Polygeeniset riskisummat SBayesRC-menetelmällä – sisäisten verenvuotojen ennustaminen verenohennuslääkitystä saavilla yksilöillä

Esa Satomaa

Pro gradu -tutkielma
Toukokuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
Kari Auranen, Joni Virta, Aleksi Winstén

TURUN YLIOPISTO, Matematiikan ja tilastotieteen laitos

Pro gradu -tutkielma
Pääaine: Tilastotiede
Tekijä: Esa Satomaa
Otsikko: Polygeeniset riskisummat SBayesRC-menetelmällä – sisäisten verenvuotojen ennustaminen verenohennuslääkitystä saavilla yksilöillä
Ohjaaja: Kari Auranen, Joni Virta, Aleksi Winstén
Sivumäärä: 65 sivua + liitteet 15 sivua
Aika: Toukokuu 2026

Tutkielma käsittelee polygeenista prediktiota eli ihmisen perimätiedon hyödyntämistä tilastollisessa mallintamisessa. Tutkielmassa esitetään polygeenisen prediktion keskeiset käsitteet, kuten perimän rakenne, SNP eli snippi perimämuuttujana, GWAS-tutkimukset, joissa estimoidaan snippien ja vastetapahtuman väliset assosiaatiot, sekä polygeeninen riskisumma, joka kuvaa yhdellä luvulla yksilön perinnöllistä alttiutta kohdata vastetapahtuma.

Tutkielmassa esitellään erityisesti yksi polygeenisen prediktion menetelmä. SBayesRC on uusi polygeenista prediktiota varten kehitetty menetelmä, jolla estimoidaan GWAS-tutkimuksista saatujen GWAS-tunnuslukujen avulla snippien yhteisvaikutukset, joita käytetään polygeenisten riskisummien laskemiseen. Tutkielmassa selvitetään SBayesRC-menetelmän tilastoteoreettista pohjaa, kuten menetelmän taustalla olevaa hierarkkista Bayes-mallia, sekä menetelmässä käytettävää Gibbsin otantaa, jolla mallin parametrit estimoidaan.

Tutkielman sovelluksessa selvitetään polygeenisten riskisummien hyödyllisyyttä kallon- ja suolistonsisäisen verenvuodon ennustamisessa verenohennuslääkitystä käyttävillä yksilöillä. SBayesRC-menetelmää käytetään polygeenisten riskisummien laskemiseksi FinnGen-rekisterin verenohennuslääkitystä käyttävillä yksilöillä, kun vastetapahtumana on joko kallon- tai suolistonsisäinen verenvuoto. Riskisummat lasketaan SBayesRC- ja PRS-CS-menetelmillä käyttäen kahden eri GWAS-tutkimuksen tunnuslukuja. FinnGen-rekisteridatan perusteella muodostetaan elinaika-aineistot kallon- ja suolistonsisäiselle verenvuodolle, ja polygeenisten riskisummien yhteyttä aikaan kohdata verenvuototapahtuma analysoidaan Coxin regressiomallilla. Riskisumman tuomaa lisää mallin ennustetarkkuuteen tarkastellaan käyttäen bootstrap-validoitua concordance- eli C-indeksiä. Vaikka joissakin tapauksissa riskisumma olikin tilastollisesti merkitsevässä yhteydessä aikaan kohdata verenvuototapahtuma (p-arvo < 0.1), ei riskisumman käyttäminen malleissa selittävänä muuttujana parantanut mallien ennustetarkkuutta.

Asiasanat: SBayesRC, polygeeninen riskisumma, GWAS, kallonsisäinen verenvuoto, suolistonsisäinen verenvuoto, FinnGen.

Sisällys

1 Johdanto

Kun ymmärrys ihmisen geeniperimästä on lisääntynyt, on luonnollisesti noussut esiin kysymys, kuinka tätä tietoa voisi hyödyntää lääketieteessä ja terveydenhuollossa. Jos esimerkiksi tietty geneettinen variaatio ihmisen perimässä on yhteydessä tietyn sairauden puhkeamisen todennäköisyyteen, voisi yksilöiden perimätietoa ehkä hyödyntää sairauden riskiryhmän kartoittamisessa. Tällaisia genotyyppien ja fenotyyppien välisiä yhteyksiä onkin selvitetty genominlaajuisilla assosiaatiokartoituksilla eli GWAS-tutkimuksilla, joissa estimoidaan eri geneettisten varianttien yhteyden voimakkuutta valittuun fenotyyppiin tilastollisia menetelmiä käyttäen [32].

Yksilöille voidaan laskea heidän perimätietojensa ja GWAS-tutkimuksen tulosten perusteella polygeeninen riskisumma (*polygenic risk score, PRS*), jonka tarkoituksena on kertoa yksilön geneettisestä alttiudesta kohdata tietty vastetapahtuma tai sairaus [7]. Polygeenisen riskisumman laskemista varten on kehitety useita menetelmiä, kuten PRS-CS [14], AnnoPred [17] ja LDpred2 [25]. Tämä tutkielma käsittelee yhtä melko uutta tällaista menetelmää nimeltä SBayesRC, josta julkaistiin artikkeli vuonna 2024 [38]. Tutkielmassa esitellään polygeenisen prediktion – eli perimään pohjautuvien ennusteiden tekemisen – oleelliset taustatiedot ja polygeenisten riskisummien laskemisen periaatteet. Tämän jälkeen käsitellään SBayesRC-menetelmää ja sen tilastoteoreettista pohjaa. Lopuksi menetelmää sovelletaan laskemalla polygeeniset riskisummat suomalaisen FinnGen-aineiston verenohennuslääkitystä käyttävälle osajoukolle, kun vastetapahtumana on joko kallonsisäinen tai suolistonsisäinen verenvuoto. Riskisummia käytetään tämän jälkeen selittävinä muuttujina elinaikaanalyysissä, jossa vastetapahtumana on aika kallon- tai suolistonsisäiseen verenvuotoon verenohennuslääkitystä käyttävillä yksilöillä. Tutkimuskysymyksenä on, lisääkö polygeenisten riskisummien käyttäminen mallien ennustetarkkuutta. Mallia, jossa selittävinä muuttujina ovat polygeeninen riskisumma sekä muut kovariaatit, verrataan muuten samaan malliin, mutta ilman polygeenista riskisummaa. Tutkielman kuvat 1, 2, 3 ja 4 on tehty hyödyntäen tekoälyä (ChatGPT versiot 5–5.2).

2 Polygeeninen prediktio

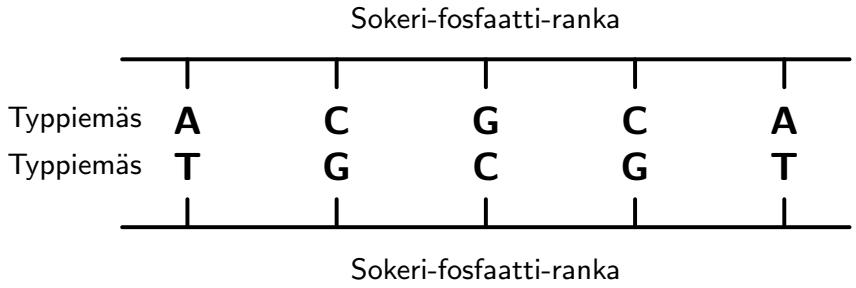
Polygeenisella prediktiolla tarkoitetaan eri vastemuuttujiin liittyvien ennusteiden tekemistä ihmisten perimätietojen pohjalta. Poly-etuliite viittaa siihen, että vastemuuttujia ennustetaan hyödyntäen suurta määrää – mahdollisesti miljoonia – perimämuuttujia laajasta osasta ihmisen perimää. Perimään pohjautuvan alttiuden kohdata vastetapahtuma katsotaan johtuvan näiden lukuisten osatekijöiden yhteisvaikutuksesta. Vastaavasti monogeenisyydellä viitataan siihen, kun ominaisuus tai sairaus johtuu vain yhdestä geenistä. Tässä luvussa käydään läpi polygeenisen prediktion perustiedot alkaen perimän rakenteesta ja päätyen polygeenisen riskisumman laskemiseen.

2.1 Perimän rakenne

Vaikka tutkielman pääasiallisena mielenkiinnon kohteena ei olekaan perimän taustalla oleva biologia, seuraavaksi esitetään lyhyesti perimän biokemiallinen rakenne

kokonaisvaltaisemman käsityksen saamiseksi. Ihmisen perimä eli genomi on pitkä sarja deoksiribonukleiinihappoa eli DNA:ta [20]. Se sisältää kaiken ihmisen periytyvän informaation ja toimii perinnöllisyyden pohjana. Perimä pitää sisällään monta eri DNA-molekyyliä eli kromosomia, mutta myös ihmissolujen mitokondrioiden sisältämän DNA:n.

DNA on lineaarinen ja ei-haarautuva molekyyli, joka koostuu nukleotideista [5]. Nukleotideissa on kolme komponenttia: typpiemäs, pentoosisokeri ja fosfaattiryhmä. Monien nukleotidien muodostamaa ketjua kutsutaan polynukleotidiksi. Vuorottelevat pentoosisokerit ja fosfaattiryhmät muodostavat sokeri-fosfaatti-rangan, jonka sokereihin typpiemäkset ovat kiinnittyneet. Nukleotidin typpiemäs on aina yksi neljästä: sytosiini (C), tymiini (T), adeniini (A) tai guaniini (G). DNA:n rakenne on kaksoisjuoste, joka koostuu kahdesta tällaisesta polynukleotidista. Polynukleotidien typpiemäkset ovat pareittain sitoutuneet toisiinsa typpisidoksin, emäsparit ovat aina joko adeniini-tymiini-pareja (A-T tai T-A) tai sytosiini-guaniini-pareja (C-G tai G-C). Kuvassa 1 on esitetty DNA-ketjun rakenne.



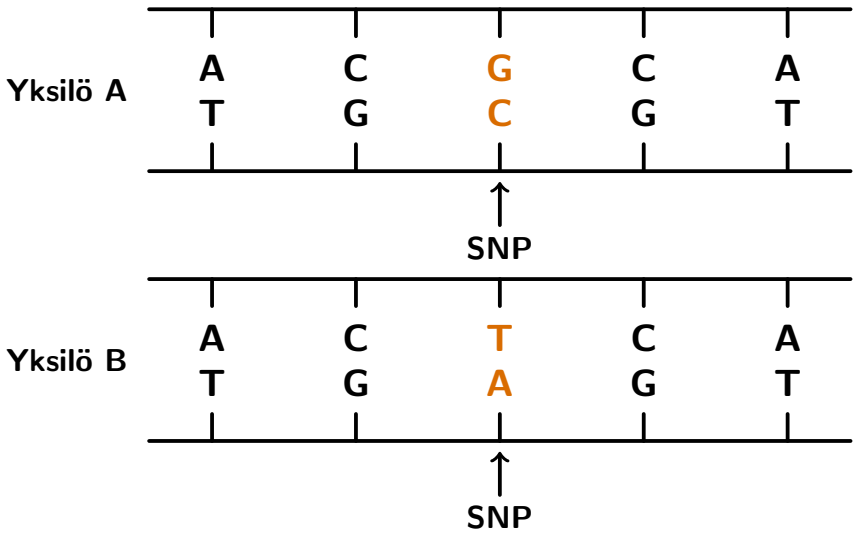
Kuva 1: DNA:n rakenne.

Geenit ovat jaksoja DNA:n kaksoisjuosteessa [20]. Ne ovat perimän toiminnallisia yksiköitä, jotka osallistuvat RNA-juosteiden (toinen nukleiinihappo) ja proteiinien tuottamiseen. Tietty geeni sijaitsee aina tietyssä kohdassa kromosomia, ja tästä sijainnista käytetään nimitystä lokus. Lokuksessa sijaitseva geeni ei ole aina samanlainen, vaan samasta geenistä voi olla vaihtoehtoisia versioita eli alleeleja. Tämä variaatio koskee sitä, mitä neljästä typpiemäksestä geenissä esiintyy ja missä järjestyksessä. Käsitteitä lokus ja alleeli voidaan käyttää myös yhdestä nukleotidista, jolloin lokus on nukleotidin sijainti perimässä ja alleeli on kyseisessä sijainnissa oleva typpiemäs, esimerkiksi A tai G.

2.2 Yhden nukleotidin polymorfismi

Yhden nukleotidin polymorfismi (*SNP, single-nucleotide polymorphism*) eli snippi on sellainen sijainti perimässä, jossa yhdessä nukleotidissa esiintyy vaihtelua populaation yksilöiden välillä [5]. Tällaisessa lokuksessa joillain ihmisillä voi olla esimerkiksi C-typpiemäs ja toisilla A-typpiemäs, jolloin snipillä on kaksi alleelia eli vaihtoehtoista versiota. Kuva 2 kuvaa tällaista tilannetta, snipin alleelit poikkeavat yksilöillä A ja B. Vaikka teoriassa onkin mahdollista, että yhdellä snipillä on neljä eri alleelia (A, T, C ja G), on käytännössä selvästi suurimmassa osassa snippejä alleeleja vain kaksi.

Tämä johtuu siitä, että tapahtumasarja, jossa yksilön nukleotidi pistemutatoituu, siirtyy hänen jälkeläisilleen ja sukupolvien saatossa yleistyy populaatiossa on varsin harvinainen. Todennäköisyys, että vastaava tapahtumasarja tapahtuu uudestaan samassa nukleotidissa aiheuttaen kolmannen variantin yleistymisen populaatiossa, on ymmärrettävästi hyvin pieni, ja siksi suurin osa snipeistä on kaksivarianttisia. GWAS-tutkimuksissa käytetyt geneettiset markkerit (*genetic marker*), joiden assosiaatiota vastemuuttujaan tutkitaan, ovat käytännössä aina snippejä.



Kuva 2: Esimerkki snipistä.

Koska A-typpiemäs on aina parina T-typpiemäkselle ja G-typpiemäs C-typpiemäkselle, perimän rakennetta kartoittaessa tai muuten perimää tarkasteltaessa riittää käsitellä vain toista DNA-juosteista. Kun tiedetään yhden juosteen nukleotidien alleelit, määräytyvät toisen juosteen alleelit implisiittisesti tästä yhteydestä. Jos siis sanotaan esimerkiksi "snipin j alleeli on A", on toinen kromosomin DNA-juosteista valittu referenssijuosteeksi ja ilmaisulla tarkoitetaan, että "snipin j alleeli on A referenssijuosteen nukleotidissa". Myös tässä tutkielmassa puhutaan jatkossa lyhyesti snipin alleelista viitaten referenssijuosteen alleeliin.

2.3 Genotyyppimatriisi

Genotyypillä eli perimätyypillä tarkoitetaan yksilön geneettistä kokonaisuutta, toisin sanoen sitä, millainen perimä yksilöllä käytännössä on. Seuraavaksi esitetään, kuinka yksilön genotyyppi koodataan genotyyppimatriisiin $\mathbf{X}$.

GWAS-tutkimuksissa ja polygeenisia riskisummia laskettaessa genotyyppidata koskee joukkoa snippejä, olkoon niitä m kappaletta. Kuten aikaisemmin mainittiin, snipillä on lähes aina kaksi vaihtoehtoista alleelia. GWAS-tutkimuksessa estimoidaan alleelien yhteys vastemuuttujaan, toinen alleeleista toimii vaikutusalleelina (*effect allele*) ja toinen verrokkina. Periaatteessa kumpi tahansa alleeleista voidaan valita vaikutusalleeliksi, mutta usein siksi valitaan ei-referenssiperimän mukainen alleeli.

Tällöin ollaan kiinnostuttu referenssiperimästä poikkeavan perimävariantin yhteydestä vastemuuttujaan. Ihmisen referenssiperimä on kartoitettu esimerkiksi Human Genome -projektissa [34].

Ihminen on diploidi eliö, eli ihmisellä on kaksi parittaista joukkoa kromosomeja, joista toisen hän perii äidiltään ja toisen isältään [20]. Tämä tarkoittaa, että jokaisen snipin kohdalla yksilöllä voi olla vaikutusalleelista joko 0 kopiota (hän ei ole perinyt kyseistä alleelia kummaltakaan vanhemmaltaan), 1 kopio (hän on perinyt kyseisen alleelin vain toiselta vanhemmaltaan) tai 2 kopiota (hän on perinyt kyseisen alleelin molemmilta vanhemmiltaan). Tämä vaikutusalleelin kopioiden määrä on koodattu genotyyppimatriisiin, usein koodaus on $0/1/2$, jossa luku vastaa suoraan kopioiden määrää, mutta myös toisenlainen koodaus on mahdollinen, esimerkiksi $-1/0/1$ [31]. Tässä tutkielmassa koodauksen oletetaan olevan $0/1/2$. Jos genotyyppimatriisi $\mathbf{X}$ koskee yksilöitä, joiden lukumäärä on n, se on $(n \times m)$ matriisi, jossa rivit vastaavat yksilöitä ja sarakkeet snippejä. Matriisin alkio x_{ij} kertoo siis, kuinka monta snipin j vaikutusalleelin kopiota yksilöllä i on: 0, 1 tai 2.

2.4 GWAS

Genominlaajuinen assosiaatiokartoitus (*genome-wide association study*) eli GWAS on metodi, joka pyrkii löytämään genomin laajuudelta ne geneettiset markkerit, jotka ovat yhteydessä johonkin fenotyyppiin, usein johonkin sairauteen [5]. Markkerit ovat käytännössä aina snippejä.

Tyypillisesti assosiaatiota etsitään lineaarisella tai logistisella regressiomallilla [32]. Lineaarista regressiota käytetään, kun vastemuuttuja eli fenotyyppi on jatkuva (esim. verenpaine) ja logistista regressiota kun vastemuuttuja on binäärinen (esim. sairauden olemassaolo). Assosiaatiota testataan erikseen jokaiselle snipille.

Olkoon $\mathbf{Y}$ vastemuuttujan arvojen vektori, $\mathbf{W}$ kovariaattiarvojen matriisi sisältäen vakiotermisarakkeen ja $\mathbf{X}_j$ snipin j vaikutusalleelin arvojen vektori GWAS-aineiston yksilöille (vastaa genotyyppimatriisin $\mathbf{X}$ saraketta). Tällöin vektorien $\mathbf{Y}$ ja $\mathbf{X}_j$ pituus on GWAS:n otoskoko n ja matriisin $\mathbf{W}$ dimensio on $n \times (l+1)$, jossa l on kovariaattien lukumäärä. GWAS:n lineaarinen regressiomalli jatkuva-arvoisen fenotyypin ja snipin j assosiaation testaamiseksi voidaan kirjoittaa muodossa

$$\begin{aligned} \mathbf{Y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{X}_j b_j + \boldsymbol{g} + \boldsymbol{e}, \\ \boldsymbol{g} \sim N(\mathbf{0}, \sigma_A^2 \boldsymbol{\Psi}), \\ \boldsymbol{e} \sim N(\mathbf{0}, \sigma_e^2 \boldsymbol{I}). \end{aligned}$$

Vektori $\boldsymbol{\alpha}$ sisältää kovariaattien vaikutukset ja b_j on vaikutusalleelin vaikutus. Satunnaisvaikutusten n-vektori $\boldsymbol{g}$ sisältää muiden geenivarianttien polygeenisen vaikutuksen. Se voidaan sisällyttää malliin selittämään yksilöiden välistä geneettistä sukulaisuutta. Varianssi σ_A^2 kuvaa fenotyypin additiivista geneettistä variaatiota ja $\boldsymbol{\Psi}$ on geneettisten yhteyksien matriisi, joka kuvaa yksilöiden välistä geneettistä samankaltaisuutta. Lisäksi n-vektori $\boldsymbol{e}$ on residuaalivektori ja σ_e^2 on residuaalivarianssi. Mallissa on usein kovariaatteja, kuten sukupuoli ja ikä, vastaamassa osituksesta ja ehkäisemässä demografisten tekijöiden sekoittavia vaikutuksia. Keskeinen syy siihen, miksi kaikkien snippien assosiaatioita ei testata kerralla, on snippien suuri määrä.

Koska snippejä voi olla tutkimuksessa useita miljoonia, olisi mallissa moninkertainen määrä muuttujia verrattuna otoskokoon. Koska jokaisen snipin j kohdalla mallissa on mukana vain kyseinen snippi, ovat vaikutukset b_j snippien marginaalivaikutuksia.

Edellä mainittu geneettisten yhteyksien matriisi (*GRM, genetic relationship matrix*) $\mathbf{\Psi}$ voidaan estimoida seuraavalla yhtälöllä [37]:

$$\Psi_{ik} = \frac{1}{N} \sum_{j=1}^{m} \frac{(x_{ij} - 2p_j)(x_{kj} - 2p_j)}{2p_j(1 - p_j)}.$$

Yllä Ψ_{ik} on geneettinen yhteys yksilöiden i ja k välillä. Indeksi j käy läpi kaikki m snippiä, x_{ij} ja x_{kj} ovat yksilöiden i ja k vaikutusalleelien kopioiden määrät snipille j ja p_j on snipin j vaikutusalleelin suhteellinen frekvenssi.

GWAS-tutkimuksen tulokset kootaan yleensä kokonaisuudeksi, josta käytetään nimitystä GWAS-tunnusluvut (*GWAS summary statistics*) [32]. Nämä tunnusluvut sisältävät mm. testattujen snippien tunnukset ja niiden sijainnit perimässä, snippien estimoidut marginaalivaikutukset sekä vaikutusten keskivirheet ja p-arvot.

2.5 Kytkentäepätasapaino

Kytkentäepätasapaino (*Linkage Disequilibrium, LD*) tarkoittaa eri lokuksissa sijaitsevien alleelien välistä ei-satunnaista assosiaatiota [29]. LD voidaan määritellä usealla eri tavalla, mutta kaikkien määritelmien taustalla on suure

$$D_{AB} = p_{AB} - p_A p_B,$$

jossa p_A ja p_B ovat alleelien A ja B esiintyvyydet populaatiossa ja p_{AB} on molempien alleelien yhtäaikainen esiintyvyys populaatiossa. Tässä A ja B sijaitsevat eri lokuksissa. Kun $D_{AB} = 0$, vallitsee alleelien välillä kytkentätasapaino (*linkage equilibrium*), ja muussa tapauksessa kytkentäepätasapaino eli LD. Yksi yleisesti käytetty tapa kvantifioida LD on suureella r^2:

$$r^2 = \frac{D_{AB}^2}{p_A(1 - p_A)p_B(1 - p_B)}.$$

Tämä on alleelien A ja B läsnäolosta kertovien indikaattorimuuttujien (1/0) välinen neliöity korrelaatiokerroin.

Useat polygeenisessa prediktiossa käytettävät menetelmät, kuten PRS-CS [14] ja myöhemmin tarkemmin käsiteltävä SBayesRC [38], tarvitsevat syötteenä LD-referenssimatriisin tai -referenssipaneelin. Vaikka määritelmällisesti LD ei olekaan synonyymi alleelien väliselle korrelaatiolle, vaikuttaa ero näiden välillä olevan käytännössä hyvin pieni. Esimerkiksi Reich ja kumppanit sanovat artikkelissaan LD:llä tarkoitettavan lähekkäisten alleelien välistä korrelaatiota [27] ja Zheng ja kumppanit puhuvat artikkelissaan LD-korrelaatiomatriisista [38].

2.6 Funktionaalinen annotaatio

Yksi tässä tutkielmassa käsiteltävän SBayesRC-metodin huomattavista ominaisuuksista on funktionaalisten annotaatioiden hyödyntäminen [38]. Perimäannotaatio (*genome annotation*) on prosessi, jossa selvitetään DNA-sekvenssin biologista merkitystä [30]. Annotaatioissa perimän kontekstissa on siis kyse käytännön merkitysten antamisesta perimän osatekijöille kuten geeneille.

Perimäannotaatio on jaettavissa kahteen toisiinsa liittyvään tyyppiin: rakenteelliseen ja funktionaaliseen annotaatioon. Rakenteellinen annotaatio käsittää perimän osatekijöiden kuten geenien kuvailun ja rajaamisen, ja funktionaalinen annotaatio käsittää tehtävän tai toiminnallisuuden määrittelemisen näille osatekijöille. [4]

Edellä esitetty tapa kategorisoida perimäannotaatio ei ole ainoa mahdollinen, esimerkiksi Reed ja kumppanit [26] käsittelevät artikkelissaan moniulotteista perimäannotaatiota (*multidimensional genome annotation*), jossa yhden ulottuvuuden tapaus vastaa edellä esitettyä. Joka tapauksessa funktionaalisella annotaatiolla viitataan tietoon perimän osatekijöiden, kuten geenien tai snippien, toiminnallisuudesta.

Jatkossa SBayesRC-menetelmässä käytettävää annotaatiodataa merkitään matriisilla $\mathbf{A}$. Matriisin dimensio on $m \times C$, jokainen rivi vastaa yhtä snippiä (m kappaletta) ja sarakkeet vastaavat annotaatioita (C kappaletta). Esimerkkinä $\mathbf{A}$ voisi sisältää sarakkeen eli annotaatiomuuttujan, joka kertoo, sijaitseeko snippi lähimmän geenin koodaavalla alueella. Tällöin annotaatiomuuttuja olisi binäärimuuttuja, jossa arvo 1 tarkoittaa, että snippi sijaitsee lähimmän geenin koodaavalla alueella, ja arvo 0, että ei sijaitse.

2.7 Polygeeninen riskisumma

Polygeeninen riskisumma (*PRS, polygenic risk score*) on yksilön perimän ja sopivien GWAS-tunnuslukujen pohjalta laskettu estimaatti yksilön geneettiselle alttiudelle johonkin piirteeseen (*trait*) kuten sairauteen [7]. Kuten nimi antaa ymmärtää, geneettisen alttiuden oletetaan perustuvan monien geneettisten markkereiden eli snippien yhteisvaikutukseen.

Klassisessa PRS-metodissa yksilön i PRS lasketaan painotettuna summana

$$PRS_i = \sum_{j=1}^{m} x_{ij} b_j.$$

Yksilön vaikutusalleelien kopioiden määrät x_{ij} summataan yhteen painotettuina vastaavien alleelien marginaalivaikutuksilla b_j, jotka on estimoitu aiemmin vastaavan fenotyypin GWAS-tutkimuksessa [7]. PRS:n laskemiseksi on kehitetty myös monimutkaisempia menetelmiä. Choi, Mak ja O'Reilly listaavat keskeisiksi tekijöiksi PRS:n laskemiseen tarkoitettuja metodeja kehitettäessä 1) mahdollisen GWAS:n vaikutusten säätämisen esimerkiksi kutistamisella (*shrinkage*), 2) PRS:n kohdistamisen tietylle populaatiolle ja 3) LD:n eli snippien välisen korrelaation huomioonottamisen [7]. Esimerkkinä kohdista 1) ja 3) on tutkielmassa tarkemmin käsiteltävä metodi SBayesRC, jossa GWAS-tutkimuksesta saatujen snippien marginaalivaikutuksien avulla estimoidaan snippien yhteisvaikutukset [38].

Polygeeninen riskisumma on polygeenisen prediktion lopputuote. Se on yksilölle laskettava suure, joka yhdellä luvulla kuvaa hänen geneettistä alttiuttaan johonkin piirteeseen kuten sairauteen. Keskeinen kliininen sovellus PRS-muuttujalle on, että sitä voitaisiin käyttää sairauksien riskiryhmien kartoittamisessa. Jos yksilöllä on PRS-muuttujan perusteella korkea perinnöllinen alttius johonkin sairauteen, voitaisiin tämä ottaa huomioon hoitosuunnitelmaa tehdessä tai ohjata yksilö ennaltaehkäisevän toiminnan pariin. Eri kysymys on, kuinka hyvin PRS-muuttuja käytännössä toimii vastetapahtuman ennustamisessa verrattuna muihin mahdollisiin selittäviin tekijöihin. Tämän tutkielman myöhemmässä sovellusosassa etsitään vastausta tähän kysymykseen kallonsisäisen ja suolistonsisäisen verenvuodon tapauksessa.

2.8 Polygeenisen prediktion yhteenveto

Suurin osa ihmisen perimästä koostuu DNA-molekyyleistä eli kromosomeista, jotka sisältävät pitkiä emäsparisekvenssejä. Osassa näistä emäspareista esiintyy populaatiossa variaatiota, tällaisia sijainteja kutsutaan snipeiksi. GWAS-tutkimuksissa estimoidaan snippien tiettyjen realisaatioiden eli alleelien ja jonkin vastemuuttujan välinen yhteys, ja tutkimuksen tulokset julkaistaan GWAS-tunnuslukujen muodossa. GWAS-tunnusluvuista saatavia snippien marginaalivaikutuksia voidaan vielä jatkokäsitellä jonkin sopivan menetelmän avulla, esimerkiksi estimoimalla niiden avulla snippien yhteisvaikutukset. Tällöin tarvitaan usein myös tietoa snippien välisestä kytkentäepätasapainosta eli korrelaatiosta tai snippien funktionaalisista annotaatioista, jotka sisältävät ylimääräistä tietoa snippien eri aspekteista. Lopulta GWAS-tunnuslukujen ja yksilön perimädatan eli genotyypin avulla yksilölle voidaan laskea polygeeninen riskisumma, joka kuvaa yksilön perinnöllistä taipumusta tiettyyn piirteeseen kuten johonkin sairauteen. Toiveena on, että polygeenista riskisummaa voidaan hyödyntää esimerkiksi sairauden riskiryhmän kartoittamisessa.

3 SBayesRC

Seuraavaksi tutkielmassa syvennytään tarkemmin yhteen melko uuteen polygeenisessa prediktiossa hyödynnettävään menetelmään. SBayesRC on Zhili Zhengin ja kumppanien kehittämä polygeenisen prediktion menetelmä, jolla voidaan estimoida snippien yhteisvaikutukset käyttämällä GWAS-tutkimuksessa estimoituja snippien marginaalivaikutuksia [38]. Snippien yhteisvaikutukset toimivat marginaalivaikutuksia paremmin polygeenisten riskisummien laskennassa, koska ne ottavat huomioon snippien korrelaatiorakenteen. Menetelmällä voidaan estimoida myös muita geneettiseen arkkitehtuuriin liittyviä parametreja.

