin perimäosuus pieneni huomattavasti 22% (0-44%). Ruoansulatuselinten syöpien perimäosuuksien estimoinnissa havaittu pieni SIR:n estimaatti saattaa viitata siihen, että asetelmakorjauksen jälkeen aineistoon on jäänyt korkean syöpäriskin perheitä, jolloin perimän komponentti selittää vaihtelua hyvin. Puolestaan rintasyövässä satunnaismuuttujien hajonnat pienivät samassa suhteessa ja perimäosuuden estimaatti säilyi ennallaan.

Mallin 12 kaltaisissa sekamalleissa perimän komponentti yliestimointuu helposti, sillä kaikkea ympäristön vaikutusta ei usein tunneta tai huomioida mallissa (Gjessing ja Lie, 2008). Tässä tutkielmassa ympäristön vaikutus oletettiin yhtä suureksi koko perheelle, jolloin esimerkiksi vanhempien ja lasten erilaiset lapsuusympäristöt voivat identifioitua perimän komponenttin.

Tutkielmassa arvioitiin ensimmäistä kertaa Suomessa nuoruusiän rinta- ja ruoansulatuelinten syöpien perimäosuutta syöpärekisteriin perustuvassa perheaineistossa. Perimäosuuden estimointi toteutettiin suhteellisen ilmaantuvuuden tilastollisella mallilla valikoituneessa syöpäperheaineistossa. Jatkossa tätä menetelmää voidaan hyödyntää myös muiden syöpien perimäosuuksien estimointiin.

# 9 Ohjelmakoodi

### JAGS ohjelmakoodi

```
model {
  for( i in 1:n.fam ) {
    for( j in offset[i]:(offset[i+1] - 1) ) {

      y[j] ~ dpois(mu[j])
      u[t] ~ dnorm( 0, tau.u)
      od[t]~ dnorm( 0, tau.od)
      log(mu[j]) <- log(e[j]) + b0 +
                    a*f[i] +
                    b*inprod(G[j,1:(offset[i+1]-offset[i])],
                             u[offset[i]:(offset[i+1]-1)]) +
                    c*od[j]
    }
    f[i]  ~ dnorm( 0, tau.f)
  }

  b0  ~ dnorm(0, 0.001)

  a ~ dnorm(0, 0.000001)
  tau.f  ~ dgamma(0.001, 0.001)
  sigma.f2 <- (1/tau.f)*pow(a,2)

  b ~ dnorm(0, 0.000001)
  tau.g  ~ dgamma(0.001, 0.001)
  sigma.g2 <- (1/tau.g)*pow(b,2)

  c ~ dnorm(0, 0.000001)
  tau.od  ~ dgamma(0.001, 0.001)
  sigma.od2 <- (1/tau.od)*pow(c,2)
}
```

# Lähdeluettelo


Bae HT, Perls TT ja Sebastiani P (2014) An efficient technique for bayesian modeling of family data using the bugs software. Frontiers in genetics 5(390).

Breslow NE, Day NE *et al.* (1987) Statistical methods in cancer research, volume 2. International Agency for Research on Cancer Lyon.

Burton PR (2003) Correcting for nonrandom ascertainment in generalized linear mixed models (glmms), fitted using gibbs sampling. Genetic epidemiology 24(1): 24–35.

Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olson JM ja Elston RC (2000) Ascertainment adjustment: where does it take us? The American Journal of Human Genetics 67(6): 1505–1514.

Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW ja Palmer LJ (1999) Genetic variance components analysis for binary phenotypes using generalized linear mixed models (glmms) and gibbs sampling. Genetic epidemiology 17(2): 118–140.

Demidenko E (2005) Mixed Models; Theory and Applications. Wiley.

Dobson AJ ja Barnett A (2008) An introduction to generalized linear models. CRC press.

Dong C ja Hemminki K (2001) Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. International journal of cancer 92(1): 144–150.

Fisher R (1934) The effect of methods of ascertainment upon the estimation of frequencies. Annals of eugenics 6(1): 13–25.

Gelman A ja Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Gelman A ja Rubin DB (1992) Inference from iterative simulation using multiple sequences. Statistical science 7(4): 457–472.

Gelman A *et al.* (2005) Bayesian Data Analysis, second edition. Chapman et. Hall.

Gjessing HK ja Lie RT (2008) Biometrical modelling in genetics: are complex traits too complex? Statistical methods in medical research 17(1): 75–96.

Holford TR (1980) The analysis of rates and of survivorship using log-linear models. Biometrics 36(2): 299–305.

Jamsen KM, Zaloumis SG, Scurrah KJ ja Gurrin LC (2012) Specification of generalized linear mixed models for family data using markov chain monte carlo methods. Journal of Biometrics & Biostatistics 2013(4).

Keiding N (1990) Statistical inference in the lexis diagram. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 332(1627): 487–509.

Khoury MJ, Beaty TH ja Cohen BH (1993) Fundamentals of genetic epidemiology, volume 22. Oxford University Press, USA.

Koistinen P (2013) Course material: Computational statistics.

Koskenvuo L, Pitkäniemi J, Rantanen M ja Lepistö A (2016) Impact of screening on survival in familial adenomatous polyposis. Journal of clinical gastroenterology 50(1): 40–44.

Lange K (2003) Mathematical and statistical methods for genetic analysis. Springer Science & Business Media.

Lévesque LE, Hanley JA, Kezouh A ja Suissa S (2010) Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. Bmj 340: b5087.

McCullagh P ja Nelder JA (1989) Generalized linear models, volume 37. CRC press.

Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, Graff RE, Holst K, Möller S, Unger RH *et al.* (2016) Familial risk and heritability of cancer among twins in nordic countries. JAMA 315(1): 68–76.

Plummer M *et al.* (2003) Jags: A program for analysis of bayesian graphical models using gibbs sampling. Proc. Proceedings of the 3rd international workshop on distributed statistical computing, Technische Universit at Wien Wien, Austria, 124: 125.

Scurrah KJ, Palmer LJ ja Burton PR (2000) Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (glmms) and gibbs sampling in bugs. Genetic epidemiology 19(2): 127–148.

Spiegelhalter D, Thomas A, Best N ja Gilks W (1996) Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK 1–59.

Therneau T, Atkinson E, Sinnwell J, Matsumoto M, Schaid D ja McDonnell S (2012) kinship2: Pedigree functions. R package version 1(7).

Thomas DC *et al.* (2004) Statistical methods in genetic epidemiology. Oxford University Press.