Tilastollinen malli, johon SBayesRC-menetelmä pohjautuu, on Bayes-malli, ja malliparametrien estimointi tapahtuu Markov Chain Monte Carlo -otannalla eli MCMC-otannalla [38]. MCMC-otanta on menetelmäperhe, jossa tuntematonta jakaumaa tutkitaan tekemällä siitä numeerinen otos Markovin ketju -prosessin avulla. Bayes-päättelyssä estimoidaan parametrien posteriorijakauma, ja tämän jakauman tunnuslukuja voidaan laskea MCMC-otoksen perusteella. SBayesRC-menetelmässä käytetään Gibbsin otantaa, jossa MCMC-otanta perustuu mallin parametrien täysehdollisiin jakaumiin.

Metodi ottaa syötteenä kolme asiaa: 1) GWAS-tunnusluvut, 2) LD-referenssimatriisin ja 3) funktionaalisen annotaatiodatan [38]. GWAS-tunnusluvut sisältävät seuraavat tiedot snipeistä: snippitunnuksen, vaikutusalleelin, vaihtoehtoisen alleelin, vaikutusalleelin suhteellisen osuuden populaatiossa, vaikutusalleelin marginaalivaikutuksen, keskivirheen ja p-arvon, sekä per-SNP otoskoot. Funktionaalisten annotaatioiden käyttäminen ei ole välttämätöntä, sillä menetelmää voidaan käyttää myös ilman niitä, jolloin se muistuttaa aikaisempaa SBayesR-menetelmää [21].

Luvussa käsitellään seuraavaksi SBayesRC-menetelmän tilastollisen mallin perustana toimivan matala-asteisen mallin muodostus. Tämän jälkeen siirrytään hierarkkiseen Bayes-malliin, joka saadaan tekemällä matala-asteisen mallin parametreille priorijakaumaoletuksia. Mallin parametrien täysehdolliset jakaumat esitellään, koska näitä hyödynnetään Gibbsin otannassa. Viimeisenä käsitellään Gibbsin otanta-algoritmi, jolla mallin parametrit estimoidaan. Pääasiallisena lähteenä käytettiin Zhengin ja kumppanien menetelmää käsittelevää tutkimusartikkelia ja sen liiteosaa [38]. Itse artikkeliin ei alaluvuissa viitata enää erikseen.

3.1 Matala-asteisen mallin muodostus

Estimointi aloitetaan yksinkertaisesta lineaarisesta regressiomallista

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

jossa $\mathbf{Y}$ on kovariaattien suhteen vakioitu fenotyyppivektori, $\mathbf{X}$ on standardoitu genotyyppimatriisi (sarakkeiden keskiarvo 0 ja varianssi 1) ja $\boldsymbol{\varepsilon}$ on normaalijakautunut residuaalivektori, jonka odotusarvo on $\mathbf{0}$ ja kovarianssimatriisi $\mathrm{Var}(\boldsymbol{\varepsilon})=\mathbf{I}\sigma^2_\varepsilon$. Parametri $\boldsymbol{\beta}$ on snippien yhteisvaikutusten vektori. Kun yksilöiden lukumäärä on n ja snippien määrä on m, matriisin $\mathbf{X}$ dimensio on $n \times m$, vektorin $\boldsymbol{\beta}$ pituus m ja vektoreiden $\mathbf{Y}$ ja $\boldsymbol{\varepsilon}$ pituus n.

Kun yhtälö (1) kerrotaan vasemmalta puolelta matriisilla $\frac{1}{n}\mathbf{X}^{'}$, saadaan malli muotoon

$$\frac{1}{n}\mathbf{X}^{'}\mathbf{Y} = \frac{1}{n}\mathbf{X}^{'}\mathbf{X}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}^{'}\boldsymbol{\varepsilon}.$$

Yhtälön vasen puoli $\frac{1}{n}\mathbf{X}^{'}\mathbf{Y}$ korvataan nyt aiemman GWAS-tutkimuksen perusteella marginaalivaikutuksilla $\mathbf{b}$. Perustelu tähän on annettu Lloyd-Jonesin ja kumppanien artikkelissa [21]: Merkitään matriisin $\mathbf{X}$ sarakkeita $\mathbf{x}_j$. Lineaarisen regressiomallin marginaalivaikutuksen b_j pienimmän neliösumman estimaatti on

$$\hat{b_j} = \frac{\mathbf{x}_j'\mathbf{Y}}{\mathbf{x}_j'\mathbf{x}_j}.$$

Tämä vastaa tilannetta, jossa muuttujavektoria $\mathbf{Y}$ selitetään pelkästään snipillä j. Yhtälössä ei ole mukana vakiokerrointa, koska sarakkeet $\mathbf{x}_j$ ja vektori $\mathbf{Y}$ oletetaan keskistetyiksi. Tällöin kaikkien marginaalivaikutusten vektori $\hat{\mathbf{b}}$ voidaan esittää matriisitulona

$$\hat{\mathbf{b}} = \mathbf{D}^{-1}\mathbf{X}'\mathbf{Y}, \quad \mathbf{D} = \mathrm{diag}\left(\mathbf{x}_1'\mathbf{x}_1, ..., \mathbf{x}_m'\mathbf{x}_m\right).$$

Koska matriisi $\mathbf{X}$ on standardoitu, pätee kaikilla j

$$\frac{1}{n}\mathbf{x}_j'\mathbf{x}_j = 1 \Rightarrow \ \mathbf{x}_j'\mathbf{x}_j = n,$$

eli $\mathbf{D}^{-1} = \frac{1}{n}\mathbf{I}$. Näin saadaan lopulta

$$\frac{1}{n}\mathbf{X}'\mathbf{Y} = \ \mathbf{D}^{-1}\mathbf{X}'\mathbf{Y} = \ \hat{\mathbf{b}}.$$

Koska matriisi $\mathbf{X}$ on standardoitu, on $\frac{1}{n}\mathbf{X}'\mathbf{X}$ sen korrelaatiomatriisi. Se korvataan nyt menetelmälle syötteenä annetulla LD-referenssimatriisilla, olkoon tämä $\mathbf{R}$. Malli (1) voidaan nyt esittää muodossa

$$\mathbf{b} = \mathbf{R}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}. \tag{2}$$

Monet vastaavat menetelmät kuten PRS-CS [14] pohjautuvat tähän tiivistelmädataan perustuvaan malliin (*summary-data-based model*). Mallin (2) residuaaleille pätee

$$\mathrm{Var}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right) = \frac{1}{n}\mathbf{X}'\left(\mathbf{I}\sigma_\varepsilon^2\right)\left(\frac{1}{n}\mathbf{X}'\right)' = \frac{1}{n^2}\mathbf{X}'\mathbf{X}\sigma_\varepsilon^2 = \frac{1}{n}\mathbf{R}\sigma_\varepsilon^2,$$

eli residuaalien kovarianssimatriisi on verrannollinen LD-matriisiin. Täyden LD-matriisin laskeminen ei usein ole käytännössä mahdollista eikä tarpeellistakaan. Sen sijaan snipit jaetaan lohkoihin, joiden on todettu olevan populaatiossa likimäärin toisistaan riippumattomia, ja jokaiselle tällaiselle lohkolle i lasketaan vastaava matriisi $\mathbf{R}_i$. Nyt sopivasti järjestettynä koko perimänlaajuinen LD-matriisi $\mathbf{R}$ voidaan siis olettaa lohkodiagonaaliseksi matriisiksi, joka koostuu L lohkokohtaisesta matriisista:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{R}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{R}_L \end{pmatrix}.$$

Seuraava askel on tehdä jokaisesta lohkokohtaisesta LD-matriisista ominaisarvohajotelma

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}',$$

jossa $\mathbf{U}$ on matriisin $\mathbf{R}$ ominaisvektoreiden matriisi ja $\mathbf{\Lambda}$ on matriisin $\mathbf{R}$ ominaisarvojen diagonaalimatriisi. Tässä lohkoa osoittavasta alaindeksistä on luovuttu notaation yksinkertaistamiseksi. Ominaisarvohajotelma voidaan tehdä lohkoille erikseen. Sijoittamalla malliin (2) ominaisarvohajotelma ja kertomalla vasemmalta puolelta matriisilla $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'$ saadaan

$$\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{b} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{U}\mathbf{\Lambda}\mathbf{U}'\boldsymbol{\beta} + \frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{X}'\boldsymbol{\varepsilon}.$$

Matriisi $\mathbf{U}$ on korrelaatiomatriisin ominaisvektorien matriisina ortogonaalinen, jolloin $\mathbf{U}'\mathbf{U} = \mathbf{U}^{-1}\mathbf{U} = \mathbf{I}$. Lisäksi koska $\mathbf{\Lambda}$ on diagonaalimatriisi diagonaalialkioilla a_i, on matriisi $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}$ diagonaalimatriisi, jonka diagonaalialkiot ovat $\frac{a_i}{\sqrt{a_i}} = \sqrt{a_i}$, eli $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda} = \mathbf{\Lambda}^{\frac{1}{2}}$. Näin ollen malli sieventyy muotoon

$$\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{b} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}'\boldsymbol{\beta} + \frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{X}'\boldsymbol{\varepsilon}.$$

Vektori $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{b}$ koostuu marginaalisten SNP-vaikutusten lineaarikombinaatioista, $\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}'$ on uusi kerroinmatriisi yhteisvaikutuksille $\boldsymbol{\beta}$ ja $\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{X}'\boldsymbol{\varepsilon}$ on uusi residuaalivektori. Antamalla näille merkinnät $\mathbf{w}$, $\mathbf{Q}$ ja $\mathbf{e}$ saadaan malli muotoon

$$\mathbf{w} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{e}. \tag{3}$$

Uusien residuaalien $\mathbf{e}$ kovarianssimatriisi on nyt

$$\begin{aligned}
\mathrm{Var}(\mathbf{e}) &= \frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{X}'\mathbf{I}\sigma^2_\varepsilon\left(\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{X}'\right)' = \frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{X}'\mathbf{I}\sigma^2_\varepsilon\frac{1}{n}\mathbf{X}\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}} \\
&= \sigma^2_\varepsilon\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}} = \sigma^2_\varepsilon\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{R}\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}} \\
&= \sigma^2_\varepsilon\frac{1}{n}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}'\mathbf{U}\mathbf{\Lambda}\mathbf{U}'\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}} = \frac{1}{n}\sigma^2_\varepsilon\mathbf{\Lambda}^{-1}\mathbf{\Lambda} = \frac{1}{n}\mathbf{I}\sigma^2_\varepsilon,
\end{aligned}$$

kun muistetaan, että $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_\varepsilon)$, $\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$ ja $\mathbf{U}'\mathbf{U} = \mathbf{I}$. Residuaalit $\mathbf{e}$ ovat siis riippumattomia ja samoin jakautuneita.

SBayesRC-menetelmän keskeinen osa on seuraavaksi esiteltävä mallin dimensionpienennys. Ominaisarvojen diagonaalimatriisi jaetaan osiin

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_0 \end{pmatrix}.$$

Diagonaalimatriisi $\mathbf{\Lambda}_q$ sisältää ensimmäiset q laskevassa järjestyksessä olevaa ominaisarvoa, jotka yhdessä selittävät LD-matriisin varianssista ainakin osuuden ρ. Diagonaaliatriisi $\mathbf{\Lambda}_0$ sisältää jäljelle jäävät ominaisarvot. Jaottelun avulla malli (3) voidaan kirjoittaa muodossa

$$\begin{bmatrix} \mathbf{w}_q \\ \mathbf{w}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_q \\ \mathbf{Q}_0 \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \mathbf{e}_q \\ \mathbf{e}_0 \end{bmatrix}. \tag{4}$$

Kun mallista (4) jätetään pois osaa $\mathbf{w}_0$ vastaavat yhtälöt, saadaan matala-asteinen malli

$$\mathbf{w}_q = \mathbf{Q}_q\boldsymbol{\beta} + \mathbf{e}_q. \tag{5}$$

Matriisin $\mathbf{Q}_q$ dimesio on $q \times m$, jossa q on huomattavasti pienempi kuin m. Sen sijaan, että todelliset snippien yhteisvaikutukset sovitettaisiin kaikkiin keskenään hyvin korreloituneisiin havaintopisteisiin, ne sovitetaan pienempään määrään informatiivisiä datapisteitä. Jatkossa käsiteltäessä matala-asteista mallia (5) alaindekseistä q luovutaan.

Menetelmän käytön kannalta merkittävä kysymys on, mikä on sopiva valinta parametrin ρ arvoksi. Artikkelissaan Zheng ja kumppanit suorittavat ennen muita analyysejä simulaatioon perustuvan matala-asteisen mallin kalibroinnin, jossa yhtenä kalibroitavana parametrina oli kyseinen ρ. He esittävät, että ennustetarkkuuden kannalta parametrin ρ optimaalinen arvo riippuu periaatteessa GWAS- ja LD-referensseihin käytetystä aineistosta. Siksi Zheng ja kumppanit kehittivät menetelmään alustusvaiheen, jossa optimaalinen parametrin ρ arvo määritetään GWAS-tunnuslukuihin pohjautuvan pseudovalidaation avulla. Tässä alustusvaiheessa rakennetaan ensin GWAS-tunnuslukujen ja LD-referenssin pohjalta joukko uusia tunnuslukuja mallin koulutusta ja validointia varten. Sitten SBayesRC-menetelmä suoritetaan ilman funktionaalisia annotaatioita pienellä iteraatiomäärällä (150) koulutustunnusluvuille eri parametrin ρ arvoilla (oletusarvona käytetään arvoja 0.9, 0.95, 0.99 ja 0.995). Tämän jälkeen ennustetarkkuus lasketaan käyttäen validaatioon varattuja tunnuslukuja, ja lopulta valitaan se parametrin ρ arvo, jolla prediktiotarkkuus on suurin. Kun SBayesRC-menetelmän pääfunktio (jolla estimoidaan snippien yhteisvaikutukset) suoritetaan R-ohjelmalla oletusasetuksilla, ρ valitaan edellä esitetyllä tavalla. Testattavia parametrin ρ arvoja on mahdollista vaihtaa tai ohittaa alustusvaihe kokonaan ja valita käytettävä parametrin ρ arvo suoraan.

3.2 Hierarkkinen Bayes-malli

Tässä luvussa käsitellään hierarkkinen Bayes-malli, johon SBayesRC-menetelmä perustuu. Zhengin ja kumppanien artikkelissa on esitetty Bayes-mallin yhteistiheysja-

kauma:

$$
\begin{aligned}
f(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\mu}, {\boldsymbol{\sigma}_{\boldsymbol{\alpha}}}^2, \sigma_\varepsilon^2) &\propto (\sigma_\varepsilon^2)^{-\frac{g}{2}} \exp\left\{ -\frac{(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})'(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})}{2\frac{\sigma_\varepsilon^2}{n}} \right\} \\
&\times \prod_{j=1}^{m} \left\{ \sum_{k=1}^{5} \pi_{jk} \left[\exp\left\{ -\frac{\beta_j^2}{2\gamma_k \sigma_g^2} \right\} \right] \right\} \\
&\times \prod_{k=2}^{5} \prod_{j=1}^{m} \Phi(\mu_k + \mathbf{A}_j \boldsymbol{\alpha}_k)^{z_{jk}} [1 - \Phi(\mu_k + \mathbf{A}_j \boldsymbol{\alpha}_k)]^{(1-z_{jk})} \\
&\times \prod_{k=2}^{5} \prod_{c=1}^{C} (\sigma_{\alpha_k}^2)^{-\frac{1}{2}} \exp\left\{ -\frac{\alpha_{kc}^2}{2\sigma_{\alpha_k}^2} \right\} \\
&\times \prod_{k=2}^{5} (\sigma_{\alpha_k}^2)^{-\frac{2+\nu_\alpha}{2}} \exp\left\{ -\frac{\nu_\alpha \tau_\alpha^2}{2\sigma_{\alpha_k}^2} \right\} \\
&\times (\sigma_\varepsilon^2)^{-\frac{2+\nu_\varepsilon}{2}} \exp\left\{ -\frac{\nu_\varepsilon \tau_\varepsilon^2}{2\sigma_\varepsilon^2} \right\}.
\end{aligned}
\tag{6}
$$

Tässä esityksessä on korjattu joitakin epäjohdonmukaisuuksia Zhengin ja kumppanien artikkelin esityksestä, eikä se siksi vastaa sitä täysin. Symbolilla $\propto$ tarkoitetaan verrannollisuutta. Mallin parametreista esitettiin jo aiemmin matala-asteisen mallin (5) SNP-vaikutusten vektori $\boldsymbol{\beta}$ ja residuaalien varianssiparametri σ_ε^2. Muut parametrit esitetään seuraavissa alaluvuissa. Seuraavaksi luvuissa 3.2.1 – 3.2.7 käydään yksi rivi kerrallaan läpi yhteistiheysjakauman (6) osat, ja samalla estimoitavat parametrit ja Bayes-mallin hierarkkinen rakenne.

3.2.1 Havaintojen regressiomalli

Bayes-mallin alimmalla tasolla on havaintojen otantamallin taso, joka perustuu matala-asteiseen malliin (5). Tämä taso käsittää havaintojen $\mathbf{w}$ normaalisen otantamallin, jonka parametreina ovat SNP-vaikutukset $\boldsymbol{\beta}$ ja residuaalivarianssi σ_ε^2. Matala-asteinen malli on lineaarinen regressiomalli, jonka tiheysfunktion yleinen muoto on

$$
\frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right\}. \tag{7}
$$

Tämä on multinormaalijakauman tiheysfunktio, jossa selitettävän muuttujan arvojen $\mathbf{Y}$ odotusarvot määräytyvät selittävien muuttujien arvojen ja regressiokertoimien tulona $\mathbf{X}\boldsymbol{\beta}$. Vektorin $\mathbf{Y}$ kovarianssimatriisi on $\boldsymbol{\Sigma}$ ja alkioiden määrä n. Kun tiheysfunktioon (7) sijoitetaan matala-asteisen mallin (5) selitettävän muuttujan arvojen vektori $\mathbf{w}$, odotusarvovektori $\mathbf{Q}\boldsymbol{\beta}$ ja matala-asteisen mallin kovarianssimatriisi eli residuaalivektorin kovarianssimatriisi $\frac{1}{n}\mathbf{I}\sigma_\varepsilon^2$, sekä lisäksi pudotetaan pois uskottavuuspäättelyn kannalta epäoleellinen mallin parametreista riippumaton kerroin

$\frac{1}{\sqrt{(2\pi)^n}}$, saadaan matala-asteisen mallin tiheysfunktio muotoon

$$\frac{1}{\sqrt{|\frac{1}{n}\mathbf{I}\sigma_\varepsilon^2|}}\exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})'(\frac{1}{n}\mathbf{I}\sigma_\varepsilon^2)^{-1}(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})\right\}.$$

Diagonaalimatriisin determinantti on diagonaalialkioiden tulo, joten

$$\frac{1}{\sqrt{|\frac{1}{n}\mathbf{I}\sigma_\varepsilon^2|}} = \frac{1}{\sqrt{\frac{(\sigma_\varepsilon^2)^q}{n^q}}} = n^{\frac{q}{2}}(\sigma_\varepsilon^2)^{-\frac{q}{2}}.$$

$n^{\frac{q}{2}}$ on mallin parametreista riippumaton vakio, joka voidaan uskottavuuspäättelyssä tiputtaa pois. Eksponenttifunktion sisällä kovarianssimatriisin käänteismatriisin $\boldsymbol{\Sigma}^{-1} = \left(\frac{1}{n}\mathbf{I}\sigma_\varepsilon^2\right)^{-1}$ identiteettimatriisi katoaa matriisitulossa jättäen vain skalaarikertoimen $\left(\frac{1}{n}\sigma_\varepsilon^2\right)^{-1} = \frac{1}{\sigma_\varepsilon^2/n}$. Näin lopulliseksi matala-asteisen mallin tiheysfunktioksi saadaan

$$(\sigma_\varepsilon^2)^{-\frac{q}{2}}\exp\left\{-\frac{(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})'(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})}{2\frac{\sigma_\varepsilon^2}{n}}\right\},$$

joka on yhteistiheysjakauman (6) ensimmäisellä rivillä.

3.2.2 SNP-vaikutusten priorijakaumat

Hierarkkisen Bayes-mallin seuraava taso sisältää snippien vaikutusten $\boldsymbol{\beta}$ priorijakaumat. Näiksi priorijakaumiksi on valittu normaalinen sekoitejakauma:

$$\beta_j \sim \sum_{k=1}^{5}\pi_{jk}N(0,\sigma_k^2), \quad f(\pi_{jk}) = \mathbf{A}_j\boldsymbol{\alpha}_k, \quad j = 1,...,m. \tag{8}$$

Tässä f on linkkifunktio, $\mathbf{A}_j$ on annotaatiomatriisin snippiä j vastaava rivi ja $\boldsymbol{\alpha}_k$ on tuntemattomien annotaatiovaikutusten vektori. Yksi SBayesRC:n keskeisistä piirteistä on funktionaalisen annotaatiodatan hyödyntäminen polygeenisen ennustekyvyn tehostamisessa. Annotaatiodata sisällytetään tilastolliseen malliin tällä tavalla yhteisvaikutusten priorijakaumien kautta. Sekoitejakauma (8) koostuu k:sta nollakeskisestä normaalijakaumasta, joilla kullakin on todennäköisyys π_{jk}, että yhteisvaikutus β_j kuuluu kyseiseen jakaumaan. Normaalijakaumien varianssiparametrit σ_k^2 ovat k-ryhmäkohtaiset, ja ne parametrisoidaan Bayes-mallissa tulona $\sigma_k^2 = \gamma_k\sigma_g^2$. Tässä σ_g^2 on geneettinen varianssi, joka estimoidaan Gibbsin otannassa joka iteraatiolla erikseen kuten tutkielman liitteessä B on esitetty. Geneettisen varianssin ajatellaan hierarkkisessa Bayes-mallissa olevan tunnettu kiinteä arvo, ja k-kohtaiset painot γ_k ovat myös tunnettuja kiinteitä arvoja.

Normaalisekoitejakauman yleinen muoto on yksiulotteisessa tapauksessa

$$y = \sum_{k=1}^{K}\pi_k N(\mu_k,\sigma_k^2),$$

jolloin normaalisekoitejakauman tiheysfunktio on

$$\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{ -\frac{1}{2\sigma_k^2}(y-\mu_k)^2 \right\}. \tag{9}$$

Tässä K on sekoitejakauman luokkien lukumäärä. SBayesRC-metodissa näitä luokkia on viisi eli $K = 5$. Nyt satunnaismuuttujana on snipin j yhteisvaikutus β_j eli $y = \beta_j$. Todennäköisyysparametri on j-kohtainen eli π_{jk}. Koska yhteisvaikutusten odotusarvoksi oletetaan 0 kaikissa luokissa, $\mu_k = 0$ kaikilla k. Koska k-kohtainen varianssiparametri $\sigma_k^2 = \gamma_k \sigma_g^2$ on estimoitavista parametreista riippumaton vakio, voidaan tiheysfunktion (9) eksponenttifunktiota edeltävä kerroin

$$\frac{1}{\sqrt{2\pi\sigma_k^2}} = \frac{1}{\sqrt{2\pi\gamma_k\sigma_g^2}}$$

jättää pois, jolloin yhteisvaikutuksen β_j tiheysfunktioksi jää

$$\sum_{k=1}^{K} \pi_{jk} \exp\left\{ -\frac{\beta_j^2}{2\gamma_k\sigma_g^2} \right\}.$$

Kaikkien yhteisvaikutusten $\boldsymbol{\beta}$ priorijakaumat voidaan kirjoittaa muodossa

$$\prod_{j=1}^{m} \sum_{k=1}^{K} \pi_{jk} \exp\left\{ -\frac{\beta_j^2}{2\gamma_k\sigma_g^2} \right\}.$$

Nämä yhteisvaikutusten priorijakaumat ovat yhteistiheysfunktion (6) toisella rivillä.

3.2.3 Todennäköisyyksien $\boldsymbol{\pi}_j$ uudelleenparametrisointi ja indikaattorimuuttujat δ_j ja z_{jk}

Parametrien $\boldsymbol{\pi}_j$ alkiot eivät ole toisistaan riippumattomia, koska ne ovat todennäköisyyksiä, jotka summautuvat arvoon 1 jokaisella j. Annotaatiovaikutuksille $\boldsymbol{\alpha}_k$ ja sekoitejakaumakomponenttien todennäköisyysparametreille π_{jk} käytetään vaihtoehtoista parametrisointia, jotta näistä riippuvuuksista päästäisiin eroon. Indikaattorimuuttuja δ_j määritellään snipin j sekoitejakaumaryhmään kuuluvuutena, jolloin

$$P(\delta_j = k) = \pi_{jk}, \quad k \in \{1, 2, 3, 4, 5\}.$$

Seuraavaksi määritellään ehdollinen todennäköisyys, että snipin j vaikutus kuuluu ryhmään k tai sitä korkeampaan ryhmään sillä ehdolla, että vaikutus kuuluu ryhmään $k-1$ tai sitä korkeampaan ryhmään:

$$p_{jk} = P(\delta_j \geq k \mid \delta_j \geq k-1), \quad k \geq 2.$$

Nyt pätee

$$\begin{aligned}
\pi_{j1} &= 1 - p_{j2}, \\
\pi_{j2} &= (1 - p_{j3})p_{j2}, \\
\pi_{j3} &= (1 - p_{j4})p_{j3}p_{j2}, \\
\pi_{j4} &= (1 - p_{j5})p_{j4}p_{j3}p_{j2}, \\
\pi_{j5} &= p_{j5}p_{j4}p_{j3}p_{j2}.
\end{aligned}$$

Ensimmäinen yhtäsuuruus pätee, koska todennäköisyys että snipin j vaikutus kuuluu ryhmään 1 on komplementtitodennäköisyys todennäköisyydelle $P(\delta_j \geq 2)$, eli todennäköisyydelle, että jakauma kuuluu johonkin toiseen ryhmään. Koska ehto $\delta_j \geq 1$ toteutuu varmasti, pätee

$$p_{j2} = P(\delta_j \geq 2 \mid \delta_j \geq 1) = P(\delta_j \geq 2),$$

jolloin

$$p_{j1} = 1 - P(\delta_j \geq 2) = 1 - p_{j2}.$$

Toinen yhtäsuuruus pätee, koska

$$\begin{aligned}
\pi_{j2} &= P(\delta_j = 2) = P(\delta_j < 3 \ \cap \ \delta_j \geq 2) \\
&= P(\delta_j \geq 2)P(\delta_j < 3 \mid \delta_j \geq 2) \\
&= P(\delta_j \geq 2 \mid \delta_j \geq 1)(1 - P(\delta_j \geq 3 \mid \delta_j \geq 2)) \\
&= p_{j2}(1 - p_{j3}).
\end{aligned}$$

Muut yhtäsuuruudet voidaan johtaa vastaavalla tavalla. Nyt jokainen p_{jk} yhdistetään vektoriin $\boldsymbol{\alpha}_k$ yleistetyllä lineaarisella mallilla

$$g(p_{jk}) = \mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc}. \tag{10}$$

Todennäköisyyden p_{jk} linkkifunktion määräämä muunnos määräytyy siis snipin j annotaatiomuuttujien arvojen A_{jc}, annotaatiomuuttujien k-ryhmäkohtaisten vaikutusten α_{kc} ja k-kohtaisen vakiotermin μ_k lineaarisena funktiona. Annotaatiomuuttujien lukumäärä on C. Parametrisaatio määrittää implisiittisesti aiemmin priorijakauman (8) yhteydessä esitetyn linkkifunktion f. Parametrisaatiossa parametrit p_{jk} ovat riippumattomia toisistaan, ja vektoreille $\boldsymbol{\alpha}_k$ voidaan suorittaa otantaa rinnakkain jokaisessa MCMC-algoritmin iteraatiossa. Yleistetyn lineaarisen mallin (10) linkkifunktioksi g valitaan probit-funktio eli

$$g(p_{jk}) = \Phi^{-1}(p_{jk}),$$

jossa Φ on standardinormaalijakauman kertymäfunktio.

On hyvä huomata, että tässä parametrisaatiossa ryhmä $k = 1$ toimii pohja- tai referenssitasona, ja todennäköisyydet p_{jk} sekä parametrit μ_k ja α_{kc} on määritelty k:n arvoilla $k \in \{2, 3, 4, 5\}$. Tämä näkyy yhteistiheysfunktiossa (6) siten, että annotaatiovaikutuksiin liittyvissä osioissa käydään läpi k:n arvot kahdesta viiteen eikä yhdestä viiteen.

Lisäksi määritellään indikaattorimuuttuja z_{jk}, joka kertoo, millä todennäköisyydellä snipin j annotaatiovaikutus kuuluu korkeampaan sekoitejakauman luokkaan. Toisin sanoen $z_{jk} = 1$, jos $\delta_j = k$ $(k \geq 2)$.

3.2.4 Annotaatiosekoitteen malli

Hierarkkisen Bayes-mallin seuraava taso koskee annotaatiomallia, jossa snippien annotaatiotiedot $\mathbf{A}_j$ ja annotaatiovaikutukset α_{kc} määrittävät todennäköisyyden, että snipin yhteisvaikutuksen β_j priorijakauma kuuluu ryhmään k. Kuten edellisessä alaluvussa esitettiin, annotaatioiden yhteys todennäköisyyksiin p_{jk} on

$$p_{jk} = \Phi(\mu_k + \mathbf{A}_j \boldsymbol{\alpha}_k).$$

Koska indikaattorimuuttuja z_{jk} noudattaa Bernoulli-jakaumaa todennäköisyysparametrilla p_{jk}, sen Bernoulli-mallin mukainen todennäköisyys on

$$p_{jk}^{z_{jk}}(1 - p_{jk})^{1-z_{jk}} = \Phi(\mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc})^{z_{jk}}[1 - \Phi(\mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc})]^{1-z_{jk}}.$$

Indikaattorimuuttujien z_{jk} yhteistodennäköisyys saadaan käymällä läpi ryhmät $k \in \{2, 3, 4, 5\}$ kaikilla snipeillä j, jolloin saadaan

$$\prod_{k=2}^{5}\prod_{j=1}^{m} \Phi(\mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc})^{z_{jk}}[1 - \Phi(\mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc})]^{1-z_{jk}},$$

joka on yhteistiheysfunktion (6) kolmannella rivillä.

Tässä tutkielmassa merkinnällä $\mathbf{A}_j$ tarkoitetaan matriisin $\mathbf{A}$ snippiä j vastaavaa riviä eli rivivektoria, Zhengin ja kumppanien artikkelissa merkinnällä $\mathbf{A}_j$ tarkoitetaan välillä samaa vektoria pystyvektorina. Siksi kun tässä tutkielmassa kirjoitetaan esimerkiksi kuten yllä $\mathbf{A}_j\boldsymbol{\alpha}_k$, kirjoitetaan Zhengin ja kumppanien artikkelissa vastaavasta vektoritulosta usein $\mathbf{A}_j'\boldsymbol{\alpha}_k$.

3.2.5 Annotaatiovaikutusten α_{kc} priorijakaumat

Hierarkkisen mallin seuraava taso koskee annotaatiovaikutusten α_{kc} priorijakaumia. Kuten edellisessä luvussa esitettiin, nämä jakaumat ovat nollakeskisiä normaalijakaumia, joiden hajontaparametrit $\sigma^2_{\alpha_k}$ ovat sekoitejakaumaryhmä k -kohtaisia. Toisin sanoen kaikilla annotaatiovaikutuksilla α_{kc} hajontaparametri määräytyy ryhmän k eikä annotaatiomuuttujan c mukaan. Normaalijakauman tiheysfunktio on tällöin

$$\frac{1}{\sqrt{2\pi\sigma^2_{\alpha_k}}}\exp\left\{-\frac{(\alpha_{kc}-0)^2}{2\sigma^2_{\alpha_k}}\right\},$$

ja kun tiheysfunktiosta pudotetaan pois epäoleellinen parametreista riippumaton vakio $\frac{1}{\sqrt{2\pi}}$, jäljelle jää

$$(\sigma^2_{\alpha_k})^{-\frac{1}{2}}\exp\left\{-\frac{\alpha^2_{kc}}{2\sigma^2_{\alpha_k}}\right\}.$$

Kaikkien annotaatiovaikutusten yhteisjakauma saadaan käymällä kaikkien luokkien $k \in \{2,3,4,5\}$ ja annotaatiomuuttujien $c \in \{1,2,...,C\}$ kombinaatiot, jolloin yhteisjakauma on

$$\prod_{k=2}^{5}\prod_{c=1}^{C}(\sigma^2_{\alpha_k})^{-\frac{1}{2}}\exp\left\{-\frac{\alpha^2_{kc}}{2\sigma^2_{\alpha_k}}\right\},$$

joka on yhteisjakauman (6) neljännellä rivillä. Zhengin ja kumppanien artikkelin lisämateriaalin yhteisjakaumassa on tässä kohtaa ilmeisesti virhe, koska hajontaparametreiksi on merkitty $\sigma^2_{\alpha_c}$ eikä $\sigma^2_{\alpha_k}$. Heidän aikaisemmassa esityksessään annotaatiovaikutusten priorijakauman todetaan olevan

$$\alpha_{kc} \sim N(0, \sigma^2_{\alpha_k}),$$

eli hajontaparametri riippuu ryhmästä k eikä c.

Yhteistiheysjakaumassa (6) ei näy annotaatiomallin vakiotermien μ_k priorijakaumia. Tämä johtuu luultavasti siitä, että näille on valittu ei-informatiiviset priorit joille pätee $f(\mu_k) \propto 1$, joten näillä prioreilla ei ole vaikutusta yhteistiheysjakaumaan.

3.2.6 Hajontaparametrien priorijakaumat

Hierarkkisen mallin viimeinen taso koskee mallin hajontaparametrien priorijakaumia. Ensimmäiset hajontaparametrit ovat annotaatiovaikutusten priorijakaumien hajontaparametrit $\sigma^2_{\alpha_k}$. Näiden priorijakauma on skaalattu käänteinen χ^2-jakauma (tai vaihtoehtoisella parametrisoinnilla käänteinen gammajakauma), jonka perusmuoto on

$$\frac{\left(\frac{\nu\tau^2}{2}\right)}{\Gamma(\nu/2)}x^{-(\nu/2+1)}\exp\left(-\frac{\nu\tau^2}{2x}\right), x > 0.$$

Parametri ν on jakauman vapausaste-parametri ja τ^2 on jakauman skaalaparametri. Hierarkkisessa Bayes-mallissa parametrien $\sigma^2_{\alpha_k}$ priorijakaumissa parametrit ν_α ja τ^2_α ovat kiinteitä. Zhengin ja kumppanien artikkelissa ei kerrota suoraan näiden

parametrien arvoja, mutta ne on luultavasti valittu niin, että priori on heikosti informatiivinen.

Termi $(\nu\tau^2/2)\ /\ \Gamma(\nu/2)$ voidaan jättää pois estimoitavista parametreista. Sijoittamalla muuttujaksi hajontaparametri $\sigma^2_{\alpha_k}$ jäljelle jää

$$\left(\sigma^2_{\alpha_k}\right)^{-(\nu_\alpha/2+1)} \exp\left(-\frac{\nu_\alpha \tau^2_\alpha}{2\sigma^2_{\alpha_k}}\right) = \left(\sigma^2_{\alpha_k}\right)^{-\frac{2+\nu_\alpha}{2}} \exp\left(-\frac{\nu_\alpha \tau^2_\alpha}{2\sigma^2_{\alpha_k}}\right).$$

Priorijakaumien yhteistiheys on tällöin

$$\prod_{k=2}^{5} \left(\sigma^2_{\alpha_k}\right)^{-\frac{2+\nu_\alpha}{2}} \exp\left(-\frac{\nu_\alpha \tau^2_\alpha}{2\sigma^2_{\alpha_k}}\right).$$

Tämä on yhteistiheysjakauman (6) viidennellä rivillä. Myös tässä kohtaa Zhengin ja kumppanien kirjoittamassa yhteisjakaumassa on ilmeisesti virhe. Hajontaparametriksi on merkitty $\sigma^2_{\alpha_c}$ eikä $\sigma^2_{\alpha_k}$, vaikka hajontaparametri riippuu ryhmästä k eikä annotaatiomuuttujasta c.

Mallin viimeinen parametri on matala-asteisen mallin residuaalien hajontaparametri σ^2_ε. Tämän priorijakauma on myös skaalattu käänteinen χ^2-jakauma vapausasteparametrilla ν_ε ja skaalausparametrilla τ^2_ε. Samoilla perusteilla kuin edellisten hajontaparametrien tapauksissa, residuaalien hajontaparametrin tiheysfunktio on

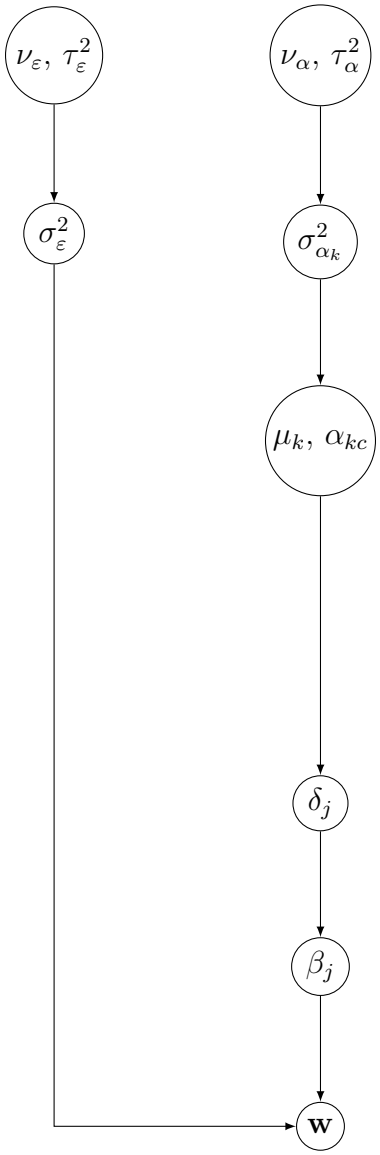
$$\left(\sigma^2_{\alpha_\varepsilon}\right)^{-\frac{2+\nu_\varepsilon}{2}} \exp\left(-\frac{\nu_\varepsilon \tau^2_\varepsilon}{2\sigma^2_\varepsilon}\right),$$

joka on yhteystiheysjakauman (6) viimeisellä rivillä.

3.2.7 Yhteenveto estimoitavista parametreista ja hierarkkisesta Bayes-mallista

Hierarkkisen Bayes-mallin estimoitavia tuntemattomia parametreja ovat snippien yhteisvaikutukset β_j, $j = 1, ..., m$, snippien yhteisvaikutusten sekoitejakaumaryhmän indikaattorimuuttujat δ_j, $j = 1, ..., m$, annotaatiovaikutukset α_{kc} ja annotaatiomallin k-kohtaiset vakiotermit μ_k, $k = 2, 3, 4, 5$ ja $c = 1, ..., C$, annotaatiovaikutusten hajontaparametrit $\sigma^2_{\alpha_k}$, $k = 2, 3, 4, 5$ sekä matala-asteisen mallin residuaalien hajontaparametri σ^2_ε. Hierarkkisen mallin tunnettuja kiinteitä parametreja ovat matala-asteisen mallin kerroinmatriisi $\mathbf{Q}$ $(q \times m)$, geneettinen varianssiparametri σ^2_g, joka estimoidaan Gibbsin otannan jokaisella iteraatiolla erikseen empiirisesti ja jota hierarkkisessa mallissa pidetään tunnettuna vakiona, annotaatiomuuttujien arvot matriisista $\mathbf{A}$ $(m \times C)$ sekä hajontaparametrien priorijakaumien parametrit ν_α, τ^2_α, ν_ε ja τ^2_ε. Matala-asteisen mallin marginaalivaikutusten lineaarikombinaatioiden vektoria $\mathbf{w}$ $(q \times 1)$ voidaan ajatella mallin aineistona.

Bayes-mallin hierarkkinen rakenne ja parametrien väliset yhteydet on esitetty kuvaajassa 3. On hyvä huomata, että SNP-vaikutusten sekoitejakaumaryhmään kuuluvuudet voidaan esittää indikaattorimuuttujan δ_j sijaan myös indikaattorimuuttujilla z_{jk}. Esimerkiksi yhteistiheysjakaumassa (6) esiintyy nimenomaan muuttuja z_{jk}.



$$\sigma_\varepsilon^2 \sim \text{Scaled Inv-}\chi^2(\nu_\varepsilon, \tau_\varepsilon^2)$$

$$\sigma_\alpha^2 \sim \text{Scaled Inv-}\chi^2(\nu_\alpha, \tau_\alpha^2)$$

$$\alpha_{kc} \sim N(0, \sigma_{\alpha_k}^2)$$

$$f(\mu_k) \propto 1$$

$$p_{jk} = \Phi\left(\mu_k + \textstyle\sum_{c=1}^{C} A_{jc}\alpha_{kc}\right),$$
$$\pi_{jk} = f(p_{jk})$$

$$P(\delta_j) = \pi_{jk}$$

$$\beta_j \sim \textstyle\sum_{k=1}^{5} \pi_{jk} N(0, \gamma_k \sigma_g^2),$$

$$\beta_j | (\delta_j = k) \sim N(0, \gamma_k \sigma_g^2),$$

$$\mathbf{w} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{e}$$

Kuva 3: Hierarkkisen Bayes-mallin parametrien väliset suhteet.

3.3 Mallin täysehdolliset jakaumat

Malliparametrien estimointi tapahtuu SBayesRC-menetelmässä Gibbsin otannalla. Gibbsin otannan jokaisella iteraatiokierroksella kaikille estimoitaville parametreille poimitaan vuorollaan uusi arvo kyseisen parametrin täysehdollisesta jakaumasta. Täysehdollisessa jakaumassa kaikkien muiden parametrien arvot ehdollistetaan tunnetuiksi kiinteiksi arvoiksi, ja Gibbsin otannassa näinä arvoina toimivat kunkin parametrin viimeisin poimittu arvo. Seuraavaksi käydään läpi jokaisen estimoitavan parametrin Gibbsin otannassa käytettävä täysehdollinen jakauma, sekä Gibbsin otannassa käytettävä apumuuttuja l_{jk}.

3.3.1 Apumuuttuja l_{jk}

SBayesRC-menetelmässä käytetään Albertin ja Chibin [1] esittelemää tapaa ottaa käyttöön latentti apumuuttuja l_{jk}, jonka avulla muodostetaan annotaatiovaikutusten α_{kc} täysehdolliset jakaumat, joita käytetään Gibbsin otannassa. Zhengin ja kumppanien mukaan on tunnettua, että apumuuttujalle l_{jk} joka määräytyy relaatiosta

$$z_{jk} = \begin{cases} 0, & \text{jos } l_{jk} > 0 \\ 1, & \text{jos } l_{jk} \leq 0 \end{cases}$$

voidaan muodostaa lineaarinen malli

$$l_{jk} = \mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc} + \varepsilon_{jk},$$

missä $\varepsilon_{jk} \sim N(0,1)$. Kun annotaatiokomponenteille α_{kc} oletetaan normaalinen prioriajakauma $\alpha_{kc} \sim N(0, \sigma^2_{\alpha_k})$, on niiden täysehdollinen jakauma yksiulotteinen normaalijakauma, koska α_{kc} on riippumaton vektorista $\boldsymbol{z}_k$ ehdolla vektori $\boldsymbol{l}_k$. Latentin muuttujan l_{jk} täysehdollinen jakauma ehdolla z_{jk} ja $\boldsymbol{\alpha}_k$ on katkaistu (*truncated*) normaalijakauma.

Latentti apumuuttuja l_{jk} ei ole itsessään kiinnostuksen kohteena, vaan se on apuväline Gibbsin otannan suorittamiseksi. Siksi se ei ole mukana hierarkkisen Bayes-mallin yhteistiheysjakaumassa (6). Gibbsin otanta suoritetaan laajennetulla mallilla, jossa l_{jk} on mukana, ja valitsemalla otannasta kaikki muut parametrit kuin l_{jk}, marginaalinen posteriorijakauma koskee alkuperäistä hierarkkista Bayes-mallia.

3.3.2 SNP-vaikutusten täysehdolliset jakaumat

Zhengin ja kumppanien tutkimusartikkelin liitteessä snipin yhteisvaikutuksen β_j täysehdollisen jakauman todetaan olevan verrannollinen normaalijakaumaan

$$\begin{aligned} f(\beta_j \mid \mathbf{w},\, \boldsymbol{\beta_{-j}},\, \delta_j = k,\, \sigma^2_g,\, \sigma^2_\varepsilon) &\propto N\left(\frac{r_j}{C_j}, \frac{\sigma^2_\varepsilon}{C_j}\right), \\ r_j = \mathbf{Q}'\left(\mathbf{w} - \sum_{t\neq j} \mathbf{Q}'_t\beta_t\right), &\quad C_j = 1 + \frac{\sigma^2_\varepsilon}{\gamma_k\sigma^2_g}. \end{aligned} \tag{11}$$

Merkinnällä $\boldsymbol{\beta}_{-j}$ tarkoitetaan vektoria $\boldsymbol{\beta}$, josta on poistettu sen j:nnes alkio. Seuraavaksi käydään läpi kuinka tähän jakaumaan päädytään.

Ensin täysehdolliseen jakaumaan valitaan mukaan vain termit joissa β_j on mukana, jolloin saadaan

$$\begin{aligned} & f(\beta_j \mid \mathbf{w},\, \boldsymbol{\beta_{-j}},\, \delta_j = k,\, \sigma_g^2,\, \sigma_\varepsilon^2) \\ \propto\ & \exp\left\{ -\frac{(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})'(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})}{2\frac{\sigma_\varepsilon^2}{n}} \right\} \times \prod_{j=1}^{m} \left\{ \sum_{k=1}^{5} \pi_{jk} \left[\exp\left\{ -\frac{\beta_j^2}{2\gamma_k\sigma_g^2} \right\} \right] \right\}. \end{aligned} \tag{12}$$

Tarkastellaan ensin jälkimmäistä termiä eli yhteivaikutusten β_j priorijakaumia. Koska nyt oletetaan vaikutukset β_t, $t \neq j$, kiinteiksi vakioiksi, voidaan niitä vastaavat termit jättää huomiotta, jolloin jäljelle jää vain termi

$$\sum_{k=1}^{5} \pi_{jk} \left[\exp\left\{ -\frac{\beta_j^2}{2\gamma_k\sigma_g^2} \right\} \right].$$

Koska lisäksi oletetaan, että β_j kuuluu sekoitejakaumaryhmään k eli $\delta_j = k$, jää lopulta jäljelle vain tätä vastaava komponentti

$$\exp\left\{ -\frac{\beta_j^2}{2\gamma_k\sigma_g^2} \right\}. \tag{13}$$

Tarkastellaan seuraavaksi tiheysfunktion (12) ensimmäistä termiä eli matalaasteisen mallin (5) uskottavuusfunktiota. Tässä kohtaa täytyy huomauttaa, että jotta päädyttäisiin Zhengin ja kumppanien antamaan normaalijakaumaan, tulee tästä termistä pudottaa varianssiparametrin σ_ε^2 kerroin $1/n$. Vastaava tilanne nähdään myöhemmin luvun 3.3.6 lopussa, jossa Zhengin ja kumppanien antamaan jakaumaan päädytään pudottamalla kyseinen kerroin. On epäselvää miksi kertoimen $1/n$ pudottaminen on perusteltua. Matriisitulossa $\mathbf{Q}\boldsymbol{\beta}$ voidaan muuttuja β_j erottaa toisista kirjoittamalla

$$\mathbf{Q}\boldsymbol{\beta} = \mathbf{Q}_j\beta_j + \sum_{t\neq j} \mathbf{Q}_t\beta_t,$$

Tässä $\mathbf{Q}_t$ on matriisin $\mathbf{Q}$ sarake t. Määritellään sitten ilman muuttujaa β_j muodostettu residuaalivektori

$$\mathbf{e}_{-j} = \mathbf{w} - \sum_{t\neq j} \mathbf{Q}_t\beta_t.$$

Nyt pätee

$$\begin{aligned} \mathbf{w} = \mathbf{Q}_j\beta_j + \sum_{t\neq j} \mathbf{Q}_t\beta_t + \mathbf{e} \Rightarrow \mathbf{w} - \sum_{t\neq j} \mathbf{Q}_t\beta_t = \mathbf{Q}_j\beta_j + \mathbf{e} \\ \Rightarrow \mathbf{e}_{-j} = \mathbf{Q}_j\beta_j + \mathbf{e}. \end{aligned}$$

Residuaalivektori $\mathbf{e}_{-j}$ riippuu siis vain muuttujasta β_j. Ehdollinen uskottavuus on siis verrannollinen lausekkeeseen

$$\exp\left\{-\frac{||\mathbf{e}_{-j}-\mathbf{Q}_j\beta_j||^2}{2\sigma_\varepsilon^2}\right\}. \tag{14}$$

Täysehdollinen jakauma on verrannollinen tähän kerrottuna ehdollisella priorilla (13), jolloin saadaan

$$\begin{aligned}&\exp\left\{-\frac{||\mathbf{e}_{-j}-\mathbf{Q}_j\beta_j||^2}{2\sigma_\varepsilon^2}\right\}\exp\left\{-\frac{\beta_j^2}{2\gamma_k\sigma_g^2}\right\}\\ =\ &\exp\left\{-\frac{\mathbf{e}'_{-j}\mathbf{e}_{-j}-2\beta_j\mathbf{Q}'_j\mathbf{e}_{-j}+\beta_j^2\mathbf{Q}'_j\mathbf{Q}_j}{2\sigma_\varepsilon^2}-\frac{\beta_j^2}{2\gamma_k\sigma_g^2}\right\}.\end{aligned}$$

Termi $\mathbf{e}'_{-j}\mathbf{e}_{-j}/2\sigma_\varepsilon^2$ ei sisällä muuttujaa β_j ja voidaan tiputtaa pois lausekkeesta. Järjestelemällä termejä saadaan lauseke muotoon

$$\exp\left\{-\frac{-2\beta_j\mathbf{Q}'_j\mathbf{e}_{-j}+\beta_j^2\mathbf{Q}'_j\mathbf{Q}_j}{2\sigma_\varepsilon^2}-\frac{\beta_j^2}{2\gamma_k\sigma_g^2}\right\}=\exp\left\{-\frac{1}{2}\left[\frac{-2\beta_j\mathbf{Q}'_j\mathbf{e}_{-j}+\beta_j^2\mathbf{Q}'_j\mathbf{Q}_j}{\sigma_\varepsilon^2}+\frac{\beta_j^2}{\gamma_k\sigma_g^2}\right]\right\}$$

$$\exp\left\{-\frac{1}{2}\left[\frac{\beta_j^2\mathbf{Q}'_j\mathbf{Q}_j}{\sigma_\varepsilon^2}+\frac{\beta_j^2}{\gamma_k\sigma_g^2}-2\beta_j\frac{\mathbf{Q}'_j\mathbf{e}_{-j}}{\sigma_\varepsilon^2}\right]\right\}=\exp\left\{-\frac{1}{2}\left[\beta_j^2\left(\frac{\mathbf{Q}'_j\mathbf{Q}_j}{\sigma_\varepsilon^2}+\frac{1}{\gamma_k\sigma_g^2}\right)-2\beta_j\frac{\mathbf{Q}'_j\mathbf{e}_{-j}}{\sigma_\varepsilon^2}\right]\right\}.$$

Tässä kohtaa otetaan käyttöön merkinnät

$$\begin{aligned}r_j&=\mathbf{Q}'_j\mathbf{e}_{-j},\\ C_j&=\mathbf{Q}'_j\mathbf{Q}_j+\frac{\sigma_\varepsilon^2}{\gamma_k\sigma_g^2}.\end{aligned}$$

Lauseke saadaan nyt muotoon

$$\exp\left\{-\frac{1}{2}\left[\beta_j^2\left(\frac{C_j}{\sigma_\varepsilon^2}\right)-2\beta_j\frac{r_j}{\sigma_\epsilon^2}\right]\right\}=\ \exp\left\{-\frac{C_j}{2\sigma_\varepsilon^2}\left(\beta_j^2-2\beta_j\frac{r_j}{C_j}\right)\right\}.$$

Tämä täydennetään lopulta neliöksi:

$$\begin{aligned}&\exp\left\{-\frac{C_j}{2\sigma_\varepsilon^2}\left(\beta_j^2-2\beta_j\frac{r_j}{C_j}\right)\right\}=\exp\left\{-\frac{C_j}{2\sigma_\varepsilon^2}\left(\beta_j^2-2\beta_j\frac{r_j}{C_j}+\left(\frac{r_j}{C_j}\right)^2-\left(\frac{r_j}{C_j}\right)^2\right)\right\}\\ =&\exp\left\{-\frac{C_j}{2\sigma_\varepsilon^2}\left(\beta_j-\frac{r_j}{C_j}\right)^2+\frac{r_j^2}{2\sigma_\varepsilon^2C_j}\right\}\propto\exp\left\{-\frac{C_j}{2\sigma_\varepsilon^2}\left(\beta_j-\frac{r_j}{C_j}\right)^2\right\}.\end{aligned}$$

Tämä lauseke tunnistetaan vakiota vaille $N(r_j/C_j, \sigma_\varepsilon^2/C_j)$-jakauman tiheysfunktioksi. Näin ollaan päästy kappaleen alussa mainittuun Zhengin ja kumppanien artikkelissa esitettyyn jakaumaan (11). Artikkelin mukaan

$$C_j = 1 + \frac{\sigma_\varepsilon^2}{\gamma_k \sigma_g^2},$$

kun taas tässä esityksessä

$$C_j = \mathbf{Q}_j'\mathbf{Q}_j + \frac{\sigma_\varepsilon^2}{\gamma_k \sigma_g^2}.$$

Seuraavaksi näytetään miksi $\mathbf{Q}_j'\mathbf{Q}_j = 1$. Koska matriisi $\mathbf{X}$ on standardoitu, $\mathbf{R} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ on sen korrelaatiomatriisi, jonka kaikkien diagonaalialkioiden arvo on 1. Luvussa 3.1 määriteltiin $\mathbf{Q} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{U}'$ ja $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$, josta seuraa

$$\mathbf{Q}_j'\mathbf{Q}_j = \mathbf{u}_j'\mathbf{\Lambda}\mathbf{u}_j = \mathbf{R}_{jj} = 1.$$

Suureelle r_j pätee

$$\begin{aligned} r_j = \mathbf{Q}_j'\mathbf{e}_{-j} =& \mathbf{Q}_j'\left(\mathbf{w} - \sum_{t \neq j} \mathbf{Q}_t'\beta_t\right) \\ =& \mathbf{Q}_j'(\mathbf{Q}_j\beta_j + \mathbf{e}) \\ =& \mathbf{Q}_j'\mathbf{e} + \mathbf{Q}_j'\mathbf{Q}_j\beta_j \\ =& \mathbf{Q}_j'\mathbf{e} + \beta_j. \end{aligned}$$

Tämä muoto esiintyy Gibbsin otanta-algoritmissa, joka esitellään myöhemmin.

3.3.3 Indikaattorimuuttujan δ_j täysehdollinen jakauma

Indikaattorimuuttuja δ_j kertoo mihin sekoitejakaumaryhmään $k \in \{1, 2, 3, 4, 5\}$ snippi j kuuluu. Zhengin ja kumppanien artikkelin liitteessä muuttujan δ_j täysehdolliseksi jakaumaksi on annettu

$$P(\delta_j = k \mid \mathbf{w}, \boldsymbol{\beta}, \sigma_g^2, \sigma_\varepsilon^2) = \frac{f(\mathbf{w} \mid \delta_j = k, \boldsymbol{\beta}, \sigma_g^2, \sigma_\varepsilon^2)\, f(\delta_j = k)}{\sum_{k'} f(\mathbf{w} \mid \delta_j = k', \boldsymbol{\beta}, \sigma_g^2, \sigma_\varepsilon^2)\, f(\delta_j = k')}. \tag{15}$$

Tässä $f(\delta_j = k) = \pi_{jk}$, ja kuten aikaisemmin on esitetty, π_{jk} on parametrien p_{jk} funktio, ja parametrit p_{jk} saadaan probit-linkkifunktion kautta kaavalla

$$\Phi^{-1}(p_{jk}) = \mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc}.$$

Termiä $f(\mathbf{w} \mid \delta_j = k, \boldsymbol{\beta}, \sigma_g^2, \sigma_\varepsilon^2)$ ei ole Zhengin ja kumppanien artikkelissa selitetty tämän enempää, mutta SBayesR-menetelmän artikkelissa aihetta on käsitelty tarkemmin [21]. Termi $f(\mathbf{w} \mid \delta_j = k, \boldsymbol{\beta}, \sigma_g^2, \sigma_\varepsilon^2)$ saadaan muodostamalla datan ja yhteisvaikutuksen β_j yhteisjakauma ehdoilla $\delta_j = k$ ja $\boldsymbol{\beta}_{-j}$, ja integroimalla β_j tästä pois. Koska β_j integroidaan pois, tulisi lausekkeen (15) merkinnässä $f(\mathbf{w} \mid \delta_j = k, \boldsymbol{\beta}, \sigma_g^2, \sigma_\varepsilon^2)$ luultavasti olla vektori $\boldsymbol{\beta}_{-j}$, merkintää $\boldsymbol{\beta}$ on luultavasti käytetty luettavuuden lisäämiseksi. Datan ja yhteisvaikutuksen β_j yhteisjakauma muodostetaan kertomalla ehdollinen uskottavuus muuttujan β_j ehdollisella priorilla. Jatkossa merkinnällä θ tarkoitetaan kaikkia muita parametreja, joita ei ole mainittu. Ehdollinen uskottavuus, kun kaikkien muiden snippien kuin snipin j yhteisvaikutukset $\boldsymbol{\beta}_{-j}$ oletetaan tunnetuiksi, johdettiin luvussa 3.3.2, ja se on verrannollinen lausekkeeseen

$$f(\mathbf{w}|\boldsymbol{\beta}_{-j}, \delta_j = k, \sigma_\varepsilon^2) \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}||\mathbf{e}_{-j} - \mathbf{Q}_j\beta_j||^2\right\}.$$

Luvussa 3.3.2 johdettiin implisiittisesti myös, että merkinnällä $r_j = \mathbf{Q}_j'\mathbf{e}_{-j}$ pätee

$$\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}||\mathbf{e}_{-j} - \mathbf{Q}_j\beta_j||^2\right\} \propto \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\beta_j^2 - 2r_j\beta_j)\right\}.$$

Muuttujan β_j priori ehdolla $\delta_j = k$ on normaalijakauma

$$f(\beta_j|\delta_j = k) \propto (2\pi\gamma_k\sigma_g^2)^{-\frac{1}{2}}\exp\left\{-\frac{\beta_j^2}{2\gamma_k\sigma_g^2}\right\} \propto \exp\left\{-\frac{\beta_j^2}{2\gamma_k\sigma_g^2}\right\}.$$

Datan $\mathbf{w}$ ja muuttujan β_j yhteisjakauma ehdoilla $\delta_j = k$ ja $\boldsymbol{\beta}_{-j}$ on nyt ehdollisen uskottavuuden ja muuttujan β_j ehdollisen priorin tulo, eli

$$\begin{aligned} f(\mathbf{w}, \beta_j|\delta_j = k, \theta) &\propto \exp\left\{-\frac{1}{2}\left[\frac{\beta_j^2}{\sigma_\varepsilon^2} - \frac{2r_j\beta_j}{\sigma_\varepsilon^2} + \frac{\beta_j^2}{\gamma_k\sigma_g^2}\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\beta_j^2\left[\frac{1}{\sigma_\varepsilon^2} + \frac{1}{\gamma_k\sigma_g^2}\right] + \frac{r_j}{\sigma_\varepsilon^2}\beta_j\right\}. \end{aligned}$$

Merkinnöillä

$$A = \frac{1}{\sigma_\varepsilon^2} + \frac{1}{\gamma_k\sigma_g^2}, \quad B = \frac{r_j}{\sigma_\varepsilon^2}$$

tämä yksinkertaistuu muotoon

$$\begin{aligned} &\exp\left\{-\frac{1}{2}A\beta_j^2 + B\beta_j\right\} = \exp\left\{-\frac{A}{2}\left[\beta_j^2 - \frac{2B}{A}\beta_j\right]\right\} \\ &= \exp\left\{-\frac{A}{2}\left[\beta_j^2 - \frac{2B}{A}\beta_j + \left(\frac{B}{A}\right)^2 - \left(\frac{B}{A}\right)^2\right]\right\} \\ &= \exp\left\{-\frac{A}{2}\left[\left(\beta_j - \frac{B}{A}\right)^2 - \left(\frac{B}{A}\right)^2\right]\right\} = \exp\left\{-\frac{A}{2}\left(\beta_j - \frac{B}{A}\right)^2\right\}\exp\left\{\frac{B^2}{2A}\right\}. \end{aligned}$$

Seuraavaksi yhteisjakaumasta integroidaan pois muuttuja β_j:

$$\int_{-\infty}^{\infty} \exp\Big\{-\tfrac{1}{2}A\left(\beta_j - \tfrac{B}{A}\right)^2\Big\}\, d\beta_j \, \exp\left\{\frac{B^2}{2A}\right\}.$$

Merkitään

$$y = \sqrt{A}\left(\beta_j - \frac{B}{A}\right) \Rightarrow \beta_j - \frac{B}{A} = \frac{y}{\sqrt{A}}, \quad d\beta_j = \frac{dy}{\sqrt{A}}.$$

Nyt integraali saadaan muotoon

$$\begin{aligned} &\int_{-\infty}^{\infty} \exp\left\{-\tfrac{1}{2}A\left(\frac{y}{\sqrt{A}}\right)^2\right\} \frac{dy}{\sqrt{A}} \, \exp\left\{\frac{B^2}{2A}\right\} \\ = \;& \frac{1}{\sqrt{A}} \int_{-\infty}^{\infty} \exp\big\{-\tfrac{1}{2}y^2\big\}\, dy \, \exp\left\{\frac{B^2}{2A}\right\}. \end{aligned}$$

Standardinormaalijakauman tiheysfunktiosta saadaan

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\big\{-\tfrac{1}{2}y^2\big\} = 1 \Rightarrow \int_{-\infty}^{\infty} \exp\big\{-\tfrac{1}{2}y^2\big\} = \sqrt{2\pi},$$

joten integraaliksi saadaan lopulta

$$\sqrt{\frac{2\pi}{A}} \exp\left\{\frac{B^2}{2A}\right\} \propto \; A^{-\frac{1}{2}} \exp\left\{\frac{B^2}{2A}\right\}.$$

Näin ollen

$$\begin{aligned} &f(\mathbf{w} \mid \delta_j = k, \boldsymbol{\beta}_{-j}, \sigma_g^2, \sigma_\varepsilon^2) \propto A^{-\frac{1}{2}} \exp\left\{\frac{B^2}{2A}\right\}, \\ &A = \frac{1}{\sigma_\varepsilon^2} + \frac{1}{\gamma_k \sigma_g^2}, \quad B = \frac{r_j}{\sigma_\varepsilon^2}, \quad r_j = \mathbf{Q}_j' \mathbf{e}_{-j}. \end{aligned}$$

Gibbsin otannassa lasketaan todennäköisyyskomponentit

$$P(\delta_j = k \mid \mathbf{w}, \boldsymbol{\beta}, \sigma_g^2, \sigma_\epsilon^2) = \frac{f(\mathbf{w} \mid \delta_j = k, \boldsymbol{\beta}, \sigma_g^2, \sigma_\epsilon^2)\, f(\delta_j = k)}{\sum_{k'} f(\mathbf{w} \mid \delta_j = k', \boldsymbol{\beta}, \sigma_g^2, \sigma_\epsilon^2)\, f(\delta_j = k')},$$

ja nämä muodostavat kategorisen jakauman, josta uusi δ_j poimitaan. SBayesR-menetelmässä tämä tehdään muodostamalla kumulatiivinen todennäköisyysvektori $(P(\delta_j = 1|\theta), ..., P(\delta_j = 5|\theta))$, poimimalla arvo u tasajakaumasta $U(0, 1)$ ja valitsemalla pienin luokka k niin, että kumulatiivinen todennäköisyys on suurempi tai yhtä suuri kuin u [21]. SBayesR-menetelmän artikkelissa nämä laskutoimitukset suoritetaan logaritmeilla.

3.3.4 Annotaatiovaikutusten α_{kc} täysehdollinen jakauma

Zhengin ja kumppanien artikkelissa annotaatiovaikutusten α_{kc} täysehdollisen jakauman todetaan olevan

$$\begin{aligned}
&f(\alpha_{kc} \mid \boldsymbol{l}_k, \boldsymbol{\alpha}_{-kc}, \sigma^2_{\alpha_k}) \\
&\propto \exp\Big\{-\tfrac{1}{2}\big(\boldsymbol{l}_k - \textstyle\sum_{c' \neq c} \mathbf{A}_{c'}\alpha_{kc'}\big)'\big(\boldsymbol{l}_k - \textstyle\sum_{c' \neq c} \mathbf{A}_{c'}\alpha_{kc'}\big)\Big\} \exp\Big\{-\tfrac{\alpha^2_{kc}}{2\sigma^2_{\alpha_k}}\Big\} \\
&= N\left(\frac{r_{kc}}{C_{kc}},\ \frac{1}{C_{kc}}\right), C_{kc} = \mathbf{A}'_c\mathbf{A}_c + \frac{1}{\sigma^2_{\alpha_k}}, \quad r_{kc} = \mathbf{A}'_c\mathbf{r}^l_k, \quad \mathbf{r}^l_k = \boldsymbol{l}_k - \sum_{c' \neq c} \mathbf{A}_{c'}\alpha_{kc'}.
\end{aligned}$$

Seuraavaksi näytetään kuinka tähän päädytään. Gibbsin otannassa indikaattorimuuttuja z_{jk} saa arvon 1 jos vaikutus j kuuluu sekoitejakaumaryhmään k ja muissa tapauksissa arvon 0. Indikaattorimuuttujan z_{jk} taustalla ajatellaan olevan latentti muuttuja l_{jk}, muuttujien välillä on relaatio

$$z_{jk} = \begin{cases} 0, & \text{jos } l_{jk} > 0 \\ 1, & \text{jos } l_{jk} \leq 0, \end{cases}$$

ja latentille muuttujalle on muodostettu lineaarinen malli

$$l_{jk} = \mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{jc} + \varepsilon_{jk}.$$

Samalla periaatteella kuin aikaisemmin muodostettiin snippien yhteisvaikutusten β_j ehdollinen uskottavuusfunktio, muodostetaan nyt annotaatiovaikutusten α_{kc} ehdollinen uskottavuus. Merkitään residuaalivektoria josta on vähennetty muiden annotaatioiden vaikutus

$$\mathbf{r}^l_k = \boldsymbol{l}_k - \sum_{c' \neq c} \mathbf{A}_{c'}\alpha_{kc'}.$$

Nyt $\mathbf{A}_c$ on annotaatiomatriisin $\mathbf{A}$ annotaatiota c vastaava sarake, siis pystyvektori. Ehdollisen uskottavuuden suureesta $\boldsymbol{l}_k$ riippuva osa on nyt

$$f(\boldsymbol{l}_k \mid \boldsymbol{\alpha}_{-kc}, \sigma^2_{\alpha_k}) \ \propto\ \exp\Big\{-\tfrac{1}{2}\big(\mathbf{r}^l_k - \mathbf{A}_c\alpha_{kc}\big)'\big(\mathbf{r}^l_k - \mathbf{A}_c\alpha_{kc}\big)\Big\}.$$

Samoin kuin aikaisemmin, uskottavuus kerrotaan parametrin α_{kc} priorilla $N(0, \sigma^2_{\alpha_k})$, josta voidaan jättää huomiotta parametrista α_{kc} riippumattomat vakiot. Näin saadaan

$$\exp\Big\{-\tfrac{1}{2}\big(\mathbf{r}^l_k - \mathbf{A}_c\alpha_{kc}\big)'\big(\mathbf{r}^l_k - \mathbf{A}_c\alpha_{kc}\big)\Big\} \exp\left\{-\frac{1}{2\sigma^2_{\alpha_k}}\alpha^2_{kc}\right\},$$

ja tämä on kuten Zhengin ja kumppanien artikkelissa. Kuten aikaisemmin snippien yhteisvaikutusten kanssa, ensimmäisen eksponenttifunktion sisällä oleva termi avataan ja parametrista α_{jk} riippumaton termi $\left(\mathbf{r}_k^l\right)'\mathbf{r}_k^l$ jätetään pois lausekkeesta:

$$\begin{aligned}
&-\tfrac{1}{2}\left(\mathbf{r}_k^l-\mathbf{A}_c\alpha_{kc}\right)'\left(\mathbf{r}_k^l-\mathbf{A}_c\alpha_{kc}\right)=-\tfrac{1}{2}\left(\ \left(\mathbf{r}_k^l\right)'\mathbf{r}_k^l-2\alpha_{kc}\mathbf{A}_c'\mathbf{r}_k^l+\alpha_{kc}^2\mathbf{A}_c'\mathbf{A}_c\right)\\
\propto&-\tfrac{1}{2}\left(\ -2\alpha_{kc}\mathbf{A}_c'\mathbf{r}_k^l+\alpha_{kc}^2\mathbf{A}_c'\mathbf{A}_c\right)=\alpha_{kc}\mathbf{A}_c'\mathbf{r}_k^l-\tfrac{1}{2}\alpha_{kc}^2\mathbf{A}_c'\mathbf{A}_c.
\end{aligned}$$

Näin täysehdollinen jakauma saadaan muotoon

$$\begin{aligned}
&\exp\left\{\alpha_{kc}\mathbf{A}_c'\mathbf{r}_k^l-\tfrac{1}{2}\alpha_{kc}^2\mathbf{A}_c'\mathbf{A}_c\right\}\exp\left\{-\frac{1}{2\sigma_{\alpha_k}^2}\alpha_{kc}^2\right\}=\exp\left\{\alpha_{kc}\mathbf{A}_c'\mathbf{r}_k^l-\tfrac{1}{2}\alpha_{kc}^2\mathbf{A}_c'\mathbf{A}_c-\frac{1}{2\sigma_{\alpha_k}^2}\alpha_{kc}^2\right\}\\
=&\exp\left\{-\tfrac{1}{2}\alpha_{kc}^2\left(\mathbf{A}_c'\mathbf{A}_c+\frac{1}{\sigma_{\alpha_k}^2}\right)+\alpha_{kc}\mathbf{A}_c'\mathbf{r}_k^l\right\}=\exp\left\{-\tfrac{1}{2}\alpha_{kc}^2C_{kc}+\alpha_{kc}r_{kc}\right\},
\end{aligned}$$

jossa $C_{kc}=\mathbf{A}_c'\mathbf{A}_c+1/\sigma_{\alpha_k}^2$ ja $r_{kc}=\mathbf{A}_c'\mathbf{r}_k^l$. Kuten aikaisemmin SNP-vaikutusten kanssa, tämä täydennetään neliöksi:

$$\begin{aligned}
&\exp\left\{-\tfrac{1}{2}\alpha_{kc}^2C_{kc}+\alpha_{kc}r_{kc}\right\}=\exp\left\{-\tfrac{1}{2}C_{kc}\left(\alpha_{kc}^2-2\alpha_{kc}\frac{r_{kc}}{C_{kc}}\right)\right\}\\
=&\exp\left\{-\tfrac{1}{2}C_{kc}\left(\alpha_{kc}^2-2\alpha_{kc}\frac{r_{kc}}{C_{kc}}+\frac{r_{kc}^2}{C_{kc}^2}-\frac{r_{kc}^2}{C_{kc}^2}\right)\right\}=\exp\left\{-\tfrac{1}{2}C_{kc}\left(\alpha_{kc}-\frac{r_{kc}}{C_{kc}}\right)^2+\frac{r_{kc}^2}{2C_{kc}}\right\}\\
\propto&\exp\left\{-\tfrac{1}{2}C_{kc}\left(\alpha_{kc}-\frac{r_{kc}}{C_{kc}}\right)^2\right\},
\end{aligned}$$

mikä on muuttujasta α_{kc} riippuva osa normaalijakaumasta, jonka odotusarvo on r_{kc}/C_{kc} ja varianssi on $1/C_{kc}$, kuten Zhengin ja kumppanien artikkelissa on esitetty.

3.3.5 Annotaatiomallin hajontaparametrien $\sigma_{\alpha_k}^2$ täysehdollinen jakauma

Zheng ja kumppanit ovat artikkelissaan antaneet hajontaparametrien $\sigma_{\alpha_k}^2$ täysehdolliseksi jakaumaksi skaalatun käänteisen χ^2-jakauman

$$\chi^{-2}(\tilde{\nu}_\alpha,\tilde{\tau}_\alpha^2),\quad \tilde{\nu}_\alpha=C+\nu_\alpha,\quad \tilde{\tau}_\alpha^2=\left(\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k+\nu_\alpha\tau_\alpha^2\right)/\tilde{\nu}_\alpha,$$

Seuraavaksi käydään läpi kuinka tähän päädytään. Koska parametrit $\sigma_{\alpha_k}^2$ riippuvat Bayes-mallin hierarkiassa alemmalla tasolla olevista parametreista vain parametrien $\boldsymbol{\alpha}_k$ kautta, on parametrien $\sigma_{\alpha_k}^2$ täysehdollinen jakauma $f(\sigma_{\alpha_k}^2|\boldsymbol{\alpha}_k)$, ja Bayesin lauseella saadaan $f(\sigma_{\alpha_k}^2|\boldsymbol{\alpha}_k)\propto f(\boldsymbol{\alpha}_k|\sigma_{\alpha_k}^2)f(\sigma_{\alpha_k}^2)$. Ensimmäinen termi vastaa normaalijakaumaa jokaiselle annotaatiovaikutukselle $\boldsymbol{\alpha}_k$, jolloin tiheysfunktio on

$$\begin{aligned}
&\prod_{c=1}^{C}\frac{1}{\sqrt{2\pi\sigma_{\alpha_k}^2}}\exp\left\{-\frac{\alpha_{kc}^2}{2\sigma_{\alpha_k}^2}\right\}\propto\prod_{c=1}^{C}\frac{1}{\sqrt{\sigma_{\alpha_k}^2}}\exp\left\{-\frac{\alpha_{kc}^2}{2\sigma_{\alpha_k}^2}\right\}\\
=&\left(\sigma_{\alpha_k}^2\right)^{-C/2}\exp\left\{-\frac{1}{2\sigma_{\alpha_k}^2}\sum_{c=1}^{C}\alpha_{kc}^2\right\}.
\end{aligned}$$

Jälkimmäinen termi aiemmassa posteriorijakaumassa eli hajontaparametrin priori on skaalattu käänteinen χ^2-jakauma, jonka tiheysfunktio ilman parametreista riippumatonta vakiota on

$$\left(\sigma^2_{\alpha_k}\right)^{-\frac{2+\nu_\alpha}{2}} \exp\left(-\frac{\nu_\alpha \tau^2_\alpha}{2\sigma^2_{\alpha_k}}\right).$$

Hajontaparametrien täysehdollinen jakauma on siis

$$f(\sigma^2_{\alpha_k}|\boldsymbol{\alpha}_k) \propto f(\boldsymbol{\alpha}_k|\sigma^2_{\alpha_k}) f(\sigma^2_{\alpha_k}) = \left(\sigma^2_{\alpha_k}\right)^{-C/2} \exp\left\{-\frac{1}{2\sigma^2_{\alpha_k}} \sum_{c=1}^{C} \alpha^2_{kc}\right\} \left(\sigma^2_{\alpha_k}\right)^{-\frac{2+\nu_\alpha}{2}} \exp\left(-\frac{\nu_\alpha \tau^2_\alpha}{2\sigma^2_{\alpha_k}}\right)$$
$$= \left(\sigma^2_{\alpha_k}\right)^{-\frac{2+C+\nu_\alpha}{2}} \exp\left\{-\frac{1}{2\sigma^2_{\alpha_k}} \left(\sum_{c=1}^{C} \alpha^2_{kc} + \nu_\alpha \tau^2_\alpha\right)\right\}.$$

Skaalatun käänteisen χ^2-jakauman tiheysfunktion parametreista $\tilde{\nu}$ ja $\tilde{\tau}^2$ riippuva osa on

$$x^{-\frac{2+\tilde{\nu}}{2}} \exp\left\{-\frac{\tilde{\nu}\tilde{\tau}^2}{2x}\right\}.$$

Tästä nähdään, että hajontaparametrien $\sigma^2_{\alpha_k}$ täysehdollinen jakauma on skaalattu käänteinen χ^2-jakauma parametreilla

$$-\frac{2+\tilde{\nu}_\alpha}{2} = -\frac{2+C+\nu_\alpha}{2} \Rightarrow \tilde{\nu}_\alpha = C + \nu_\alpha,$$
$$\tilde{\nu}_\alpha \tilde{\tau}^2_\alpha = \sum_{c=1}^{C} \alpha^2_{kc} + \nu_\alpha \tau^2_\alpha \Rightarrow \tilde{\tau}^2_\alpha = \frac{\sum_{c=1}^{C} \alpha^2_{kc} + \nu_\alpha \tau^2_\alpha}{\tilde{\nu}_\alpha} = \frac{\boldsymbol{\alpha}'_k \boldsymbol{\alpha}_k + \nu_\alpha \tau^2_\alpha}{\tilde{\nu}_\alpha},$$

kuten Zhengin ja kumppanien artikkelissa esitettiin.

3.3.6 Matala-asteisen mallin residuaalien hajontaparametrin σ^2_ε täysehdollinen jakauma

Zhengin ja kumppanien artikkelissa parametrin σ^2_ε täysehdolliseksi jakaumaksi on annettu skaalattu käänteinen χ^2-jakauma parametreilla $\tilde{\nu}_\varepsilon = q + \nu_\varepsilon$ ja $\tilde{\tau}^2_\varepsilon = (\mathbf{e}'\mathbf{e} + \nu_\varepsilon \tau^2_\varepsilon)/\tilde{\nu}_\varepsilon$, seuraavaksi käydään läpi kuinka tähän päädytään. Koska hajontaparametri σ^2_ε prioriparametreja ν_ε ja τ^2_ε lukuun ottamatta riippuu muista parametreista vain parametrien $\mathbf{w}$ ja $\boldsymbol{\beta}$ kautta, on sen täysehdollinen jakauma $f(\sigma^2_\varepsilon|\mathbf{w}, \boldsymbol{\beta})$, ja Bayesin lauseella saadaan

$$f(\sigma^2_\varepsilon|\mathbf{w}, \boldsymbol{\beta}) \propto f(\mathbf{w}, \boldsymbol{\beta}|\sigma^2_\varepsilon) f(\sigma^2_\varepsilon) = f(\mathbf{w}|\boldsymbol{\beta}, \sigma^2_\varepsilon) f(\boldsymbol{\beta}|\sigma^2_\varepsilon) f(\sigma^2_\varepsilon).$$

Koska SNP-vaikutusten $\boldsymbol{\beta}$ priori ei Bayes-mallissa riipu parametrista σ^2_ε pätee $f(\boldsymbol{\beta}|\sigma^2_\varepsilon) = f(\boldsymbol{\beta})$, jolloin täysehdollinen jakauma on

$$f(\sigma^2_\varepsilon|\mathbf{w}, \boldsymbol{\beta}) \propto f(\mathbf{w}|\boldsymbol{\beta}, \sigma^2_\varepsilon) f(\boldsymbol{\beta}) f(\sigma^2_\varepsilon) \propto f(\mathbf{w}|\boldsymbol{\beta}, \sigma^2_\varepsilon) f(\sigma^2_\varepsilon).$$

Ensimmäinen termi on matala-asteisen mallin uskottavuusfunktio ja jälkimmäinen on hajontaparametrin priori (skaalattu käänteinen χ^2-jakauma), näiden tulo on

$$\begin{aligned}
&(\sigma_\varepsilon^2)^{-\frac{q}{2}} \exp\left\{ -\frac{(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})'(\mathbf{w}-\mathbf{Q}\boldsymbol{\beta})}{2\frac{\sigma_\varepsilon^2}{n}} \right\} \left(\sigma_{\alpha_\varepsilon}^2\right)^{-\frac{2+\nu_\varepsilon}{2}} \exp\left(-\frac{\nu_\varepsilon \tau_\varepsilon^2}{2\sigma_\varepsilon^2} \right) \\
\propto &(\sigma_\varepsilon^2)^{-\frac{2+\nu_\varepsilon+q}{2}} \exp\left\{ -\frac{n\left(\mathbf{e}'\mathbf{e}\right) + \nu_\varepsilon \tau_\varepsilon^2}{2\sigma_\varepsilon^2} \right\}.
\end{aligned}$$

Samalla tavalla kuin edellisessä luvussa myös nyt nähdään, että kyseessä on skaalatun käänteisen χ^2-jakauman tiheysfunktion parametreista $\tilde{\nu}_\varepsilon$ ja $\tilde{\tau}_\varepsilon^2$ riippuva osa, ja nämä parametrit ovat

$$\begin{aligned}
&\tilde{\nu}_\varepsilon = \nu_\varepsilon + q, \\
&\tilde{\nu}_\varepsilon \tilde{\tau}_\varepsilon^2 = n\left(\mathbf{e}'\mathbf{e}\right) + \nu_\varepsilon \tau_\varepsilon^2 \Rightarrow \tilde{\tau}_\varepsilon^2 = \frac{n\left(\mathbf{e}'\mathbf{e}\right) + \nu_\varepsilon \tau_\varepsilon^2}{\tilde{\nu}_\varepsilon}.
\end{aligned}$$

Tämä poikkeaa Zhengin ja kumppanien esityksestä siinä, että heidän mukaansa $\tilde{\tau}_\varepsilon^2 = (\mathbf{e}'\mathbf{e} + \nu_\varepsilon \tau_\varepsilon^2)/\tilde{\nu}_\varepsilon$. Kuten aiemmin luvussa 3.3.2, Zhengin ja kumppanien antamaan jakaumaan päädytään pudottamalla matala-asteisen mallin uskottavuusfunktiosta hajontaparametrin σ_ε^2 kerroin $1/n$. On epäselvää, kuinka kertoimen $1/n$ poisjättäminen on perusteltua.

3.3.7 Apumuuttujan l_{jk} täysehdollinen jakauma

Aikaisemmin esitettiin, että apumuuttujalle l_{jk} voidaan muodostaa lineaarinen regressiomalli

$$l_{jk} = \mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc} + \varepsilon_{jk}, \quad \varepsilon_{jk} \sim N(0,1).$$

Jos muuttujalla l_{jk} ei olisi rajoituksia, sen täysehdollinen jakauma olisi normaalijakauma odotusarvolla $\mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{kc}$ ja varianssilla 1. Muuttuja l_{jk} määräytyy kuitenkin relaatiosta

$$z_{jk} = \begin{cases} 0, & \text{jos } l_{jk} > 0 \\ 1, & \text{jos } l_{jk} \leq 0. \end{cases}$$

Tällöin muuttujan l_{jk} täysehdollinen jakauma – jossa z_{jk} oletetaan tunnetuksi kiinteäksi arvoksi – on katkaistu normaalijakauma

$$l_{jk} \mid z_{jk}, \mu_k, \boldsymbol{\alpha}_k = \begin{cases} \mathrm{TN}(\mu_k + \mathbf{A}_j\boldsymbol{\alpha}_k,\, 1,\, 0,\, \infty), & \text{kun } z_{jk} = 0, \\ \mathrm{TN}(\mu_k + \mathbf{A}_j\boldsymbol{\alpha}_k,\, 1,\, -\infty,\, 0), & \text{kun } z_{jk} = 1. \end{cases}$$

Merkinnällä $\mathrm{TN}(a, b, c, d)$ tarkoitetaan normaalijakaumaa, jonka odotusarvo on a ja keskihajonta b, katkaistuna välille $[c, d]$. Tämä jakauma vastaa muuten sitä, jonka Zheng ja kumppanit esittävät artikkelissaan, paitsi että heillä tapaukset $z_{jk} = 0$ ja $z_{jk} = 1$ ovat toisin päin. Tämä on luultavasti kirjoitusvirhe, koska artikkelissa aikaisemmin määriteltiin, että $z_{jk} = 0$ kun l_{jk} on positiivinen.

3.4 Gibbsin otanta-algoritmi

Tässä luvussa käydään läpi eri vaiheet Gibbsin otanta-algoritmista, jolla hierarkkisen Bayes-mallin parametrit estimoidaan. Otanta-algoritmi on esitetty Zhengin ja kumppanien artikkelin menetelmäliitteessä käyttäen pseudokoodia, ja tämän luvun esitys perustuu siihen. Taulukossa 1 on suomenkielinen mukaelma pseudokoodista.

3.4.1 Alustusvaihe

Otanta-algoritmi alkaa snippien marginaalivaikutusten b_j skaalaamisella. Tiivistelmädataan perustuva malli (2) pohjautuu marginaaliefekteihin $\mathbf{b}$, jotka ovat yksikössä per standardisoitu genotyyppi. Nämä marginaaliefektit $\mathbf{b}$ estimoidaan GWAS-tunnusluvuista, jonka marginaaliefektit $\mathbf{b}^*$ ovat yksiköissä per genotyyppi skaalalla $0/1/2$. Algoritmin alussa suoritetaan skaalaus

$$b_j = s_j b_j^*, \quad s_j = \sqrt{\frac{\sigma_y^2}{N_j \sigma_j^2 + (b_j^*)^2}}.$$

Tässä σ_y^2 on fenotyypin varianssi, N_j on snipin j otoskoko (*per-SNP sample size*, kuinka monelta tilastoyksiköltä on mitattu kyseisen snipin arvo) ja σ_j on snipin j keskivirhe. Fenotyypin varianssi σ_y^2 estimoidaan suoraan GWAS-tunnusluvuista. Yksinkertaisuuden vuoksi snippikohtaisia otoskokoja N_j voidaan approksimoida kokonaisotoskoolla N.

Tämän jälkeen muodostetaan LD-lohkojen ominaisarvohajotelmat ja matalaasteinen malli (5). Viimeiseksi kaikki Bayes-mallin parametrit alustetaan aloitusarvoilla. Tämän jälkeen alkavat varsinaiset Gibbsin otannan iteraatiot, seuraavissa luvuissa esitettävät askeleet suoritetaan uudestaan jokaisella iteraatiolla.

3.4.2 Snippeihin $j \in \{1, ..., m\}$ liittyvät askeleet

Gibbsin otannassa suoritetaan seuraavat askeleet jokaiselle snipille $j \in \{1, ..., m\}$. Ensin lasketaan snippien yhteisvaikutuksiin liittyvät suureet

$$r_j = \mathbf{Q}_j' \mathbf{w}_{corr} + \beta_j \quad \text{ja} \quad C_j = 1 + \frac{\sigma_\varepsilon^2}{\gamma_k \sigma_g^2} \quad \text{jokaiselle } \gamma_k.$$

Suure $\mathbf{w}_{corr}$ vaatii selittämistä. Kuten aikaisemmin esitettiin, tarvitaan Gibbsin otannassa residuaalivektoria, josta on poistettu muiden kuin tarkastelun alla olevan snipin j vaikutus:

$$\mathbf{e}_{-j} = \mathbf{w} - \sum_{t \neq j} \mathbf{Q}_t \beta_t.$$

Tämän laskeminen uudelleen jokaisen snipin tapauksessa olisi kuitenkin laskennallisesti hyvin työlästä. Sen sijaan SBayesRC-menetelmässä käytetään juoksevaa residuaalivektoria $\mathbf{w}_{corr}$, jota päivitetään laskennallisesti tehokkaasti aina kun jokin

	Pseudokoodi
1	**Input:** GWAS-tunnusluvut, LD-referenssimatriisi ja annotaatiotiedosto.
2	Snippien marginaaliefektien skaalaus: $b_j = s_j b_j^*$
3	Muodosta LD-lohkojen ominaisarvohajotelmat ja matala-asteinen malli.
4	Alusta malliparametrit.
5	Jokaiselle iteraatiolle i **tee**
6	Jokaiselle snipille j **tee**
7	Laske $r_j = \mathbf{Q}_j' \mathbf{w}_{corr} + \beta_j$.
8	Laske $C_j = 1 + \frac{\sigma_\varepsilon^2}{\gamma_k \sigma_g^2}$ jokaiselle γ_k.
9	Laske posterioritodennäköisyydet snipin yhteisvaikutuksen sekoitejakauman ryhmäkuuluvuuksille ja poimi δ_j.
10	Poimi snipin yhteisvaikutus β_j sen täysehdollisesta jakaumasta $N\left(\frac{r_j}{C_j}, \frac{\sigma_\varepsilon^2}{C_j}\right)$.
11	Edellisessä vaiheessa poimitun β_j^{new} perusteella päivitä $\mathbf{w}_{corr}^{new} = \mathbf{w}_{corr}^{old} + \mathbf{Q}_j \left(\beta_j^{old} - \beta_j^{new}\right)$.
12	Aseta indikaattorimuuttujien $\mathbf{z}_j$ arvot indikaattorimuuttujan δ_j perusteella.
13	Indeksin j silmukka **päättyy**.
14	Jokaiselle $k \in \{2, 3, ..., \text{ryhmien määrä}\}$ **tee**
15	Jokaiselle snipille j, joka läpäisi rajan kyseiseen komponenttiin **tee**
16	Poimi apumuuttujan l_{jk} arvo sen täysehdollisesta jakaumasta (katkaistu normaalijakauma) ehdolla z_{jk}.
17	Indeksin j silmukka **päättyy**.
18	Jokaiselle annotaatiomuuttujalle c **tee**
19	Poimi annotaatiovaikutus α_{kc} sen täysehdollisesta jakaumasta ehdolla $\boldsymbol{l}_{jk}$.
20	Indeksin c silmukka **päättyy**.
21	Poimi annotaatiovaikutuksen hajontaparametri $\sigma_{\alpha_k}^2$ sen täysehdollisesta jakaumasta ehdolla $\boldsymbol{\alpha}_k$.
22	Indeksin k silmukka **päättyy**.
23	Laske $\widehat{\mathbf{w}} = \mathbf{Q}\boldsymbol{\beta}$ ja geneettinen varianssi $\sigma_g^2 = \widehat{\mathbf{w}}'\widehat{\mathbf{w}}$.
24	Estimoi per-SNP-heritabiliteettirikasteparametri jokaiselle annotaatiomuuttujalle.
25	Poimi residuaalin hajontaparametri σ_ε^2 sen täysehdollisesta jakaumasta jokaiselle lohkolle.
26	Indeksin i silmukka **päättyy**.
27	Skaalaa SNP-yhteisvaikutusten posteriorikeskiarvot takaisin per-alleeli-skaalalle kaavalla $\beta_j^* = s_j \hat{\beta}_j$.

Taulukko 1: Gibbsin otanta-algoritmi, mukaelma Zhengin ja kumppaninen esittämästä pseudokoodista.

yhteisvaikutus β_j muuttuu. Koska kaikkien snippien yhteisvaikutukset β_j alustetaan arvolla 0, on Gibbsin otannan alussa

$$\mathbf{w}_{corr} = \mathbf{w} - \mathbf{Q}\boldsymbol{\beta} = \mathbf{w} - \mathbf{0} = \mathbf{w}.$$

Juoksevan residuaalivektorin $\mathbf{w}_{corr}$ avulla saadaan laskettua suure r_j kuten yllä kaavalla $r_j = \mathbf{Q}_j'\mathbf{w}_{corr} + \beta_j$. Myöhemmin nähdään, kuinka vektoria $\mathbf{w}_{corr}$ päivitetään snippien yhteisvaikutusten muuttuessa.

Seuraavaksi suoritetaan otanta snipin j sekoitejakaumaryhmään kuuluvuudesta, eli poimitaan otos parametrin δ_j täysehdollisesta jakaumasta. Sen jälkeen suoritetaan otanta snipin j yhteisvaikutuksesta, eli poimitaan otos parametrin β_j täysehdollisesta jakaumasta, joka on normaalijakauma $N\left(r_j/C_j, \sigma_\varepsilon^2/C_j\right)$. Tämän jälkeen suoritetaan juoksevan residuaalivektorin $\mathbf{w}_{corr}$ päivitys. Olkoon snipin j juuri poimittu yhteisvaikutus β_j^{uusi} ja vanha yhteisvaikutus β_j^{vanha}. Juoksevan residuaalivektorin päivitetty arvo on nyt

$$\mathbf{w}_{corr}^{uusi} = \mathbf{w}_{corr}^{vanha} + \mathbf{Q}_j\left(\beta_j^{vanha} - \beta_j^{uusi}\right).$$

Lopuksi indikaattorimuuttujien z_{jk} eli vektorin $\mathbf{z}_j$ arvot asetetaan suoraan poimitun δ_j:n perusteella.

$$\begin{aligned}
\delta_j = 1 &\Rightarrow \mathbf{z}_j = \{0,0,0,0\},\\
\delta_j = 2 &\Rightarrow \mathbf{z}_j = \{1,0,0,0\},\\
\delta_j = 3 &\Rightarrow \mathbf{z}_j = \{0,1,0,0\},\\
\delta_j = 4 &\Rightarrow \mathbf{z}_j = \{0,0,1,0\},\\
\delta_j = 5 &\Rightarrow \mathbf{z}_j = \{0,0,0,1\}.
\end{aligned}$$

3.4.3 Sekoitejakaumaluokkiin $k \in \{2,3,4,5\}$ liittyvät askeleet

Seuraavaksi Gibbsin otannassa suoritetaan poiminta annotaatiomallin muuttujille, kaikki tässä luvussa esitetyt askeleet suoritetaan iteratiivisesti kaikille luokille $k \in \{2,3,4,5\}$. Aluksi käydään iteratiivisesti läpi ne snipit, jotka läpäisivät rajan kyseiseen komponenttiin (*...SNP:s that passed the bar for current component*), ja näille suoritetaan parametrien l_{jk} otanta parametrien täysehdollisista jakaumista ehdoilla z_{jk} (katkaistu normaalijakauma). Tämä kuulostaa ongelmalliselta, sillä se tarkoittaisi, että snipillä j parametri l_{jk} päivitettäisiin vain siinä ryhmässä k, johon snippi sillä hetkellä kuuluu, ja muissa ryhmissä vanha parametrin l_{jk} arvo pysyisi voimassa. Jokaisen snipin j tapauksessa ryhmäkuuluvuus $\mathbf{z}_{jk}$ määrittää mitä arvoja latentit muuttujat l_{jk} voivat saada kaikilla k, jolloin ei ole riittävää, että parametri l_{jk} päivitetään jokaisella j vain yhden ryhmän k tapauksessa. Siksi onkin mahdollista, että pseudokoodia kirjoittaessa on tapahtunut virhe, ja oikeasti parametrit l_{jk} päivitetään kaikissa indeksien j ja k kombinaatioissa.

Tämän jälkeen käydään iteratiivisesti läpi kaikki annotaatiovaikutukset $\alpha_{kc,}c \in \{1,...,C\}$, ja jokaiselle suoritetaan otanta parametrin täysehdollisesta jakaumasta, joka on aikaisemmin määritelty normaalijakauma $N\left(r_{kc}/C_{kc}, 1/C_{kc}\right)$. Viimeisenä suoritetaan otanta hajontaparametrille $\sigma_{\alpha_k}^2$ sen täysehdollisesta jakaumasta, joka on aikaisemmin määritelty skaalattu käänteinen χ^2-jakauma.

3.4.4 Muut parametrit

Ennen uuden iteraatiokierroksen alkua suoritetaan vielä otanta muille parametreille. Ensimmäisenä lasketaan vektori $\widehat{\mathbf{w}} = \mathbf{Q}\boldsymbol{\beta}$ ja tämän perusteella estimaatti geneettiselle varianssille $\sigma_g^2 = \widehat{\mathbf{w}}'\widehat{\mathbf{w}}$. Tämän jälkeen lasketaan jokaiselle annotaatiomuuttujalle per-SNP-heritabiliteettirikaste. Heritabiliteettirikasteen laskenta on esitetty tutkielman liitteessä B. Viimeisenä suoritetaan LD-lohkokohtainen otanta matalaasteisen mallin residuaalien hajontaparametrille σ_ε^2 sen täysehdollisesta jakaumasta, joka on aikaisemmin esitetty skaalattu käänteinen χ^2-jakauma.

3.4.5 Yhteisvaikutusten β_j skaalaus Gibbsin otannan jälkeen

Gibbsin otanta-algoritmin alussa snippien GWAS-tunnusluvuista saadut marginaalivaikutukset b_j^* skaalattiin kaavalla $b_j = s_j b_j^*$. Gibbsin otannan päätyttyä snippien yhteisvaikutusten estimaatit (posteriorikeskiarvot) $\widehat{\beta}_j$ täytyy vielä skaalata takaisin alkuperäiselle asteikolle. Taulukon 1 Zhengin ja kumppanien vastaavaa mukailevassa pseudokoodissa rivillä 27 tämä tehdään kaavalla $\widehat{\beta}_j^* = s_j\widehat{\beta}_j$. Kun katsotaan Gibbsin otannan alussa tehtyä skaalausta kaavalla $b_j = s_j b_j^*$, otannan lopussa tehtävän skaalauksen luulisi aritmeettisesti olevan $\widehat{\beta}_j^* = \widehat{\beta}_j / s_j$, ja on luultavaa, että skaalaus tehdään oikeasti tällä kaavalla.

4 Polygeenisten riskisummien hyödyntäminen verenvuototapahtumien ennustamisessa

Tutkielmassa on tähän mennessä käsitelty teoreettisesti polygeenista prediktiota luvussa 2 ja SBayesRC-menetelmää luvussa 3. Tässä luvussa polygeenista prediktiota ja SBayesRC-menetelmää hyödynnetään sovelluksessa, jossa kokeillaan polygeenisten riskisummien hyödyllisyyttä ennustettaessa kallon- ja suolistonsisäistä verenvuotoa verenohennuslääkitystä käyttävillä FinnGen-tutkimusprojektin yksilöillä. Aluksi esitellään tutkimuskysymys, jonka jälkeen selitetään, kuinka tutkimusyksilöt valittiin ja kovariaatit muodostettiin. Sen jälkeen käsitellään sovelluksessa käytettävät GWAS-tunnusluvut, LD-referenssitiedosto ja annotaatiotiedosto. Tämän jälkeen esitetään elinaika-aineistojen muodostus ja sovelluksen elinaika-analyysin yksityiskohdat. Viimeisenä esitetään analyysin tulokset ja johtopäätökset.

4.1 Tutkimuskysymys

Verenohennuslääkkeitä eli antikoagulantteja käytetään yleisesti veritulppauman ehkäisyssä ja hoidossa [28]. Lääkitykset ovat vaikutukseltaan tehokkaita, mutta ne ovat myös yhteydessä merkittävään verenvuototapahtuman riskiin. Myös monia muita kliinisiä tekijöitä, kuten korkea ikä ja anemia, on yhdistetty verenvuodon lisääntyneeseen riskiin. Yksilön verenvuotoriskin määrittämisen avuksi on tehty kliinisiä riskinarviotyökaluja, mutta näiden kyky ennustaa tehokkaasti verenohennuslääkitykseen liittyviä verenvuotoja on toistaiseksi ollut vaatimaton.

Tutkielman sovelluksen tutkimuskysymyksenä oli, auttaako polygeenisen riskisumman (PRS-muuttujan) käyttö ennustamaan kallon- tai suolistonsisäistä veren-

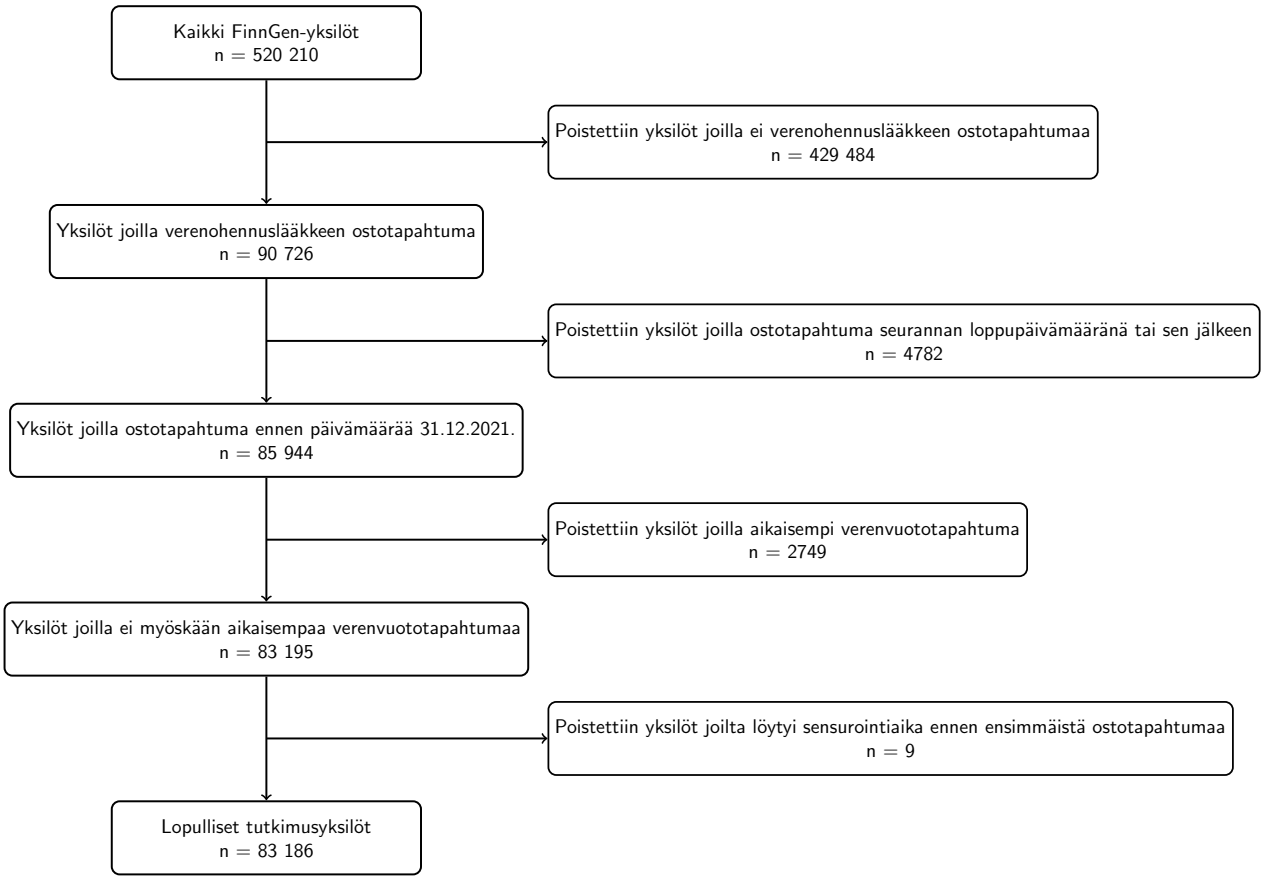
vuototapahtumaa verenohennuslääkitystä käyttävillä yksilöillä. Tutkimuskysymykseen vastattiin estimoimalla SBayesRC-menetelmällä 7 315 258 snipille joko kallon- tai suolistonsisäiseen verenvuotoon liittyvät yhteisvaikutukset, kummassakin tapauksessa käyttäen GWAS-tunnuslukuja kahdesta vaihtoehtoisesta GWAS-tutkimuksesta. Yhteisvaikutukset estimoitiin myös PRS-CS-menetelmällä, jotta SBayesRC-menetelmällä saatuja tuloksia voitaisiin vertailla vanhemmalla menetelmällä saatuihin tuloksiin. Estimoitujen yhteisvaikutusten avulla laskettiin suomalaisen FinnGen-aineiston osa-aineistolle kyseisiin verenvuototapahtumiin liittyvät polygeeniset riskisummat. Osa-aineisto muodostettiin verenohennuslääkettä käyttäneistä yksilöistä. Riskisummien yhteyttä uhkaan kohdata kallon- tai suolistonsisäinen verenvuototapahtuma tutkittiin elinaika-analyysissä, jossa vastemuuttujana oli aika kyseiseen verenvuototapahtumaan päivinä alkaen verenohennuslääkityksen ostotapahtumasta. Analyyseissä käytettiin Coxin regressiomallia, jossa selittävinä muuttujina olivat polygeeninen riskisumma ja joukko muita kovariaatteja. Mikäli PRS-muuttujan ja vastemuuttujan väliltä löytyi tilastollisesti merkitsevä yhteys, PRS-muuttujan tuomaa lisää mallin ennustetarkkuuteen jatkotarkasteltiin käyttäen konkordanssi- eli C-indeksiä.

4.2 FinnGen-aineisto

FinnGen on vuonna 2017 käynnistynyt tutkimushanke, joka pyrkii lisäämään ymmärrystä tavallisten sairauksien geneettisestä taustasta hyödyntäen suomalaisten genomitietoa ja terveysrekistereistä kerättyä sairaustietoa [24]. Tutkimushanke on esikilpailullinen ja tieteellinen, ja sen osapuolina ovat suomalaiset yliopistot, hyvinvointialueet, Veripalvelu, THL ja yhdeksän kansainvälistä lääkeyritystä. FinnGenin terveystiedot perustuvat suomalaisiin terveysrekistereihin, ja genomitiedot on analysoitu suomalaisten biopankkien DNA-näytteistä. Tämän tutkielman kirjoitushetkellä FinnGen-hanke sisältää yli 500 000 suomalaisen genotyyppi- ja fenotyyppitiedot [9].

Tutkielmassa käytettävä osa-aineisto muodostettiin valitsemalla mukaan yksilöt, jotka olivat Kelan lääkekorvausrekisterin mukaan ostaneet jotain seuraavista verenohennuslääkkeistä: varfariini (B01AA03), dabigatraanieteksilaatti (B01AE07), rivaroksabaani (B01AF01), apiksabaani (B01AF02) tai edoksabaani (B01AF03). Tällaisia yksilöitä oli 90 726 kappaletta. Joukosta karsittiin yksilöt, jotka olivat ostaneet verenohennuslääkettä ensimmäistä kertaa vasta elinaika-analyysin seurannan loppupäivämääränä tai sen jälkeen. Näin poistettiin 4782 yksilöä, jolloin jäljelle jäi 85 944 yksilöä. Seurannan lopetuspäivämääräksi valittiin 31.12.2021, koska FinnGen-aineiston kuolinrekisteri päättyi siihen. Seurannan aloitusajankohtana voidaan pitää päivämäärää 17.12.1992, joka on Kelan lääkeostorekisterin aikaisimman ostotapahtuman päivämäärä. Osa-aineistosta jätettiin pois myös yksilöt, joilta löytyi kallon- tai suolistonsisäinen verenvuototapahtuma yksilön ensimmäisen lääkeostotapahtuman päivämääränä tai tätä aikaisemmin. Näin poistettiin 2749 yksilöä, jolloin jäljelle jäi 83 195 yksilöä. Tämä tehtiin, koska käytetty FinnGenin sairauspäätepisterekisteri [11] sisältää vain jokaisen yksilön ensimmäisen vastetapahtuman. Tämä tarkoittaa, että jos kyseinen verenvuototapahtuma tapahtui ennen seuranta-ajan alkua (ensimmäinen lääkeostotapahtuma), oli yksilölle mahdotonta löytää verenvuototapahtu-

maa seuranta-aikana, jolloin seurannassa olisi ollut yksilöitä, joiden oli mahdotonta kohdata vastetapahtuma. Aineistosta poistettiin vielä 9 yksilöä, joilta löytyi FinnGenin kovariaattirekisteristä sensurointiaika ennen ensimmäistä lääkeostotapahtumaa. Aineiston karsinnan jälkeen jäljelle jäi 83 186 verenohennuslääkettä ostanutta yksilöä, jotka muodostivat analyysin tilastoyksiköt. Samoja tilastoyksiköitä käytettiin sekä analyysissä, jossa vastemuuttujana oli aika kallonsisäiseen verenvuototapahtumaan, että analyysissä, jossa vastemuuttujana oli aika suolistonsisäiseen verenvuototapahtumaan. Tutkimusyksilöiden valikoitumisprosessi on esitetty kuvaajassa 4.



Kuva 4: Vuokaavio tutkimusjoukon muodostuksesta.

4.3 Vastetapahtuma- ja kovariaattitietojen haku

Analyysissä käytetyt kovariaatit polygeenisten riskisummien lisäksi olivat ikä, sukupuoli, aikaisempi psykiatrinen häiriö, dementia, maksakirroosi tai maksan vajaatoiminta, munuaisen vajaatoiminta tai dialyysi, alkoholin väärinkäyttö, aivohalvaus, diabetes, sydämen vajaatoiminta, rasva-aineenvaihdunnan häiriö, korkea verenpaine sekä verisuonitauti. Verenohennuslääkkeen ostotapahtuman (haettiin Kelan lääkeostorekisteristä) yhteydessä oli myös tieto yksilön iästä (vuosina) ostohetkellä, josta saatiin yksilön ikä jokaiselle lääkeostotapahtuman hetkelle. Ikä kategorisoitiin kolmiarvoiseksi muuttujaksi: alle 65-vuotiaat, 65-75-vuotiaat sekä yli 75-vuotiaat.

FinnGenin kovariaattirekisteri sisältää sukupuolimuuttujan, jonka arvo oli saatavilla kaikille paitsi 2129 yksilölle. Näille yksilöille sukupuolimuuttujan arvoksi annettiin "tuntematon", koska R:n coxph-funktio, jolla Coxin regressiomalli sovitettiin, pudottaa pois havaintorivit, joissa on puuttuvia arvoja selittävissä muuttujissa, ja yksilöt haluttiin pitää mukana analyysissä.

Muiden kovariaattien kohdalla yksilöille etsittiin rekistereistä aikaisin ajankohta, jossa kyseisen kovariaatin positiivinen arvo oli diagnosoitu tai muuten todennettu. Näiden diagnoosi- ja todennusaikojen avulla myöhemmin luvussa 4.8 käsiteltäviin elinaika-ainestoihin muodostettiin binääriset kovariaattimuuttujat. Muuttujissa arvo 1 merkitsee, että yksilöllä oli todennettu kovariaatin positiivinen arvo kyseisen lääkkeenkäyttöjakson aloituspäivänä tai aikaisemmin, ja arvo 0 merkitsee, että positiivinen arvo oli todennettu vasta myöhemmin tai ei ollenkaan. Yksilöiden diagnoosiaikoja etsittiin ensisijaisesti erikoissairaanhoidon hoitoilmoitusrekisteristä (HILMO) sekä avohoitoilmoitusrekisteristä (Avohilmo) kovariaatteja vastaavilla ICD-10-koodeilla. Osalle kovariaateista diagnoosiaikoja etsittiin näistä rekistereistä myös kovariaattia vastaavilla ICPC-2-koodeilla. Lisäksi osalle kovariaateista todennusaikoja etsittiin myös Kelan lääkeostorekisteristä kovariaatteihin liittyvien ATC-koodien avulla (jos yksilö osti kovariaattiin liittyvää lääkitystä tietyllä hetkellä, todensi tämä kovariaatin positiivisen arvon sillä hetkellä). Osalle kovariaateista todennusaikoja etsittiin vielä Kelan korvausrekisteristä kovariaatteihin liittyvillä koodeilla (jos yksilö sai tiettyyn kovariaattiin liittyvää korvausta, todensi tämä kovariaatin positiivisen arvon sillä hetkellä). Käytetyt kovariaattien hakukoodit löytyvät taulukosta 2. Jos samalle yksilölle löytyi kovariaatista monta diagnoosi- tai todennusaikaa, näistä valittiin kaikista aikaisin.

ICD-10 (*International Statistical Classification of Diseases and Related Health Problems, 10th Revision*) on WHO:n kehittämä kansainvälinen tautiluokitusjärjestelmä, ICPC-2 (*International Classification of Primary Care, Second Edition*) on perusterveydenhuollossa käytettävä luokitusjärjestelmä ja ATC (Anatomical Therapeutic Chemical Classification System) on WHO:n kehittämä lääkeluokitusjärjestelmä. Taulukon 2 ICD-10- ja ATC-koodit ovat koodien alkuosia. Esimerkiksi rasvaaineenvaihdunnan häiriön tapauksessa diagnooseja etsittiin HILMO- ja Avohilmorekistereistä hakemalla ICD-10-koodeja, joiden alku on E78, ja Kelan lääkeostorekisteristä hakemalla ATC-koodeja, joiden alku on C10. ICD-10 ja ICPC-2 hakujen kohdalla mukaan otettiin myös diagnoosit, joissa kyseistä koodia vastaava diagnoosi ei ollut ensisijainen diagnoosi.

Verenvuototapahtumat haettiin FinnGenin sairauspäätepisterekisteristä [11]. Rekisterin päätetapahtumat on muodostettu kansallisista terveysrekistereistä saadun tiedon avulla. Kallonsisäiset verenvuototapahtumat haettiin rekisteristä koodilla I9_INTRACRA ja suolistonsisäiset verenvuototapahtumat koodilla K11_GIBLEEDING. FinnGenin Risteys-verkkosivulla [10] on tarkemmin määritelty, miten eri sairauspäätetapahtumat on muodostettu, tutkielmassa käytettiin risteyksen versiota R12. Taulukossa 3 on kunkin kovariaatin kohdalla kuvattu, kuinka monella yksilöllä oli kovariaatista positiivinen arvo yksilön ensimmäisen lääkejakson aloitushetkellä, toisin sanoen, kuinka monella yksillöllä oli kovariaatin diagnoosi- tai todennusaika, joka oli ensimmäisen lääkejakson aloituspäivä tai sitä aikaisempi päivä. Taulukossa 3 on myös näiden yksilöiden suhteellinen osuus kaikista yksilöistä.

Kovariaatti	**ICD-10**	**ICPC-2**	**Kelan korvauskoodi**	**ATC**
Verisuonitauti	I20–I25, I65–I66, I672, I70	K74, K75, K76, K91, K92	206	–
Korkea verenpaine	I10–I15	K85, K86, K87	205	C03A, C03B, C03DB, C03EA, C07A, C08CA, C08D, C09
Rasva-aineenvaihdunnan häiriö	E78	T93	206	C10
Sydämen vajaatoiminta	I50, I110, I130, I132	K77	201	–
Diabetes	E10–E14	T89, T90	103, 215	A10
Aivohalvaus	I63, I64, I693–I698	K90	–	–
Alkoholin väärinkäyttö	F10	–	–	–
Munuaisen vajaatoiminta tai dialyysi	N18, Z49	–	–	–
Maksan vajaatoiminta tai kirroosi	K702–K704, K717, K718, K72, K74		–	–
Dementia	F00–F03, G30	–	–	–
Psykiatrinen häiriö	F04–F99	–	–	–

Taulukko 2: Kovariaatit ja käytetyt hakukoodit

4.4 Tutkielmassa käytetyt GWAS-tunnusluvut

Luvussa 2.4 käsiteltiin GWAS-tutkimuksia ja sitä, kuinka näiden tulokset – mukaanlukien snippien marginaalivaikutukset – kootaan GWAS-tunnusluvuiksi. GWAS Catalog -verkkosivulle [15] on kerätty eri vastemuuttujiin liittyviä GWAS-tutkimuksia ja -tunnuslukuja. Tutkielmaa varten sivustolta etsittiin GWAS-tutkimuksia, joissa

Kovariaatti	Pos. arvojen lukumäärä	Suhteellinen osuus (%)
Verisuonitauti	22665	27
Kohonnut verenpaine	66743	80
Dyslipidemia	41302	50
Sydämen vajaatoiminta	9556	11
Diabetes	17293	21
Aivohalvaus	6894	8
Alkoholin väärinkäyttö	2549	3
Munuaisen vajaatoiminta tai dialyysi	1481	2
Maksan vajaatoiminta tai kirroosi	357	0.4
Dementia	1505	2
Psykiatrinen häiriö	12706	15

Taulukko 3: Niiden tutkimusyksilöiden lukumäärät, joilla oli kovariaatin mukainen diagnoosi ensimmäisen lääkejakson alussa. Suhteellinen osuus on laskettu kaikista 83 186 tutkimusyksilöstä.

vastemuutujana on joko kallon- tai suolistonsisäinen verenvuoto, ja joiden GWAS-tunnusluvut olivat vapaasti saatavissa. Kriteereinä valinnoille olivat, kuinka uusi GWAS-tutkimus oli sekä GWAS-tutkimuksen otoskoko. Oli myös toivottavaa, että GWAS-tutkimuksen yksilöt olisivat syntyperältään mahdollisimman lähellä yksilöitä, joille polygeeniset riskisummat lasketaan, toisin sanoen eurooppalaisia. Lisäksi kriteerinä oli, ettei GWAS-tutkimuksessa oltu käytetty FinnGen-aineistoa, koska muuten polygeenisten riskisummien tehoa olisi testattu yksilöihin, joita oli käytetty myös GWAS-tunnuslukujen kautta riskisummien muodostamiseen. Tämä olisi verrattavissa tilanteeseen, jossa mallin toimivuutta testattaisiin samoihin yksilöihin, joista malli on estimoitu. Lopulta tutkielmaan otettiin mukaan yhteensä neljät GWAS-tunnusluvut kahdesta eri tutkimuksesta, jotka esitellään seuraavaksi.

4.4.1 Zhou ja kumppanit (2018)

Vuonna 2018 julkaistussa tutkimuksessaan Zhou ja kumppanit tuottivat GWAS-tunnusluvut 1403 phecode-johdetulle binääriselle fenotyypille käyttäen UK biopankin dataa [39]. Aineisto sisälsi 408961 valkoihoista, eurooppalaista syntyperää olevaa yksilöä Britanniasta. Phe-koodit (*phecode*) ovat manuaalisesti muodostettuja samaa konseptia (kuten fenotyyppi) kuvaavien ICD-koodien ryhmiä [2]. Analyyseissä on käytetty SAIGE (*Scalable and Accurate Implementation of GEneralized mixed model*) -menetelmää, jonka Zhou ja kumppanit esittelevät artikkelissaan. Artikkelin mukaan toisin kuin kaksi usein käytettyä menetelmää (lineaarinen sekamalli ja logistinen sekamalli), SAIGE antaa tarkkoja p-arvoja tilanteessa, jossa tapausryhmän koko on huomattavasti pienempi kuin verrokkiryhmän, kuten tilanne usein on. Lisäksi SAIGE ottaa artikkelin mukaan huomioon yksilöiden välisen sukulaisuuden, joka on huomattava sekoittava tekijä geneettisissä assosiaatiotutkimuksissa. Analyyseissä käytettiin noin 28 miljoonaa osittain imputoitua snippiä.

Zhoun ja kumppanien GWAS-tutkimuksessa [39] muodostettiin GWAS-tunnusluvut sekä kallonsisäiselle- että suolistonsisäiselle verenvuodolle. Kallonsisäisen verenvuodon GWAS-tunnusluvut löytyivät GWAS-katalogista [15] koodilla GCST-

90436134, tunnuslukujen raportoiduksi piirteeksi oli merkitty intracranial hemorrhage (PheCode 430). Tapausryhmän koko oli 1796, verrokkiryhmän koko oli 399017 ja yksilöiden määrä yhteensä oli 400813. GWAS-tunnusluvuista löytyi sekä harmonisoitu että harmonisoimaton versio, tässä tutkielmassa käytettiin harmonisoitua versiota. Tunnuslukutiedoston nimi oli GCST90436134.h.tsv.gz.

Zhoun ja kumppanien suolistonsisäisen verenvuodon GWAS-tunnusluvut löytyivät GWAS-katalogista koodilla GCST90436398. Tunnuslukujen raportoiduksi piirteeksi oli merkitty gastrointestinal hemorrhage (PheCode 578). Tapausryhmän koko oli 21137, verrokkiryhmän koko oli 385157 ja yksilöiden määrä yhteensä oli 406294. Myös suolistonsisäisen verenvuodon tapauksessa valittiin harmonisoitu GWAS-tunnuslukutiedosto, jonka nimi oli GCST90436398.h.tsv.gz.

4.4.2 Jiang ja kumppanit (2021)

Toinen tässä tutkielmassa käytetty GWAS-tutkimus oli Jiangin ja kumppanien vuonna 2021 julkaistu tutkimus [19]. Myös tässä tutkimuksessa käytettiin UK biopankin dataa, ja aineisto koostui 456348 eurooppalaista syntyperää olevasta yksilöstä. Tutkimuksessa tuotettiin GWAS-tunnusluvut 2989 binääriselle piirteelle, ja analyyseissä käytettiin 11 842 647:ää osittain imputoitua snippiä. Artikkelissaan Jiang ja kumppanit esittelevät analyyseissä käytetyn yleistettyyn lineaariseen sekamalliin pohjautuvan menetelmän nimeltään fastGWA-GLMM. Kirjoittajien mukaan myös tämä menetelmä on toimiva tilanteessa, jossa tapausryhmän koko on huomattavasti pienempi kuin verrokkiryhmän.

Jiangin ja kumppanien GWAS-tutkimuksessa [19] muodostettiin GWAS-tunnusluvut kallonsisäiselle verenvuodolle, jotka löytyivät GWAS-katalogista [15] koodilla GCST90043996. Raportoiduksi piirteeksi oli merkitty intracranial hemorrhage (PheCode430). Tapausryhmän koko oli 158, verrokkiryhmän koko oli 456190 ja yksilöitä oli yhteensä 456348. Myös näistä GWAS-tunnusluvuista löytyi sekä harmonisoitu että harmonisoimaton versio, tässä tutkielmassa käytettiin harmonisoitua versiota. Tiedoston nimi oli 34737426-GCST90043996-EFO_0000551.h.tsv.gz.

GWAS-katalogin mukaan Jiangin ja kumppanien tutkimuksessa ei valitettavasti laskettu GWAS-tunnuslukuja piirteelle gastrointestinal hemorrhage (PheCode 578), mutta sen sijaan tunnusluvut laskettiin verrattavissa olevalle piirteelle hematemesis (PheCode 578.1) eli veren oksentamiselle. Nämä tunnusluvut löytyivät GWAS-katalogista koodilla GCST90044208, ja ne valittiin viimeisiksi käytettäviksi GWAS-tunnusluvuiksi. Tapausryhmän koko oli 1377, verrokkiryhmän koko oli 454971 ja yksilöitä oli yhteensä 456348. Tässäkin tapauksessa valittiin harmonisoitu versio tunnusluvuista tiedostonimellä 34737426-GCST90044208-HP_0002248.h.tsv.gz.

4.5 Tutkielmassa käytetty LD-referenssitiedosto

SBayesRC-menetelmä tarvitsee toimiakseen GWAS-tunnuslukujen lisäksi LD-referenssitiedoston. Zheng kumppaneineen on laskenut kahdelle eri snippikokoelmalle (HapMap3 SNPs ja imputed SNPs) LD-referenssit käyttäen kolmea eri UK biopankin datasta löytyvää syntyperää (Eurooppa, Itä-Aasia ja Afrikka). Nämä referenssitiedostot ovat ladattavissa SBayesRC-menetelmän GitHub-verkkosivulta [38]. Tässä tutkielmassa käytettiin näistä imputed SNPs -referenssitiedostoa, joka on laskettu

eurooppalaista syntyperää olevista UK biopankin yksilöistä. Tämä LD-referenssi käsittää 7 356 518 snippiä, joita Zheng ja kumppanit käyttivät suuressa osassa tutkimuksensa analyysejä.

Usea tekijä puolsi tämän LD-referenssitiedoston käyttöä. Saavutettavan prediktiokyvyn kannalta LD-referenssin tulisi olla laskettu samaa syntyperää olevilta yksilöiltä kuin GWAS-tutkimuksen yksilöt [38]. Koska sekä molemmat tutkielmassa käytetyt GWAS-tunnusluvut että käytetty LD-referenssi on laskettu UK biopankin eurooppalaista syntyperää olevista yksilöistä, tämä ehto täyttyi. Lisäksi LD-referenssitiedostot eivät aina ole samassa formaatissa. Koska tämä referenssitiedosto oli tuotettu Zhengin ja kumppanien SBayesRC-menetelmää käsittelevän tutkimuksen [38] yhteydessä, oli se suoraan yhteensopiva menetelmän kanssa. Lisäksi LD-referenssi sisälsi varsin suuren määrän snippejä (7 356 518). Yleisesti ottaen suurempi määrä snippejä on prediktiokyvyn kannalta suotavaa. Esimerkiksi Zhengin ja kumppanien tutkimus sisälsi useita analyysejä, joissa verrataan SBayesRC-menetelmän prediktiokykyä käytettäessä joko noin miljoonaa tai noin seitsemää miljoonaa snippiä, ja niistä lähes kaikissa prediktiokyky oli parempi käytettäessä suurempaa määrää snippejä.

4.6 Tutkielmassa käytetty annotaatiotiedosto

Zheng ja kumppanit korostavat SBayesRC-menetelmää käsittelevässä artikkelissaan funktionaalisten annotaatioiden merkitystä polygeenisen ennustetarkkuuden lisäämisessä [38]. Myös tässä tutkielmassa haluttiin hyödyntää SBayesRC-menetelmän ominaisuutta käyttää snippeihin liittyvää funktionaalista annotaatiotietoa yhteisvaikutusten estimoinnissa.

Tässä tutkielmassa käytettiin samoja funktionaalisia annotaatioita kuin Zheng ja kumppanit käyttivät tutkimuksessaan [38]. Tämä annotaatiokokonaisuus, BaselineLD-malli v2.2, sisältää 96 annotaatiomuuttujaa noin kahdeksalle miljoonalle snipille. Nämä funktionaaliset annotaatiot ovat peräisin Gazalin ja kumppanien vuonna 2017 julkaistusta tutkimuksesta [13]. Tutkimusartikkelin mukaan annotaatiomuuttujia pitäisi BaselineLD-mallissa olla 75, mutta selvästikin annotaatioita on lisätty artikkelin julkaisun jälkeen. Tähän viittaa myös annotaatiomallin nimeen (BaselineLD-malli v2.2) sisältyvä versionumero. Gazalin ja kumppanien BaselineLD-malli sisälsi Finucanen ja kumppanien vuonna 2015 julkaistun tutkimuksen "full baseline -mallin" annotaatiot, joita oli 53 mukaanlukien annotaatio kaikista snipeistä [12].

Full baseline -mallin (ja näin ollen myös BaselineLD-mallin) 53 annotaatiota sisältävät 24 pääannotaatiota ja näistä johdettuja annotaatioita [12]. Pääannotaatioihin kuuluu mm. annotaatioita, jotka kertovat kuuluuko snippi tietynlaiseen alueeseen perimässä. Tällaisia alueita ovat mm. koodaavat alueet (*coding regions*), UTR-eli kääntämättömät alueet (*UTRs, Untranslated regions*), promoottorialueet (*promoter regions*) ja intronialueet (*intronic regions*). Muita pääannotaatioita ovat esimerkiksi histonimodifikaatioihin (*histone marks*) ja tehostaja-alueisiin (*enhancers*) liittyvät annotaatiot. Full baseline -malli sisältää myös jokaisesta pääannotaatiosta johdetun annotaation, joka kertoo onko snippi enintään 500:n emäsparin päässä pääannotaatioalueesta kuitenkaan kuulumatta siihen, toisin sanoen onko snippi pääannotaatioalueen läheisyydessä. Gazalin ja kumppanien BaselineLD-mallissa

full baseline -mallin annotaatioihin lisättiin mm. LD:hen liittyviä annotaatioita, kuten alleelin MAF-korjattu estimoitu ikä [13]. MAF (*minor allele frequency*) on snipin vähemmän yleisen alleelin omaavien ihmisten suhteellinen osuus populaatiossa. BaselineLD-malli sisältää myös 10 MAF-annotaatiota, jotka kertovat mihin luokkaväliin snipin MAF kuuluu (0–0.1, 0.1–0.2, ... , 0.9–1).

4.7 FinnGen-yksilöiden polygeeniset riskisummat

SBayesRC-menetelmään perustuvien polygeenisten riskisummien laskeminen FinnGen-datan yksilöille suoritettiin R-ohjelmalla käyttäen sbayesrc-paketin funktioita [38] sekä PRS-CS -menetelmän tapauksessa FinnGenin virtuaaliympäristön valmiita työkaluja. Täysi selostus riskisummien laskemisesta löytyy tutkielman liitteestä C. Lopputuloksena oli $2 \times 2 \times 2 \times 2 = 16$ riskisummaa eli PRS-muuttujaa (kallon- ja suolistonsisäiselle verenvuodolle, Jiangin ja kumppanien ja Zhengin ja kumppanien GWAS-tunnuslukuihin perustuvat, SBayesRC- ja PRS-CS-menetelmillä lasketut sekä jatkuva ja kategorisoitu versio PRS-muuttujasta). Kategorisoitu versio PRS-muuttujasta oli viisitasoinen ja kertoi, kuuluuko PRS-arvo kvantiiliväliin 0 – 0.025, 0.025 – 0.2, 0.2 – 0.8, 0.8 – 0.975 vai 0.975 – 1. Kategorisen PRS-version avulla tutkittiin mahdollisuutta, että PRS-muuttujien ääriarvot ovat erityisen merkityksellisiä riskiyksilöiden tunnistamisessa.

4.8 Elinaika-aineistot

Tutkielman sovelluksen tutkimuskysymyksenä oli, auttaako PRS-muuttujan käyttö ennustamaan kallon- ja suolistonsisäisiä verenvuototapahtumia verenohennuslääkitystä käyttävillä yksilöillä. Yksilöt saattoivat kuitenkin lopettaa verenohennuslääkkeen käytön pitkäksikin aikaa ja aloittaa lääkkeen käytön jälleen myöhempänä ajankohtana. Yksilön ensimmäinen lääkkeenkäyttöjakso alkoi ensimmäisestä verenohennuslääkkeen ostopäivästä. Yksilön katsottiin olevan verenohennuslääkityksen vaikutuksen alaisena, jos yksilön viimeisestä lääkeostosta oli kulunut enintään 120 vuorokautta. Verenohennuslääkettä myydään kerrallaan maksimissaan kolmen kuukauden tarpeiksi eli noin 90 vuorokaudeksi. Kun tähän lisättiin 30 vuorokauden lisäaika (on mahdollista etteivät yksilöt käyneet ostamassa uutta lääke-erää välittömästi vanhan loputtua tai että yksilöt jättivät joinain vuorokausina lääkkeen ottamatta, jolloin lääkemäärä riitti pidemmäksi aikaa), päädyttiin 120 vuorokauteen. Kun yksilön edellisestä lääkkeenostotapahtumasta oli kulunut yli 120 vuorokautta, yksilön katsottiin lopettaneen lääkkeen käytön. Jos yksilö osti myöhemmin uudestaan verenohennuslääkitystä, alkoi tästä uusi lääkkeenkäyttöjakso. Elinaika-analyysin seurantajaksot koostuivat yksittäisistä lääkkeenkäyttöjaksoista, joita samalla yksilöllä saattoi olla monia ja jotka loppuivat joko vastetapahtumaan (kallon- tai suolistonsisäinen verenvuoto) tai sensurointitapahtumaan (lääkkeen käytön lopettaminen, kuolema, seuranta-ajan loppuminen 31.12.2021 tai muu sensurointisyy). Vastetapahtuman kohdannut yksilö poistui seurannasta lopullisesti.

Kallonsisäiselle verenvuodolle ja suolistonsisäiselle verenvuodolle muodostettiin erikseen omat datakehikot, joissa jokainen rivi vastasi yhtä yksilön lääkejaksoa. Datakehikko sisälsi yksilötunnuksen, yksilön lääkkeenkäyttöjakson järjestysnumeron,

ID	Lääkejakso	TIME1	TIME2	EVENT	PRS 1	Kovariaatti 1	Kovariaatti 2	...
1	1	0	402	0	1.22	0	1	...
1	2	0	370	1	1.22	0	1	...
2	1	0	220	0	-1.55	0	0	...
2	2	0	780	0	-1.55	1	0	...
2	3	0	550	0	-1.55	1	0	...

Taulukko 4: Elinaika-aineiston rakenne

muuttujan TIME1 (lääkkeenkäyttöjakson aloitusaika, arvo aina 0), muuttujan TIME2 (lääkkeenkäyttöjakson kesto päivinä), muuttujan EVENT (binäärimuuttuja, joka saa arvon 1, jos lääkkeenkäyttöjakso päättyi vastetapahtumaan ja 0, jos jakso päättyi sensurointiin) sekä kovariaattimuuttujia, jotka sisälsivät yksilölle lasketut polygeeniset riskisummamuuttujat. Binäärisillä kovariaateilla arvo 1 tarkoittaa, että yksilöllä oli diagnosoitu tai muuten todennettu positiivinen arvo kyseisestä kovariaatista lääkejakson aloituspäivänä tai sitä aikaisemmin, ja arvo 0 tarkoittaa, että positiivinen arvo oli todennettu vasta lääkejakson aloituspäivän jälkeen tai ei ollenkaan. Elinaika-aineistojen rakenne on kuvattu taulukossa 4. Koska vastetapahtuman valinta vaikuttaa siihen, milloin yksilöt poistuvat seurannasta (yksilö poistui lopullisesti seurannasta kohdattuaan vastetapahtuman), oli kahdessa elinaika-aineistossa eri määrä havaintorivejä eli lääkejaksoja, ja tietysti myös eri määrä lääkejaksoja, jotka päättyivät vastetapahtumaan.

4.8.1 Elinaika-aineisto, kallonsisäinen verenvuoto

Elinaika-aineistossa, jossa vastemuuttujana oli kallonsisäinen verenvuoto, oli 397 954 havaintoriviä eli lääkejaksoa. Näistä lääkejaksoista 970 päättyi vastetapahtumaan eli kallonsisäiseen verenvuotoon. Elinaika-aineiston kokonaisseuranta-aika oli 128 154 984 päivää eli noin 351 110 vuotta.

Yksilön lääkejaksojen lukumäärän jakauma oli huomattavasti oikealle vino (taulukko 5). Suurimmalla osalla yksilöistä oli aineistossa vain yksi tai kaksi lääkejaksoa, mutta pienellä osalla yksilöistä lääkejaksojen määrä oli huomattava. Taulukossa 5 on esitetty myös kvantiilit lääkejaksojen kestoista päivinä. Myös tämä jakauma oli huomattavasti oikealle vino, sillä lääkejakson kesto oli suurimmassa osassa tapauksia enintään 120 päivää, mutta pienessä osassa tapauksia lääkejakso kesti useita vuosia, äärimmäisessä tapauksessa noin 26 vuotta. Noin 52% lääkejaksoista kesto oli 120 päivää. On siis ollut hyvin yleistä, että aloitettuaan lääkejakson yksilö on ostanut lääkettä uudestaan vasta yli 120 päivän kuluttua, vaikka lääkettä myydään kerrallaan vain kolmeksi kuukaudeksi. Elinaika-aineistossa oli mukana 8 polygeenista riskisummamuuttujaa: Zhoun ja kumppanien [39] sekä Jiangin ja kumppanien [19] GWAS-tunnusluvuista (vastetapahtumana kallonsisäinen verenvuoto) lasketut riskisummat sekä näiden kategorisoidut versiot laskettuna sekä SBayesRC- että PRS-CS-menetelmällä. Näiden lisäksi mukana olivat myös muut edellä määritetyt kovariaattimuuttujat.

	Min	10%	25%	50%	75%	90%	Max
Lääkejaksojen määrä yksilöillä	1	1	1	2	5	14	59
Lääkejaksojen kesto (päivinä)	1	120	120	120	295	744	9501

Taulukko 5: Lääkejaksoihin liittyviä kvantiileja, kun vastetapahtumana on kallonsisäinen verenvuoto.

4.8.2 Elinaika-aineisto, suolistonsisäinen verenvuoto

Elinaika-aineistossa, jossa vastemuuttujana oli suolistonsisäinen verenvuoto, oli 392 425 havaintoriviä eli lääkejaksoa. Näistä lääkejaksoista 2074 päättyi vastetapahtumaan eli suolistonsisäiseen verenvuotoon. Elinaika-aineiston kokonaisseuranta-aika oli 126 216 585 päivää eli noin 345 799 vuotta.

Yksilöiden lääkejaksojen lukumäärän ja lääkejaksojen keston jakaumat olivat lähes identtiset kallonsisäisen verenvuoden kanssa (taulukko 6). Tämä on ymmärrettävää, sillä niillä yksilöillä, joilla viimeinen lääkejakso ei pääty verenvuototapahtumaan, on kummassakin aineistossa identtiset havaintorivit, ja yli 96% tilastoyksiköistä oli tällaisia. Kuten kallonsisäisen verenvuodon tapauksessa, myös nyt elinaika-aineistossa oli mukana 8 polygeenista riskisummamuuttujaa (Zhoun ja kumppanien [39] sekä Jiangin ja kumppanien [19] GWAS-tunnusluvuista lasketut riskisummat vastetapahtumana suolistonsisäinen verenvuoto sekä näiden kategorisoidut versiot laskettuna sekä SBayesRC- että PRS-CS-menetelmällä) sekä muut edellä määritetyt kovariaattimuuttujat.

	Min	10%	25%	50%	75%	90%	Max
Lääkejaksojen määrä yksilöillä	1	1	1	2	5	14	59
Lääkejaksojen kesto (päivinä)	1	120	120	120	295	743	9501

Taulukko 6: Lääkejaksoihin liittyviä kvantiileja, kun vastetapahtumana on suolistonsisäinen verenvuoto.

4.9 Elinaika-analyysin menetelmät

Jokaisen vaihtoehtoisen polygeenisten riskisumman kohdalla riskisumman yhteyttä uhkaan kohdata kallonsisäinen tai suolistonsisäinen verenvuoto estimoitiin Coxin regressiomallilla, jossa polygeeninen riskisumma sekä muut aikaisemmin kuvatut kovariaatit selittivät aikaa kohdata kyseinen verenvuototapahtuma. Coxin regressiomallissa käytettiin myös frailty-parametria eli yksilöille estimoitiin yksilökohtaiset satunnaisvaikutukset. Jos polygeeniselle riskisummalle löytyi tilastollisesti merkitsevä yhteys verenvuototapahtumaan ($p < 0.1$), selvitettiin seuraavaksi, nostaako polygeenisen riskisumman käyttö mallin ennustetarkkuutta verrattuna malliin, jossa polygeeninen riskisumma ei ollut mukana.

Polygeenisen riskisumman prediktiokyvyn arvioimiseksi käytettiin concordance-indeksiä eli C-indeksiä. C-indeksi on yleisesti käytetty mitta ennustemallien arvioinnille elinaika-analyysissa [22]. Sen esitteli alunperin Harrell ja kumppanit vuonna

1982 [16]. C-indeksi voidaan määritellä todennäköisyytenä

$$C = P(M_j > M_i | T_j < T_i), \tag{16}$$

jossa M_j ja M_i ovat mallin satunnaisesti valituille yksilöille j ja i ennustamat riskit kohdata vastetapahtuma ja T_j ja Ti ovat näiden yksilöiden havaitut tapahtuma-ajat. Kyseessä on siis todennäköisyys, että kahdesta satunnaisesti valitusta yksilöstä sillä, joka kohtaa vastetapahtuman nopeammin, on myös suurempi mallin ennustama riski kohdata vastetapahtuma. Jos malli on ennustekyvyltään täydellinen, on mallin ennustama riski aina suurempi sillä yksilöllä, joka kohtaa vastetapahtuman ensimmäisenä, ja C-indeksi saa arvon 1. C-indeksin arvo 0.5 taas vastaa tilannetta, jossa mallin ennustama riski on täysin riippumaton siitä kumpi yksilöistä kohtaa vastetapahtuman ensimmäisenä. Tällöin mallin ennustekyky on samalla tasolla kuin kolikonheitto. Lähtökohtainen tapa estimoida C-indeksin arvo on Harrelin estimaattori, joka voidaan kirjoittaa

$$\widehat{C} = \frac{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} \Big[I(T_i < T_j) + (1 - \Delta_j) I(T_i = T_j) \Big] \Big[I(M_i > M_j) + \frac{1}{2} I(M_i = M_j) \Big]}{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} \Big[I(T_i < T_j) + (1 - \Delta_j) I(T_i = T_j) \Big]}, \tag{17}$$

jossa N on aineiston koko, Δ_i on indikaattorimuuttuja, joka kertoo päättyykö rivi i vastetapahtumaan, T_i on aika jossa vastetapahtuma tai sensurointi kohdattiin rivillä i ja M_i on mallin estimoima riski kohdata vastetapahtuma riville i. Käytännössä tämä tarkoittaa, että ensin elinaika-aineiston riveistä muodostetaan kaikki mahdolliset parit, joista tiedetään kummassa rivissä vastetapahtuma kohdattiin ensimmäisenä. Näiden parien määrä vastaa lausekkeen (17) nimittäjää. Lausekkeen osoittajassa taas on 1) nimittäjän pareista niiden parien määrä, joissa sillä rivillä jolla vastetapahtuma kohdataan aikaisemmin mallin estimoima riski on suurempi, toisin sanoen niiden parien määrä, joissa mallin ennuste piti paikkansa summattuna 2) nimittäjän pareista niiden määrällä, joissa mallin ennustamat riskit olivat yhtä suuret painotettuna arvolla 0.5. Yksinkertaisessa tilanteessa, jossa ennustetut riskit eivät koskaan ole yhtä suuret, Harrellin estimaattori on yksinkertaisesti kaikista aineiston vertailukelpoisista pareista niiden parien osuus, joissa mallin ennuste piti paikkansa (rivillä jolla ennustettu riski oli suurempi vastetapahtuma kohdattiin aikaisemmin). Tässä tutkielmassa käytetty C-indeksin estimaattori on Harrelin estimaattori.

On tunnettu tosiasia, että mallin ennustekyvystä saadaan ylioptimistinen kuva, jos sitä testataan samaan aineistoon kuin mistä malli on estimoitu. Realistisempi kuva saadaan käyttämällä jotain validointitekniikkaa, kuten ristiinvalidointia. Tässä tutkielmassa C-indeksien eroa mallin, joka sisältää polygeenisen riskisumman, ja mallin, joka ei tätä sisällä välillä testataan käyttämällä seuraavanlaista bootstrap-validointia. Elinaika-aineistosta poimitaan bootstrap-otos, joka on ositettu niin, että bootstrap-otokseen tulee aina sama määrä vastetapahtumia kuin alkuperäisessä elinaika-aineistossa. Keskimäärin noin 37% riveistä ei tule valituksi bootstrap-otokseen, tästä osasta aineistoa käytetään nimitystä OOB-otos (*out-of-bag*) [3]. Molemmat Coxin regressiomallit (malli joka sisältää polygeenisen riskisumman ja muu-

ten sama malli mutta ilman riskisummaa) estimoitiin bootstrap-otoksesta, ja mallien C-indeksit laskettiin OOB-otoksesta. Lisäksi laskettiin C-indeksien erotus. Tätä toistettiin 5000 iteraatiota, ja mallien C-indeksien erotuksen lopullinen estimaatti on näin saatujen bootstrap-otoksiin perustuvien C-indeksien erotusten mediaani. Lisäksi C-indeksien erotusten 5 %:n ja 95 %:n kvantiilit muodostavat 90 %:n uusioluottamusvälin (*bootstrap confidence interval*), joka kuvaa bootstrap-estimaattiin liittyvää epävarmuutta. Jos arvo 0 kuuluu C-indeksien erotuksen estimaatin uusioluottamusväliin, viestii tämä siitä, että ennustetarkkuudet eivät eroa mallin, jossa polygeeninen riskisumma on selittävänä muuttujana, ja muuten saman mallin mutta ilman riskisummaa välillä.

Analyysivaiheessa huomattiin, että yksilökohtaisen satunnaisvaikutuksen käyttö tehosti huomattavasti Coxin regressiomallin prediktiotarkkuutta, kun C-indeksi laskettiin samasta aineistoista kuin mistä malli estimoitiin. Esimerkiksi kun kallonsisäinen verenvuoto -aineistoon sovitettiin regressiomalli, jossa riskisummamuuttuja oli Zhengin ja kumppanien GWAS-tunnuslukuihin perustuva SBayesRC-menetelmällä laskettu jatkuva PRS-muuttuja ja malli sisälsi yksilökohtaisen satunnaisvaikutuksen, oli samasta aineistosta laskettu C-indeksi 0.991. Kun mallista poistettiin yksilökohtainen satunnaisvaikutus, oli samasta aineistosta laskettu C-indeksi 0.66. Koska kliininen päämäärä on arvioida riskiä kohdata verenvuototapahtuma uusilla yksilöillä, joihin liittyvät satunnaisvaikutukset eivät ole tiedossa, päädyttiin tutkielmassa käyttämään kaikkia C-indeksejä laskettaessa Coxin regressiomalleja ilman frailty-parametria eli yksilökohtaista satunnaisvaikutusta. Jokaisen riskisummamuuttujan kohdalla estimoitiin siis ensin riskisummamuuttujan ja muiden kovariaattien ja vastemuuttujan väliset yhteydet käyttäen mallia, joka sisälsi yksilökohtaiset satunnaisvaikutukset, ja jos jatkettiin ennustetarkkuuden arviointiin C-indekseillä, siirryttiin käyttämään regressiomallia ilman yksilökohtaista satunnaisvaikutusta.

5 Elinaika-analyysi

5.1 Kallonsisäinen verenvuoto, Zhou ja kumppanit GWAS

5.1.1 SBayesRC, jatkuva PRS

Ensimmäinen käytetty PRS-muuttuja oli Zhoun ja kumppanien kallonsisäisen verenvuodon GWAS-tunnuslukuihin perustuva jatkuva PRS, jonka laskemiseen käytettiin SBayesRC-menetelmää. Coxin regressiomallin tulokset on esitetty taulukossa 7.

Polygeenisen riskisumman uhkasuhteen estimoitu arvo oli 1.056, eli mallin mukaan suurempi polygeeninen riskisumma on yhteydessä suurempaan uhkaan kohdata kallonsisäinen verenvuoto. PRS-muuttujan uhkasuhteen 90% luottamusväli sisälsi kuitenkin arvon 1, eli yhteys ei ollut tilastollisesti merkitsevä asetetulla kriteerillä p-arvo < 0.1. Muista kovariaateista kaikista suurin yhteys kallonsisäiseen verenvuotoon oli yli 75 vuoden iällä, jolla uhkasuhde oli noin 3 (90% luottamusväli, LV, 2.5–3.5). Tulkinta on, että verrattuna alle 65-vuotiaisiin, yli 75 vuotiailla on noin kolminkertainen uhka kohdata kallonsisäinen verenvuoto. Muut tilastollisesti merkitsevät kovariaatit olivat dementia (US 2.1, 90% LV 1.7–2.6), 65-75 vuoden ikä (US 1.7, 90% LV 1.5–2.1), aivohalvaus (US 1.7, 90% LV 1.5–2.0), alkoholiongelma (US 1.5, 90%

Muuttuja	Log-US	US	US 90% LV	P-arvo
PRS (Zhou IC SBayesRC)	0.055	1.056	(0.999, 1.117)	0.107
Ikä 65–75	0.550	1.733	(1.459, 2.059)	<0.001
Ikä yli 75	1.094	2.986	(2.517, 3.543)	<0.001
Mies	0.173	1.189	(1.054, 1.342)	0.019
Sukupuoli tuntematon	0.250	1.283	(0.908, 1.815)	0.236
Verisuonitauti	0.054	1.056	(0.929, 1.199)	0.484
Kohonnut verenpaine	-0.036	0.965	(0.759, 1.228)	0.808
Dyslipidemia	0.037	1.037	(0.903, 1.192)	0.663
Sydämen vajaatoiminta	0.138	1.148	(1.007, 1.308)	0.083
Diabetes	0.046	1.047	(0.924, 1.187)	0.543
Aivohalvaus	0.536	1.710	(1.485, 1.968)	<0.001
Alkoholiongelma	0.433	1.542	(1.144, 2.079)	0.017
Munuaisen vajaatoiminta tai dialyysi	0.190	1.209	(0.917, 1.595)	0.259
Maksan vajaatoiminta tai kirroosi	0.336	1.400	(0.721, 2.718)	0.404
Dementia	0.752	2.121	(1.720, 2.616)	<0.001
Psykiatrinen häiriö	0.126	1.134	(0.965, 1.333)	0.199

Taulukko 7: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, SBayesRC-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

LV 1.1–2.1), miessukupuoli (US 1.2, 90% LV 1.1–1.3) ja sydämen vajaatoiminta (US 1.1, 90% LV 1.0–1.3). Näiden kovariaattien uhkasuhteet muuttuivat hyvin vähän myöhemmin käsiteltävissä malleissa, joissa PRS-muuttuja on korvattu toisella, eikä niitä kommentoida enää tämän jälkeen. Koska PRS-muuttuja ei ollut tilastollisesti merkitsevä, ei ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla suoritettu. Vastaavan mallin ilman yksilökohtaista satunnaisvaikutusta samasta aineistosta laskettu C-indeksi oli 0.659.

5.1.2 SBayesRC, kategorinen PRS

Seuraavaksi PRS-muuttuja korvattiin saman muuttujan kategorisoidulla versiolla, Coxin regressiomallin tulokset on esitetty taulukossa 8. Tuloksista huomattiin kaksi asiaa, ensinnäkin korkeimmalla PRS-luokalla uhkasuhteen luottamusväli ei sisältänyt arvoa 1, eli korkein PRS-luokka oli tilastollisesti merkitsevässä yhteydessä kallonsisäiseen verenvuotoon. Luokan estimoitu uhkasuhde oli 1.4, mallin mukaan siis yksilöillä, joiden PRS kuuluu ylimpään 2.5 prosenttiin, on 1.4 kertainen uhka kohdata kallonsisäinen verenvuoto kuin yksilöillä, joiden PRS on keskimääräistä tasoa (kvantiiliväli 0.2 – 0.8). Toiseksi huomattiin, että kokonaisuutena kategorisen PRS-muuttujan yhteys riskiin kohdata verenvuototapahtuma ei ollut täysin johdonmukainen, sillä referenssitasoon verrattuna PRS-muuttujan korkeammassa kvantiililuokassa 0.8–0.975 oli pienempi uhka kohdata vastetapahtuma, ja luokassa 0.8–0.975 uhka kohdata vastetapahtuma oli pienempi kuin kaikista matalimmassa PRS-luokassa. Lisäksi kaikilla muilla kuin korkeimmalla PRS-tasolla uhkasuhteen 90% luottamusväli sisälsi arvon 1, eli yhteydet eivät olleet tilastollisesti merkitseviä. Kun C-indeksi las-

kettiin samalla mallilla ilman frailty-parametria samalle elinaika-aineistolle saatiin C-indeksin arvoksi 0.662.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (0 - 0.025)	-0.027	0.973	(0.666, 1.421)	0.905
PRS (0.025 - 0.2)	-0.044	0.957	(0.820, 1.116)	0.637
PRS (0.8 - 0.975)	-0.104	0.901	(0.771, 1.053)	0.273
PRS (0.975 - 1)	0.366	1.442	(1.060, 1.960)	0.050
Ikä 65–75	0.551	1.735	(1.461, 2.061)	<0.001
Ikä yli 75	1.093	2.984	(2.514, 3.540)	<0.001
Mies	0.173	1.189	(1.053, 1.342)	0.019
Sukupuoli tuntematon	0.249	1.283	(0.907, 1.815)	0.238
Verisuonitauti	0.055	1.057	(0.930, 1.201)	0.475
Kohonnut verenpaine	-0.035	0.966	(0.759, 1.229)	0.812
Dyslipidemia	0.036	1.037	(0.902, 1.191)	0.668
Sydämen vajaatoiminta	0.138	1.148	(1.007, 1.308)	0.083
Diabetes	0.046	1.047	(0.924, 1.187)	0.544
Aivohalvaus	0.538	1.713	(1.488, 1.972)	<0.001
Alkoholiongelma	0.429	1.536	(1.139, 2.071)	0.018
Munuaisen vajaatoiminta tai dialyysi	0.192	1.212	(0.919, 1.599)	0.254
Maksan vajaatoiminta tai kirroosi	0.344	1.411	(0.726, 2.742)	0.394
Dementia	0.755	2.127	(1.724, 2.625)	<0.001
Psykiatrinen häiriö	0.127	1.136	(0.966, 1.335)	0.195

Taulukko 8: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, SBayesRC-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

Koska kategorisen PRS-muuttujan ylin luokka oli tilastollisesti merkitsevässä yhteydessä uhkaan kohdata kallonsisäinen verenvuoto, suoritettiin ennustetarkkuuden jatkotarkastelu bootstrap-validoinnilla. Bootstrap-validoinnilla C-indeksin estimaatti oli 0.651 ja 90 %:n uusioluottamusväli oli [0.626, 0.675]. C-indeksien erotuksen (malli jossa PRS-muuttuja oli mukana ja muuten sama malli mutta ilman PRS-muuttujaa) bootstrap-estimaatti taas oli -0.001 ja 90 %:n uusioluottamusväli [-0.007, 0.003]. Koska arvo 0 kuuluu uusioluottamusväliin, ei saatu näyttöä sille, että ennustetarkkuus olisi parempi mallissa, jossa PRS-muuttuja on mukana.

5.1.3 PRS-CS, jatkuva PRS

Seuraavaksi kokeiltiin jatkuvaa PRS-muuttujaa, jonka laskemisessa käytettiin PRS-CS-menetelmää. Coxin regressiomallin tulokset on esitetty taulukossa 9. PRS-muuttujan uhkasuhde oli 1.03, eli mallin mukaan suurempi PRS on yhteydessä suurempaan uhkaan kohdata kallonsisäinen verenvuoto. Uhkasuhteen 90% luottamusväli sisälsi kuintenkin arvon 1, eli yhteys ei ollut tilastollisesti merkitsevä asetetulla kriteerillä. Saman mallin mutta ilman yksilökohtaista satunnaisvaikutusta C-indeksi laskettuna samasta aineistosta oli 0.659. Koska PRS-muuttuja ei ollut tilastollisesti

merkitsevä, ei ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla suoritettu.

Muuttuja	Log-US	US	US 90% LV	p-value
PRS (Zhou IC PRS-CS)	0.031	1.031	(0.975, 1.091)	0.37
Ikä 65–75	0.55	1.734	(1.460, 2.059)	<0.001
Ikä yli 75	1.094	2.986	(2.517, 3.543)	<0.001
Mies	0.173	1.189	(1.053, 1.342)	0.019
Sukupuoli tuntematon	0.25	1.284	(0.907, 1.816)	0.236
Verisuonitauti	0.055	1.056	(0.929, 1.200)	0.482
Kohonnut verenpaine	-0.036	0.965	(0.758, 1.228)	0.808
Dyslipidemia	0.036	1.037	(0.903, 1.191)	0.667
Sydämen vajaatoiminta	0.138	1.148	(1.007, 1.308)	0.082
Diabetes	0.046	1.047	(0.924, 1.187)	0.544
Aivohalvaus	0.538	1.712	(1.487, 1.972)	<0.001
Alkoholiongelma	0.434	1.543	(1.144, 2.080)	0.017
Munuaisen vajaatoiminta tai dialyysi	0.191	1.210	(0.917, 1.597)	0.257
Maksan vajaatoiminta tai kirroosi	0.335	1.398	(0.719, 2.716)	0.407
Dementia	0.753	2.123	(1.721, 2.619)	<0.001
Psykiatrinen häiriö	0.126	1.135	(0.965, 1.334)	0.198

Taulukko 9: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, PRS-CS-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

5.1.4 PRS-CS, kategorinen PRS

Viimeisenä estimoinnissa käytettiin PRS-CS-menetelmällä lasketun PRS-muuttujan kategorista versiota, mallin tulokset on esitetty taulukossa 10. Kuten aikaisemmin SBayesRC-menetelmällä lasketun PRS-muuttujan tapauksessa, myös nyt korkeimmalla PRS-luokalla oli suurin uhkasuhde (US = 1.3). Tällä kertaa uhkasuhteen luottamusväli kuitenkin sisälsi arvon 1, eli yhteys ei ollut tilastollisesti merkitsevä. Myöskään tässä mallissa kategorinen PRS-muuttuja ei käyttäydy täysin johdonmukaisesti, alimmalla PRS-tasolla log-uhkasuhde on positiivinen, mikä tarkoittaa sitä, että mallin mukaan niillä, joilla on erittäin alhainen PRS on suurempi uhka kohdata verenvuototapahtuma kuin niillä, joilla PRS on keskitasoa. Muuten samalla mallilla mutta ilman yksilökohtaista satunnaisvaikutusta samasta aineistosta lasketun C-indeksin arvo oli 0.661. Koska kategorisen PRS-muuttujan tasot eivät olleet tilastollisesti merkitseviä, ei ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla suoritettu.

5.2 Kallonsisäinen verenvuoto, Jiang ja kumppanit GWAS

PRS-muuttujat, joiden laskemisessa käytettiin Jiangin ja kumppanien GWAS-tunnuslukuja, eivät toimineet yhtä hyvin kuin PRS-muuttujat, joiden laskemisessa käytettiin Zhoun ja kumppanien GWAS-tunnuslukuja. Minkään PRS-muuttujan ta-

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (0 - 0.025)	0.088	1.092	(0.763, 1.565)	0.686
PRS (0.025 - 0.2)	-0.083	0.921	(0.787, 1.077)	0.386
PRS (0.8 - 0.975)	-0.004	0.996	(0.856, 1.160)	0.969
PRS (0.975 - 1)	0.264	1.302	(0.940, 1.804)	0.182
Ikä 65–75	0.550	1.733	(1.459, 2.058)	<0.001
Ikä yli 75	1.093	2.983	(2.514, 3.539)	<0.001
Mies	0.173	1.188	(1.053, 1.341)	0.019
Sukupuoli tuntematon	0.249	1.283	(0.908, 1.814)	0.236
Verisuonitauti	0.054	1.056	(0.929, 1.199)	0.483
Kohonnut verenpaine	-0.035	0.965	(0.759, 1.228)	0.809
Dyslipidemia	0.036	1.037	(0.903, 1.191)	0.669
Sydämen vajaatoiminta	0.138	1.148	(1.007, 1.308)	0.082
Diabetes	0.046	1.047	(0.924, 1.186)	0.546
Aivohalvaus	0.538	1.712	(1.487, 1.971)	<0.001
Alkoholiongelma	0.432	1.540	(1.143, 2.076)	0.017
Munuaisen vajaatoiminta tai dialyysi	0.191	1.210	(0.918, 1.596)	0.257
Maksan vajaatoiminta tai kirroosi	0.339	1.403	(0.722, 2.725)	0.401
Dementia	0.752	2.122	(1.720, 2.617)	<0.001
Psykiatrinen häiriö	0.126	1.134	(0.965, 1.333)	0.199

Taulukko 10: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, PRS-CS-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

pauksessa ei löytynyt tilastollisesti merkitsevää yhteyttä vastemuuttujaan. Tuloksia ei käsitellä tässä enempää, vaan elinaika-analyysin yksityiskohdat löytyvät tutkielman liitteestä A.

5.3 Suolistonsisäinen verenvuoto, Zheng ja kumppanit GWAS

5.3.1 SBayesRC, jatkuva PRS

Taulukossa 11 on esitetty Coxin regressiomallin tulokset, kun jatkuva PRS-muuttuja on laskettu käyttäen SBayesRC-menetelmää. Jatkuvan PRS-muuttujan uhkasuhteen estimoitu arvo oli 1.04, eli mallin mukaan suurempi PRS on yhteydessä suurempaan uhkaan kohdata suolistonsisäinen verenvuoto. Kuten aiemmin kallonsisäisen verenvuodon tapauksessa, myös nyt PRS-muuttujan uhkasuhteen luottamusväli sisälsi arvon 1 eli yhteys ei ollut tilastollisesti merkitsevä. Kun muuten saman mallin mutta ilman yksilökohtaista satunnaisvaikutusta C-indeksi laskettiin samasta aineistosta, saatiin C-indeksin arvoksi 0.658. Koska PRS-muuttuja ei ollut tilastollisesti merkitsevä, ei ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla suoritettu.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (Zhou GI SBayesRC)	0.040	1.040	(0.999, 1.084)	0.110
Ikä 65–75	0.244	1.276	(1.140, 1.429)	<0.001
Ikä yli 75	0.811	2.250	(2.010, 2.519)	<0.001
Mies	0.231	1.260	(1.154, 1.375)	<0.001
Sukupuoli tuntematon	0.066	1.068	(0.816, 1.399)	0.687
Verisuonitauti	0.237	1.267	(1.156, 1.388)	<0.001
Kohonnut verenpaine	0.323	1.382	(1.138, 1.678)	0.006
Dyslipidemia	-0.061	0.940	(0.852, 1.038)	0.306
Sydämen vajaatoiminta	0.499	1.647	(1.508, 1.798)	<0.001
Diabetes	0.231	1.260	(1.155, 1.375)	<0.001
Aivohalvaus	0.134	1.143	(1.023, 1.278)	0.048
Alkoholiongelma	0.794	2.212	(1.814, 2.697)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.667	1.949	(1.647, 2.306)	<0.001
Maksan vajaatoiminta tai kirroosi	0.638	1.892	(1.266, 2.828)	0.009
Dementia	0.196	1.216	(1.014, 1.459)	0.077
Psykiatrinen häiriö	0.151	1.163	(1.036, 1.307)	0.032

Taulukko 11: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, SBayesRC-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

Tilastollisesti merkitseviä kovariaatteja oli suurempi määrä kuin kallonsisäisen verenvuodon tapauksessa, kaikki kovariaatit paitsi tuntematon sukupuoli ja dyslipidemia olivat tilastollisesti merkitseviä kriteerillä p-arvo < 0.1. Kaikista voimakkain yhteys suolistonsisäiseen verenvuotoon (suurin log-uhkasuhteen itseisarvo) oli yli 75:n vuoden iällä, jolla estimoitu uhkasuhde oli 2.3 (90% LV 2.0–2.5). Lähes yhtä suuri uhkasuhde oli alkoholiongelmalla (US 2.2, 90% LV 1.8–2.7). Mallin mukaan siis uhka kohdata suolistonsisäinen verenvuoto on yli kaksinkertainen yli 75:n vuoden ikäisillä verrattuna alle 65-vuotiaisiin sekä yksilöillä, joilla oli aikaisempi todennettu alkoholiongelma verrattuna heihin, joilla tätä ei ollut todennettu. Seuraavaksi suurimmat yhteydet olivat munuaisen vajaatoiminnalla (US 1.9, 90% LV 1.6–2.3), maksan vajaatoiminnalla (US 1.9, 90% LV 1.3–2.8) ja sydämen vajaatoiminnalla

(US 1.6, 90% LV 1.5–1.8). Lopuilla tilastollisesti merkitsevillä kovariaateilla oli hieman pienemmät yhteydet (uhkasuhteet vaihtelivat välillä 1.1–1.4). Kovariaattien uhkasuhteet eivät muuttuneet suuresti malleissa, joissa PRS-muuttuja on korvattu toisella, eikä niitä kommentoida enää tämän jälkeen.

5.3.2 SBayesRC, kategorinen PRS

Seuraavaksi jatkuva PRS-muuttuja korvattiin kategorisella versiolla, mallin tulokset on esitetty taulukossa 12. Ainoa PRS-muuttujan luokka, jossa uhkasuhteen luottamusväli ei sisältänyt arvoa 1, oli alin taso eli kvantiiliväli 0–0.025, jonka uhkasuhde oli 0.7 (90% LV 0.5–0.9). Mallin mukaan siis yksilöillä, joiden PRS kuuluu alimpaan 2.5 prosenttiin uhka kohdata verenvuototapahtuma on 0.7 kertainen verrattu yksilöihin, joiden PRS on keskitasoa (kvantiiliväli 0.2–0.8). Kuten kallonsisäisen verenvuodon tapauksessa, ei nytkään kategorinen PRS-muuttuja käyttäytynyt täysin johdonmukaisesti. Kahdella korkeimmalla tasolla log-uhkasuhteen estimaatti oli negatiivinen, eli mallin mukaan uhka kohdata suolistonsisäinen verenvuoto on pienempää suurilla PRS:n arvoilla verrattuna keskitasoon. Kun muuten saman mallin mutta ilman yksilökohtaista satunnaisvaikutusta C-indeksi laskettiin samasta aineistosta saatiin sen arvoksi 0.658.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (0 - 0.025)	-0.405	0.667	(0.493, 0.902)	0.027
PRS (0.025 - 0.2)	-0.103	0.902	(0.807, 1.009)	0.130
PRS (0.8 - 0.975)	-0.050	0.951	(0.853, 1.060)	0.446
PRS (0.975 - 1)	-0.152	0.859	(0.659, 1.120)	0.346
Ikä 65–75	0.242	1.274	(1.138, 1.426)	<0.001
Ikä yli 75	0.807	2.241	(2.003, 2.508)	<0.001
Mies	0.230	1.258	(1.153, 1.373)	<0.001
Sukupuoli tuntematon	0.064	1.066	(0.815, 1.394)	0.696
Verisuonitauti	0.236	1.266	(1.156, 1.386)	<0.001
Kohonnut verenpaine	0.323	1.381	(1.138, 1.676)	0.006
Dyslipidemia	-0.060	0.942	(0.854, 1.039)	0.315
Sydämen vajaatoiminta	0.495	1.641	(1.503, 1.792)	<0.001
Diabetes	0.229	1.258	(1.153, 1.371)	<0.001
Aivohalvaus	0.134	1.143	(1.023, 1.277)	0.047
Alkoholiongelma	0.789	2.201	(1.807, 2.681)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.661	1.936	(1.638, 2.289)	<0.001
Maksan vajaatoiminta tai kirroosi	0.639	1.895	(1.271, 2.825)	0.008
Dementia	0.191	1.210	(1.010, 1.451)	0.083
Psykiatrinen häiriö	0.151	1.163	(1.035, 1.305)	0.032

Taulukko 12: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, SBayesRC-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

Koska yksi kategorisen PRS-muuttujan tasoista oli tilastollisesti merkitsevä, mallin ennustetarkkuutta testattiin bootstrap-validoinnilla. Mallin C-indeksin estimaa-

tiksi saatiin 0.6522 ja 90 %:n uusioluottamusväliksi [0.635, 0.669]. Bootstrap-validoinnilla saatu estimaatti mallin, jossa PRS-muuttuja oli mukana, ja muuten saman mallin mutta ilman PRS-muuttujaa C-indeksien erotukselle oli -0.001 ja 90 %:n uusioluottamusväli [-0.005, 0.002]. Koska uusioluottamusväli sisälsi arvon 0, ei ollut näyttöä sille, että PRS-muuttuja lisäisi mallin ennustetarkkuutta.

5.3.3 PRS-CS, jatkuva PRS

Seuraavaksi kokeiltiin PRS-CS-menetelmällä laskettua jatkuvaa PRS-muuttujaa, Coxin regressiomallin tulokset on esitetty taulukossa 13. PRS-muuttujan uhkasuhde oli 1.046, eli mallin mukaan suurempi PRS-arvo on yhteydessä suurempaan uhkaan kohdata suolistonsisäinen verenvuototapahtuma. Koska uhkasuhteen luottamusväli ei sisältänyt arvoa 1, oli PRS-muuttuja tilastollisesti merkitsevä asetetulla kriteerillä. Koska PRS-muuttuja on jatkuva, ei sen uhkasuhteen tulkinta ole yhtä suoraviivaista kuin kategorisen muuttujan tapauksessa. Uhkasuhteesta saadaan karkeasti vertailukelpoinen kategorisen tapauksen kanssa seuraavalla operaatiolla. Kun oletetaan että PRS-muuttuja on normaalijakautunut, on standardoidun PRS-muuttujan vaihteluväli noin 4. Kertomalla log-uhkausuhde 0.045 arvolla 4, saadaan $0.045 \times 4 = 0.18$, joka on karkeasti tulkittavissa log-uhkasuhteiden erona yksilön, jolla on erittäin alhainen PRS ja yksilön, jolla on erittäin korkea PRS välillä. Tällöin uhkasuhteiden ero samojen yksilöiden välillä on $\exp(0.18) = 1.197$, eli uhka kohdata suolistonsisäinen verenvuoto on noin 1.2 kertaa korkeampi yksilöillä, joilla on erittäin korkea PRS verrattuna yksilöihin, joilla on erittäin matala PRS. Kun muuten sama malli mutta ilman yksilökohtaista satunnaisvaikutusta sovitettiin elinaika-aineistoon ja C-indeksi laskettiin samasta aineistosta, saatiin sen arvoksi 0.658.

Koska PRS-muuttuja oli tilastollisesti merkitsevä, suoritettiin mallin ennustetarkkuuden jatkotarkastelu bootstrap-validoinnilla. Mallin bootstrap-validoiduksi C-indeksin estimaatiksi saatiin 0.653 ja 90 %:n uusioluottamusväliksi [0.637, 0.670]. Bootstrap-validoinnilla saatu C-indeksien erotuksen estimaatti mallille jossa PRS-muuttuja oli mukana ja muuten samalle mallille mutta ilman PRS-muuttujaa oli 0.0002 ja 90 %:n uusioluottamusväli [-0.002, 0.001]. Koska uusioluottamusväli sisälsi arvon 0, ei ollut näyttöä sille, että PRS-muuttuja lisäisi mallin ennustetarkkuutta.

5.3.4 PRS-CS, kategorinen PRS

Viimeiseksi PRS-CS-menetelmällä laskettu PRS-muuttuja korvattiin sen kategorisella versiolla, ja tämän mallin tulokset on esitetty taulukossa 14. PRS-muuttujan luokista tilastollisesti merkitsevä kriteerillä p-arvo < 0.1 oli vain toiseksi alin taso 0.025–0.2, jossa uhkasuhteen luottamusväli ei sisältänyt arvoa 1. Tässä luokassa uhkasuhde oli 0.9, eli yksilöillä, joiden PRS kuului kvantiiliväliin 0.025–0.2 uhka kohdata suolistonsisäinen verenvuoto oli 0.9 kertainen verrattuna yksilöihin, joiden PRS kuului kvantiiliväliin 0.2–0.8. PRS-muuttujan luokat eivät tässäkään tapauksessa käyttäydy täysin johdonmukaisesti, korkeimmassa PRS-luokassa on negatiivinen log-uhkasuhde eli mallin mukaan uhka kohdata verenvuototapahtuma niillä, joilla on kaikista korkein PRS-taso, on pienempi kuin referenssitasolla. Kun muuten sama malli mutta ilman yksilökohtaista satunnaisvaikutusta sovitettiin elinaika-aineistoon ja tälle laskettiin C-indeksi samasta aineistosta, saatiin C-indeksin arvoksi 0.658.

Muuttuja	Log-US	US	US 90% LV	P-arvo
PRS (Zhou GI PRS-CS)	0.045	1.046	(1.005, 1.090)	0.066
Ikä 65–75	0.243	1.276	(1.140, 1.428)	<0.001
Ikä yli 75	0.811	2.249	(2.010, 2.518)	<0.001
Mies	0.230	1.259	(1.154, 1.374)	<0.001
Sukupuoli tuntematon	0.065	1.067	(0.815, 1.397)	0.693
Verisuonitauti	0.236	1.266	(1.156, 1.387)	<0.001
Kohonnut verenpaine	0.323	1.381	(1.137, 1.677)	0.006
Dyslipidemia	-0.061	0.941	(0.852, 1.038)	0.306
Sydämen vajaatoiminta	0.498	1.645	(1.507, 1.797)	<0.001
Diabetes	0.231	1.260	(1.155, 1.374)	<0.001
Aivohalvaus	0.133	1.143	(1.022, 1.277)	0.049
Alkoholiongelma	0.793	2.210	(1.812, 2.694)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.667	1.947	(1.646, 2.304)	<0.001
Maksan vajaatoiminta tai kirroosi	0.638	1.892	(1.266, 2.827)	0.009
Dementia	0.195	1.215	(1.013, 1.458)	0.078
Psykiatrinen häiriö	0.151	1.163	(1.036, 1.306)	0.032

Taulukko 13: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, PRS-CS-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

Koska kategorinen PRS-muuttuja sisälsi tilastollisesti merkitsevän luokan, suoritettiin mallin ennustetarkkuuden jatkotarkastelu bootstrap-validoinnilla. Mallin C-indeksin estimaatiksi saatiin 0.652 ja 90 %:n uusioluottamusväliksi [0.636, 0.670]. Mallin jossa PRS-muuttuja oli mukana ja muuten saman mallin mutta ilman PRS-muuttujaa C-indeksien erotuksen estimaatiksi saatiin -0.0004 ja 90 %:n uusioluottamusväliksi [-0.004, 0.002]. Koska uusioluottamusväli sisälsi arvon 0, ei ollut näyttöä sille, että kategorinen PRS-muuttuja parantaisi mallin ennustetarkkuutta.

5.4 Suolistonsisäinen verenvuoto, Jiang ja kumppanit GWAS

Myöskään suolistonsisäisen verenvuodon tapauksessa PRS-muuttujat, joiden laskemisessa käytettiin Jiangin ja kumppanien GWAS-tunnuslukuja, eivät toimineet yhtä hyvin kuin PRS-muuttujat, joiden laskemisessa käytettiin Zhoun ja kumppanien GWAS-tunnuslukuja, eikä minkään PRS-muuttujan tapauksessa löytynyt tilastollisesti merkitsevää yhteyttä vastemuuttujaan. Tuloksia ei käsitellä tässä enempää, vaan elinaika-analyysin yksityiskohdat löytyvät tutkielman liitteestä A.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (0 - 0.025)	-0.229	0.796	(0.598, 1.059)	0.189
PRS (0.025 - 0.2)	-0.116	0.890	(0.795, 0.997)	0.092
PRS (0.8 - 0.975)	0.091	1.095	(0.985, 1.217)	0.158
PRS (0.975 - 1)	-0.113	0.893	(0.684, 1.167)	0.487
Ikä 65–75	0.244	1.276	(1.140, 1.428)	<0.001
Ikä yli 75	0.810	2.248	(2.009, 2.516)	<0.001
Mies	0.230	1.258	(1.153, 1.373)	<0.001
Sukupuoli tuntematon	0.062	1.064	(0.813, 1.393)	0.703
Verisuonitauti	0.236	1.266	(1.156, 1.387)	<0.001
Kohonnut verenpaine	0.321	1.379	(1.136, 1.674)	0.006
Dyslipidemia	-0.061	0.941	(0.853, 1.038)	0.308
Sydämen vajaatoiminta	0.497	1.643	(1.505, 1.794)	<0.001
Diabetes	0.230	1.259	(1.154, 1.373)	<0.001
Aivohalvaus	0.134	1.143	(1.023, 1.277)	0.048
Alkoholiongelma	0.794	2.212	(1.815, 2.696)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.664	1.942	(1.642, 2.296)	<0.001
Maksan vajaatoiminta tai kirroosi	0.639	1.894	(1.269, 2.828)	0.009
Dementia	0.193	1.213	(1.011, 1.454)	0.080
Psykiatrinen häiriö	0.149	1.160	(1.033, 1.303)	0.035

Taulukko 14: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Zhou ja kumppanit GWAS, PRS-CS-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

5.5 Johtopäätökset

Sovelluksessa käytetetyt PRS-muuttujat pohjautuivat joko Zhoun ja kumppanien tai Jiangin ja kumppanien GWAS-tunnuslukuihin. Sekä kallonsisäisen että suolistonsisäisen verenvuodon tapauksessa ensimmäisenä mainitut toimivat paremmin, Jiangin ja kumppanien GWAS-tunnuslukuihin perustuvista jatkuvista PRS-muuttujista yksikään ei ollut tilastollisesti merkitsevässä yhteydessä vastemuuttujaan asetetulla kriteeritasolla, kuten ei myöskään yksikään näiden kategoristen versioiden luokista.

Joidenkin Zhoun ja kumppanien GWAS-tunnuslukuihin perustuvien PRS-muuttujien ja vastemuuttujan väliltä löytyi vaatimaton mutta asetetulla kriteeritasolla $p < 0.1$ merkitsevä yhteys. Tällaisia olivat kallonsisäisen verenvuodon tapauksessa SBayesRC-menetelmällä lasketun PRS-muuttujan kategorisen version korkein kvantiililuokka 0.975 – 1 (US 1.4, 90% LV 1.1–2.0), ja suolistonsisäisen verenvuodon tapauksessa SBayesRC-menetelmällä lasketun PRS-muuttujan kategorisen version alin kvantiililuokka 0–0.025 (US 0.7, 90% LV 0.5–0.9), PRS-CS-menetelmällä laskettu jatkuva PRS-muuttuja (US 1.05, 90% LV 1.0–1.1) sekä tämän kategorisen version toiseksi alin kvantiililuokka 0.025–0.2 (US 0.9, 90% LV 0.8–1.0). Kuitenkin jokaisessa tapauksessa kategoristen PRS-muuttujien kaikki luokat eivät käyttäytyneet johdonmukaisesti, vaan joissakin luokissa yhteys vastemuuttujaan oli toivottuun verrattuna päinvastainen.

Kaikissa tapauksissa, joissa jatkuva PRS-muuttuja tai jokin kategorisen PRS-muuttujan luokista oli tilastollisesti merkitsevässä yhteydessä vastemuuttujaan, suoritettiin ennustetarkkuuden jatkotarkastelu. Mallin, jossa PRS-muuttuja oli mukana, ennustetarkkuutta verrattiin muuten saman mallin, mutta ilman PRS-muuttujaa ennustetarkkuuteen. Tarkastelu tehtiin vertaamalla mallien C-indeksejä bootstrap-validaatiolla. Minkään PRS-muuttujan tapauksessa ei ollut näyttöä sille, että PRS-muuttujan käyttö mallissa parantaisi mallin ennustekykyä. Tutkielman sovellusosan johtopäätös on siis, että polygeenisista riskisummista ei ole hyötyä verenohennuslääkitystä käyttävien yksilöiden kallon- tai suolistonsisäisen verenvuodon ennustamisessa.

6 Pohdinta

6.1 SBayesRC-menetelmä

SBayesRC on uusi polygeeniseen prediktioon tarkoitettu menetelmä, ja siinä on kaksi keskeistä ominaisuutta, joiden Zheng ja kumppanit [38] esittävät artikkelissaan parantavan sen suorituskykyä verrattuna muihin vastaaviin menetelmiin. Ensimmäinen ominaisuus on matala-asteinen malli, johon estimointi perustuu. Sen sijaan, että snippien yhteisvaikutukset sovitettaisiin kaikkiin GWAS-tutkimuksesta saatuihin snippien marginaalivaikutuksiin, ne sovitetaan rajattuun joukkoon marginaalivaikutusten kaikista informatiivisimpia lineaarikombinaatioita. Matala-asteisen mallin esitetään lisäävän huomattavasti menetelmän laskentatehoa. Zhengin ja kumppanien tutkimuksessa matala-asteisen mallin LD-lohkokohtainen lineaarikombinaatioiden määrä q_k oli keskimäärin noin 20% alkuperäisestä snippien lohkokohtaisesta lukumäärästä m_k, kun alkuperäisestä LD-lohkon varianssista haluttiin säilyttää vä-

hintään 99.5% eli ρ-parametrin arvo oli 0.995. Polygeenisen prediktion perusidea on, että vastetapahtumaa selitetään käyttämällä tietoa koko perimän laajuudelta, tai ainakin käyttäen huomattavaa osaa siitä. Tämä tarkoittaa sitä, että yleisesti ottaen mitä suurempi määrä informatiivisia perimämuuttujia eli snippejä on käytössä, sitä parempi. Menetelmien laskentateho asettaa kuitenkin käytännön rajoitteita sille, kuinka monia miljoonia snippejä estimointiin voidaan sisällyttää. Siksi SBayesRC-menetelmän – tai tarkemmin matala-asteisen mallin – tarjoama lisäys laskentatehoon on huomattava ominaisuus. Lisäksi Zheng ja kumppanit esittävät matala-asteisen mallin myös lisäävän menetelmän robustisuutta. Robustisuus johtuu ensinnäkin siitä, että dimensionpienennys poistaa LD-lohkoista pieniä ominaisarvoja/ominaisvektoreita, jotka ovat alttiita LD:n suurelle otoskohtaiselle variaatiolle. Matala-asteisessa mallissa havaintojen residuaalit ovat myös toisistaan riippumattomia, mikä mahdollistaa residuaalivarianssin estimoinnin Gibbsin otannassa lisäten mallin robustisuutta.

SBayesRC-menetelmän toinen oleellinen ominaisuus on funktionaalisten annotaatioiden hyödyntäminen yhteisvaikutusten estimoinnissa. Tätä ominaisuutta painotetaan erityisesti Zhengin ja kumppanien menetelmää käsittelevässä artikkelissa [38]. Tiettyyn vastemuuttujaan on merkityksellisellä tavalla yhteydessä rajallinen määrä snippejä, ja jossain vaiheessa saavutetaan vääjäämättä tilanne, jossa estimoinnissa käytettävien snippien määrän lisääminen ei enää tuo estimointiin huomionarvoista hyötyä. Siksi on oleellista kysyä, voiko polygeenista prediktiota tehostaa snippien määrän lisäämisen sijaan käyttämällä estimoinnissa muutakin informaatiota snipeistä kuin marginaalivaikutukset ja korrelaatiot. Funktionaaliset annotaatiot vastaavat tähän kysymykseen. Funktionaaliset annotaatiomuuttujat voivat olla luonteeltaan yleisluontoisia – kuten sijaitseeko snippi lähimmän geenin koodaavalla alueella – mutta on myös mahdollista luoda vahvasti tiettyyn vastemuuttujaan liittyviä annotaatioita. SBayesRC käyttää näitä annotaatiomuuttujia yhteisvaikutusten estimoinnissa antamalla niiden vaikuttaa Bayes-mallissa yhteisvaikutusten priorijakaumiin. Zheng ja kumppanit esittävät tutkimuksessaan, että funktionaaliset annotaatiot auttavat löytämään voimakkaasti korreloituneiden snippien joukosta ne, jotka ovat oikeasti kausaalisessa yhteydessä vastemuuttujaan. Heidän analyyseissaan funktionaalisten annotaatioiden käyttö lisäsi mallien ennustetarkkuutta, ja lisäys ennustetarkkuuteen oli erityisen suurta käytettäessä suurta määrää snippejä (noin 7 miljoonaa) verrattuna pienempään määrään (noin miljoona).

Tutkielmassa saavutettiin varsin kattava ymmärrys SBayesRC-menetelmästä ja tilastollisesta mallista sen taustalla. Pääasiallisena lähteenä käytettiin Zhengin ja kumppanien menetelmää käsittelevän tutkimusartikkelin menetelmäliitettä [38]. Tämä menetelmäliite ei kuitenkaan ollut kaikilta osin kovinkaan yksityiskohtainen, ja se sisälsi sekä epäselviä kohtia että joitakin suoranaisia virheitä. Siksi osa tutkielman teoriaosasta pohjautuu olettamiin ja päättelyihin, joita ei löydy lähdemateriaalista. Joitakin tällaisia kohtia teoriaosassa jo mainittiin, esimerkiksi annotaatiomallin

$$l_{jk} = \mu_k + \sum_{c=1}^{C} A_{jc}\alpha_{jc} + \varepsilon_{jk}$$

k-kohtainen vakiotermiparametri μ_k on selvästi estimoitava parametri Bayes-mallissa,

mutta lähdemateriaalissa parametrin priorijakaumaa ei löydy mallin yhteistiheysjakaumasta. Gibbsin otanta-algoritmista 1 ei myöskään löydy kohtaa, jossa parametrille poimittaisiin arvo täysehdollisesta jakaumasta. Luultavasti parametrin priorijakaumaa ei löydy yhteystiheysjakaumasta, koska parametrille on asetettu epäinformatiivinen priori, jonka tiheysfunktio on verrannollinen arvoon 1, jolloin priorijakauma katoaa yhteistiheysjakaumasta. Samoin parametrille poimitaan Gibbsin otannassa luultavasti arvo annotaatiomallin mukaisesta täysehdollisesta jakaumasta jakaumaseosryhmiä k käsittelevässä lohkossa, mutta koska vakiotermiparametrit eivät sinänsä ole kiinnostavia annotaatiovaikutuksiin verrattuna, tämä on jätetty pois otanta-algoritmin pseudokoodiesityksestä.

6.2 Sisäisten verenvuotojen ennustaminen polygeenisen riskisumman avulla

Tutkielman sovelluksessa hyödynnettiin kahden eri GWAS-tutkimuksen – Zhoun ja kumppanien sekä Jiangin ja kumppanien – GWAS-tunnuslukuja. Sekä kallonsisäisen että suolistonsisäisen verenvuodon tapauksessa Zhoun ja kumppanien GWAS-tunnuslukuihin perustuvat PRS-muuttujat toimivat paremmin kuin Jiangin ja kumppanien GWAS-tunnuslukuihin perustuvat. Ero oli selvin vastetapahtuman ollessa kallonsisäinen verenvuoto, jolloin SBayesRC-menetelmällä lasketun jatkuvan PRS-muuttujan uhkasuhde Coxin mallissa oli Zhoun ja kumppanien tapauksessa 1.056 (90% LV 0.999–1.117) ja Jiangin ja kumppanien tapauksessa 0.998 (90% LV 0.943–1.055). Tulos oli samansuuntainen myös muilla PRS-versioilla. Suolistonsisäisen verenvuodon tapauksessa erot eivät olleet yhtä suuria, mutta myös tällöin Zhoun ja kumppanien GWAS-tunnuslukuihin perustuvat PRS-muuttujat toimivat paremmin. SBayesRC-menetelmällä lasketun jatkuvan PRS-muuttujan uhkasuhde Coxin mallissa oli Zhoun ja kumppanien tapauksessa 1.040 (90% LV 0.999–1.084) ja Jiangin ja kumppanien tapauksessa 1.038 (90% LV 0.996–1.080). PRS-CS-menetelmällä lasketuilla PRS-muuttujilla uhkasuhteet olivat Zhoun ja kumppanien tapauksessa 1.046 (90% LV 1.005–1.090) ja Jiangin ja kumppanien tapauksessa 1.021 (90% LV 0.980–1.063).

GWAS-tunnuslukujen suorituskyvyn eroille voidaan esittää ainakin neljä mahdollista syytä. Ensimmäinen liittyy GWAS-tutkimusten otoskokoihin, ja varsinkin tapausryhmän kokoon. Jotta selittävien muuttujien yhteydet vastetapahtumaan saataisiin estimoitua luotettavasti, tulee aineistossa olla tarpeeksi yksilöitä, jotka ovat kohdanneet vastetapahtuman. Kallonsisäisen verenvuodon tapauksessa Zhoun ja kumppanien GWAS-tutkimuksessa oli GWAS-katalogin mukaan 400 813 yksilöä, joista 1796 (0.45%) kohtasi vastetapahtuman, kun taas Jiangin ja kumppanien tutkimuksessa oli GWAS-katalogin mukaan 456 348 yksilöä, joista 158 (0.03%) yksilöä kohtasi vastetapahtuman. Vaikka Jiangin ja kumppanien tutkimuksessa oli yli 50 000 tutkimusyksilöä enemmän kuin Zhoun ja kumppanien tutkimuksessa, oli Zhoun ja kumppanien tutkimuksessa tapausryhmän koko yli kymmenkertainen verrattuna Jiangin ja kumppanien tutkimukseen. Suolistonsisäisen verenvuodon tapauksessa Zhoun ja kumppanien tutkimuksessa oli GWAS-katalogin mukaan 406 294 yksilöä, joista 21 137 (5.20%) kohtasi vastetapahtuman, kun taas Jiangin ja kumppanien tutkimuksessa oli GWAS-katalogin mukaan 456 348 yksilöä, joista 1377 (0.30%)

kohtasi vastetapahtuman. Myös tässä tapauksessa Zhoun ja kumppanien tutkimuksessa tapausryhmän koko on yli kymmenkertainen Jiangin ja kumppanien tutkimukseen verrattuna, vaikka kokonaisotoskoko on Zhoun ja kumppanien tapauksessa pienempi. Pienet tapausryhmän otoskoot voivat siis olla syynä sille, miksi Jiangin ja kumppanien GWAS-tunnuslukuihin perustuvat PRS-muuttujat eivät sovelluksessa suoriutuneet yhtä hyvin kuin Zhoun ja kumppanien GWAS-tunnuslukuihin perustuvat.

Toinen mahdollinen selitys liittyy imputoitavien snippien määrään GWAS-tutkimuksissa. Zhoun ja kumppanien GWAS-tunnuslukujen tapauksessa marginaalivaikutukset imputoitiin 12 053 snipille, eli noin 0.16% lopullisista käytetyistä snipeistä. Jiangin ja kumppanien GWAS-tunnuslukujen tapauksessa marginaalivaikutukset imputoitiin 240 743 snipille, eli noin 3.27% lopullisista snipeistä. Vaikka 3.27% ei ole suuri osuus käytetyistä snipeistä, se on silti huomattavasti suurempi osuus kuin 0.16% Zhoun ja kumppanien tapauksessa. Koska GWAS-tutkimuksessa estimoidut marginaalivaikutukset ovat yleisesti ottaen parempia kuin korrelaatiorakenteen perusteella imputoidut marginaalivaikutukset, voi suurempi imputoitujen marginaalivaikutusten osuus osaltaan selittää eroa GWAS-tunnuslukujen toimivuudessa polygeenisessa prediktiossa.

Kolmas selvä ero Zhoun ja kumppanien sekä Jiangin ja kumppanien GWAS-tunnusluken välillä oli suolistonsisäisen verenvuodon tapauksessa ero vastetapahtumissa. Zhoun ja kumppanien GWAS-tunnusluvuissa vastetapahtuma oli suolistonsisäinen verenvuoto (*gastrointestinal hemorrhage*), PheCode 578. Jiangin ja kumppanien GWAS-tutkimuksesta ei löytynyt GWAS-tunnuslukuja täysin samalle vastetapahtumalle, vaan sen sijaan käytettiin GWAS-tunnuslukuja verrattavissa olevalle vastetapahtumalle veren oksentaminen (*hematemesis*), PheCode 578.1. Vaikka veren oksentaminen on yhteydessä suolistonsisäiseen verenvuotoon, voivat erot vastetapahtumissa selittää eroja GWAS-tunnuslukuihin perustuvien PRS-muuttujien toimivuudessa suolistonsisäisen verenvuodon tapauksessa. Neljäs ero liittyy käytettyihin estimointimenetelmiin. Zhoun ja kumppanien GWAS-tutkimuksissa käytettiin heidän kehittämäänsä yleistettyyn lineaariseen sekamalliin pohjautuvaa SAIGE-menetelmää, jonka he esittävät antavan lineaarista ja logistista sekamallia tarkempia p-arvoja tilanteessa, jossa tapausryhmän koko on huomattavasti pienempi kuin verrokkiryhmän. Jiangin ja kumppanien GWAS-tutkimuksessa taas käytettiin heidän kehittämäänsä myös yleistettyyn lineaariseen sekamalliin pohjautuvaa fastGWA-GLMM-menetelmää. Myös he esittävät menetelmän antavan tarkkoja tuloksia tilanteissa, joissa tapauksia on radikaalisti vähemmän kuin verrokkeja. Erot käytetyissä menetelmissä voivat osaltaan selittää eroja PRS-muuttujien toimivuudessa.

Sovelluksen analyysin lopputuloksena oli, että polygeeninen riskisumma (PRS) ei paranna kallon- tai suolistonsisäisen verenvuodon ennustusta. Suuressa osaa testattavia PRS-muuttujia yhteys vastemuuttujaan ei ollut tilastollisesti merkitsevä asetetulla kriteerillä. Niille PRS-muuttujille, joille löytyi tilastollisesti merkitsevä yhteys vastemuuttujaan, ennustetarkkuuden jatkotarkastelu bootstrap-validoiduilla C-indekseillä osoitti, että PRS-muuttujan lisääminen Coxin regressiomalliin ei paranna mallin ennustetarkkuutta huomattavalla tavalla.

C-indeksin käyttöä mallinvalinnassa on myös kritisoitu. Cookin mukaan C-tunnusluku (*C-statistic*) voi muuttua hyvin vaatimattomasti, vaikka malliin lisättäisiin klii-

nisesti merkittävä ja tilastollisesti merkitsevä selittävä muuttuja, eikä hän suosittele pelkästään sen käyttöä selittävien muuttujien ennustekyvyn arvioinnissa [8]. C-tunnusluku vastaa käytännössä C-indeksiä, C-tunnusluku koskee binääristä aineistoa (onko yksilö kohdannut vastetapahtuman) ja C-indeksi on sen yleistys elinaika-aineistolle (kauanko kestää että yksilö kohtaa vastetapahtuman).

Vaikka oletettaisiin, että muutos C-indeksissä aliarvioi uuden selittävän muuttujan tuomaa lisää mallin ennustekykyyn, puhuu tutkielman analyysi kuitenkin kokonaisuutena sen puolesta, että PRS ei paranna mallin ennustekykyä kallon- tai suolistonsisäisen verenvuodon tapauksessa. Ensinnäkään myöskään Coxin regressiomallit eivät osoittaneet voimakasta yhteyttä PRS-muuttujien ja vastemuuttujien välillä. Jos tarkastellaan vain Zhoun ja kumppanien GWAS-tunnuslukuihin perustuvia PRS-muuttujia, neljästä lasketusta jatkuvasta PRS-muuttujasta vain yksi oli tilastollisesti merkitsevä: PRS-CS-menetelmällä laskettu PRS suolistonsisäisen verenvuodon tapauksessa (US 1.05, 90% LV 1.01–1.09). Tässäkään tapauksessa yhteyttä ei voi kutsua erityisen voimakkaaksi. Lisäksi kaikkien neljän kategorisen PRS-muuttujan tapauksessa kvantiililuokat eivät käyttäytyneet johdonmukaisesti, mikä puhuu voimakasta yhteyttä vastaan. Toiseksi ennustetarkkuuden jatkotarkasteluissa C-indeksien bootstrap-validoitu erotus oli joko negatiivinen (PRS-muuttujan lisääminen laski ennustetarkkuutta) tai positiivinen mutta erittäin lähellä arvoa 0 ($<$0.0005); 90% uusioluottamusväli sisälsi kaikissa tapauksissa arvon 0. Tällaisten arvojen on vaikea nähdä kertovan muusta kuin, että PRS-muuttujat eivät tuo lisäarvoa mallin ennustekykyyn, vaikka oletettaisiin, että C-indeksien muutos aliarvioisi jonkin verran selittävän muuttujan ennustekykyarvoa.

Vertailumielessä kokeiltiin, kuinka kategorisen ikämuuttujan poistaminen vaikuttaa C-indeksin arvoon. Korkea ikä on sekä tutkielman analyysien että kliinisen tiedon valossa erittäin merkittävä ennustaja kallon- ja suolistonsisäiselle verenvuodolle, kuten myös monille muille sairauksille. Bootstrap-validoitu C-indeksien erotus laskettiin kallonsisäisen verenvuodon tapauksessa mallille, jossa mukana olivat Zhoun ja kumppanien GWAS-tunnuslukuihin perustuva, SBayesRC-menetelmällä laskettu jatkuva PRS-muuttuja sekä muut kovariaatit, sekä muuten samalle mallille, mutta ilman kategorista ikämuuttujaa. Bootstrap-validoitu C-indeksien erotuksen estimaatti oli 0.037, ja 90% uusioluottamusväli oli (0.015, 0.057). Kategorinen ikämuuttuja toi siis melko huomattavan lisän (0.037 yksikköä) C-indeksiin, ja koska uusioluottamusväli ei sisältänyt arvoa 0, on tämä lisä ennustetarkkuuteen tilastollisesti merkitsevä. Sama tarkastelu suoritettiin suolistonsisäiselle verenvuodolle. Käytetty PRS-muuttuja oli myös nyt Zhoun ja kumppanien GWAS-tunnuslukuihin perustuva, SBayesRC-menetelmällä laskettu jatkuva PRS. Bootstrap-validoitu C-indeksien erotuksen estimaatti oli 0.016 ja 90% uusioluottamusväli (0.005, 0.026). Kategorisen ikämuuttujan tuoma lisä C-indeksiin oli pienempi kuin kallonsisäisen verenvuodon tapauksessa, mutta uusioluottamusvälin perusteella tilastollisesti merkitsevä. Tilanteessa, jossa selittävän muuttujan (kategorinen ikä) tiedetään olevan vahvassa yhteydessä vastetapahtumaan, käytetty menetelmä siis paljasti tilastollisesti merkitsevän lisän ennustetarkkuuteen. Tämä antaa tukea sille, että PRS-muuttujien hyvin vähäinen vaikutus C-indeksiin ei johdu menetelmän epäherkkyydestä, vaan PRS-muuttujien vähäisestä vaikutuksesta ennustetarkkuuteen.

6.3 Aikaisempi tutkimus

Polygeenisen riskisumman hyödyntämisestä kallon- tai suolistonsisäisen verenvuodon ennustamisessa on hyvin vähän aikaisempaa tutkimusta. Mayerhoferin ja kumppanien tutkimuksessa polygeeniset riskisummat olivat tilastollisesti merkitsevässä yhteydessä aivoverenvuotoon (*intracerebral hemorrhage*), ja PRS-muuttujan hyödyntäminen toi lisäarvoa mallin ennustetarkkuuteen verrattuna malliin, jossa käytettiin vain kliinistä riskisummamuuttujaa [23]. Tutkielman sovelluksen tulos on siis ristiriidassa tämän aikaisemman tutkimustuloksen kanssa. Mayerhoferin ja kumppanien tutkimuksen ja tutkielman sovelluksen välillä on eroja, jotka voivat selittää eroja tulosten välillä. Ensimmäinen ero koskee vastemuuttujia, Mayerhoferin ja kumppanien tutkimuksessa vastemuuttuja oli aivoverenvuoto, joka on kallonsisäisen verenvuodon alalaji. Toinen ero on käytetyt GWAS-tunnusluvut. Mayerhoferin ja kumppanien tutkimuksessa käytettiin Woon ja kumppanien aivoverenvuodon GWAS-tunnuslukuja, jotka muodostettiin kuuden aikaisemman tutkimuksen meta-analyysissä [36]. Woon ja kumppanien GWAS-tutkimuksen lopulliset tutkimusyksiköt koostuivat 1545 (51.1%) tapausyksilöstä ja 1481 (48.9%) verrokista, yhteensä 3026 yksilöstä. Tutkimusyksilöt olivat yhdysvaltalaisia ja eurooppalaisia. Tutkielmassa hyödynnetyn Zhoun ja kumppanien kallonsisäisen verenvuodon GWAS-tutkimuksen kokonaisotoskoko 400813 oli huomattavasti suurempi kuin Woon ja kumppanien tutkimuksessa, mutta tapausryhmän koko 1796 (0.4%) oli jotakuinkin yhtä suuri, mikä tekee tapausryhmän suhteellisesta osuudesta hyvin pienen Zhoun ja kumppanien tutkimuksessa. On myös huomionarvoista, että Woon ja kumppanien GWAS-tutkimus keskittyy aivoverenvuotoon, kun taas Zhoun ja kumppanien GWAS-tutkimuksessa tuotettiin GWAS-tunnusluvut 1403 vastemuuttujalle. Herää kysymys, onko spesifimmässä GWAS-tutkimuksessa mahdollisesti otettu huomioon enemmän asiaan liittyviä aspekteja, mikä on voinut johtaa toimivampiin GWAS-tunnuslukuihin.

Kolmas ero tutkimusten välillä liittyy tutkimusyksilöihin. Mayerhoferin ja kumppanien tutkimuksessa käytettiin UK biopankin yksilöitä ja lopullinen PRS-muuttujan arviointi perustui 5530 antikoagulantteja käyttävään yksilöön, joista 940 (17.0%) kohtasi vastetapahtuman seuranta-aikana. Kokonaisotoskoko oli siis tässä tutkielmassa huomattavasti suurempi (83 186 FinnGen yksilöä), joskin tapausryhmä oli jokseenkin yhtä suuri kuin Mayerhoferin testityhmässä, sillä 970 (1.2%) FinnGen yksilöä kohtasi kallonsisäisen verenvuodon seuranta-aikana. Neljäs ero on GWAS-tunnuslukujen ja tutkimusyksilöiden välinen yhteensopivuus. Mayerhoferin ja kumppanien tutkimuksessa GWAS-tunnusluvut laskettiin yhdysvaltalaista ja eurooppalaista (tutkimukset Espanjassa, Puolassa ja Ruotsissa) syntyperää olevista yksilöistä ja tutkimusyksilöt olivat UK biopankin yksilöitä, kun taas tässä tutkielmassa GWAS-tunnusluvut laskettiin UK biopankin yksilöistä ja tutkimusyksilöt olivat suomalaisia FinnGen-yksilöitä. Viides ero koskee käytettyjä menetelmiä. Mayehofer ja kumppanit käyttivät PRS-muuttujien muodostamisessa LDpred2-menetelmää [25], muodostivat 153 eri PRS-muuttujaa käyttäen LDpred2-menetelmän eri muotoja ja parametriasetuksia ja valitsivat PRS-muuttujista parhaan validointiaineiston avulla. PRS-muuttujan yhteyden voimakkuutta uhkaan kohdata aivonsisäinen verenvuoto testattiin Coxin regressiomallilla, jossa PRS-muuttujan lisäksi selittävinä muuttujina olivat ikä, sukupuoli, genotyyppien 10 ensimmäistä pääkomponenttia sekä käytetty

genotyyppisiru. PRS-muuttujan tuomaa lisää ennustetarkkuuteen testattiin vertaamalla Unon C-indeksiä kahdessa Coxin regressiomallissa. Ensimmäisessä selittävänä muuttujana oli muokattu kliininen riskisummamuuttuja, jonka arvo määräytyi verenvuototapahtuman riskitekijöistä (yksi piste korkeasta verenpaineesta, yksi piste yli 65 vuoden iästä, yksi piste aikaisemmasta verenvuototapahtumasta jne.) Toisessa mallissa käytettiin muuten samaa riskisummamuuttujaa, mutta johon oli lisätty yksi piste korkeasta PRS-muuttujan arvosta. Tarkastelussa käytettiin siis karkeita, mutta käytännön kliinisessä työssä relevantteja selittäjiä.

Yhteenvetona mahdollisia selittäviä tekijöitä erossa tutkielman tuloksen ja aikaisemman tutkimuksen kanssa ovat lievä ero vastemuuttujissa, erot käytetyissä GWAS-tunnusluvuissa, erot käytettyjen GWAS-tunnuslukujen ja tutkimusyksilöiden yhteensopivuudessa, erot tutkimusyksilöissä ja otoskoossa sekä erot käytetyissä menetelmissä. On myös hieman kyseenalaista, osoitettiinko Mayerhoferin ja kumppanien tutkimuksessa todella, että PRS-muuttujan käyttö lisää aivoverenvuodon ennustetarkkuutta. Coxin regressiomallissa PRS-muuttujan yhteys vastemuuttujaan oli tilastollisesti merkitsevä, ja ennustetarkkuuden jatkotarkastelussa uhkasuhde sekä C-indeksi olivat suuremmat riskisummamuuttujassa, joka sisälsi PRS-komponentin. Kuitenkin kun katsotaan vertailtavien mallien uhkasuhteiden ja C-indeksien luottamusvälejä, vaikuttavat ne selvästi sisältävän verrattavana olevan mallin estimaatin. Näin ollen vaikka estimaatit ovat parempia mallissa, joka hyödyntää PRS-muuttujaa, on kyseenalaista, onko ero merkitsevä kun tilastollinen epävarmuus otetaan huomioon.

6.4 Analyysin rajoitteet

Tutkielman sovellukseen liittyy erinäisiä rajoituksia. Ensimmäinen rajoite koskee sovelluksen tutkimusyksilöiden ja käytettyjen GWAS-tutkimusten yksilöiden välistä yhteensopivuutta. Sovelluksen tutkimuskysymys koski verenohennuslääkitystä käyttäviä yksilöitä, ja voidaan ajatella että ideaalitapauksessa GWAS-tutkimuksissa olisi myös käytetty verenohennuslääkitystä käyttäviä yksilöitä. Tämä ei luonnollisesti ollut mahdollista käytettäessä valmiita saatavilla olevia GWAS-tunnuslukuja, eikä sovelluksessa hyödynnettyjen GWAS-tutkimusten yksilöiden verenohennuslääkitysten käytöstä ole tietoa. Lisäksi ideaalitapauksessa myös GWAS-tutkimusten yksilöt olisivat olleet suomalaista syntyperää, kun he nyt olivat UK biopankin yksilöitä.

Toinen rajoite koskee sovelluksessa käytettyjä GWAS-tunnuslukuja. Tutkielman sovelluksessa hyödynnettiin GWAS-tunnuslukuja kahdesta GWAS-tutkimuksesta, mutta myös muita GWAS-tunnuslukuja sopivilla vastemuuttujilla on olemassa, ja tulevaisuudessa niitä voidaan tehdä GWAS-tutkimuksissa vielä lisää. Yksi esimerkki on jo mainittu Woon ja kumppanien aivoverenvuodon GWAS-tunnusluvut [36], jota Mayerhofer ja kumppanit [23] käyttivät tutkimuksessaan. Olisi mielenkiintoista kokeilla näiden GWAS-tunnuslukujen pohjalta laskettujen PRS-muuttujien suorituskykyä tämän tutkielman sovelluksessa, mutta vaikka kyseinen tutkimus löytyy GWAS-katalogista [15] tunnuksella GCST002389, eivät GWAS-tunnusluvut ole siellä vapaasti saatavilla, eikä kyseisen GWAS-tutkimuksen artikkelistakaan löytynyt suoraa linkkiä GWAS-tunnuslukuihin. Lisäksi elokuussa 2025 UK biopankki (tarkemmin *The UK Biobank Whole-Genome Sequencing Consortium*) julkaisi uuden tut-

kimuksen, joka sisältää GWAS-tunnusluvut aivoverenvuodolle [33]. Nämä GWAS-tunnusluvut löytyvät GWAS-katalogista tunnuksella GCST90473584. Vaikka tutkielman sovelluksessa ei siis osoitettu käytetyillä GWAS-tunnusluvuilla polygeenisen riskisumman hyödyllisyyttä kallon- tai suolistonsisäisen verenvuodon ennustamisessa, ehkä joidenkin toisten GWAS-tunnuslukujen avulla tuotetut PRS-muuttujat toimivat paremmin.

Kolmas rajoite tutkielman sovelluksessa liittyy käytettävien menetelmien ja PRS-tarkastelun laajuuteen. Aiemmin mainitussa Mayerhoferin ja kumppanien tutkimuksessa [23] laskettiin 153 eri PRS-muuttujaa käyttäen LDpred2-menetelmän [25] eri asetuksia, ja näiden muuttujien toimintakykyä testattiin validointiaineistolla. Tämän tutkielman sovelluksessa laskettiin 8 PRS-muuttujaa kallonsisäiselle ja 8 PRS-muuttujaa suolistonsisäiselle verenvuodolle, jos lasketaan mukaan PRS-muuttujista johdetut kategoriset versiot. Sovelluksessa olisi ollut mahdollista ottaa huomioon useampia tekijöitä PRS-muuttujien laskemisessa ja testata useampia eri tavalla muodostettuja PRS-muuttujia. Vaihtoehtona sovelluksessa käytetylle noin 7 miljoonan snipin LD-referenssille olisi ollut Zhengin ja kumppanien [38] tutkimuksessaan käyttämä pienempi noin miljoonan snipin LD-referenssi, joka perustuu HapMap3-paneelin snippeihin [18]. PRS-muuttujat olisi voitu laskea myös tällä LD-referenssillä ja verrata PRS-muuttujien toimivuutta. Sovelluksessa käytettiin SBayesRC-menetelmää funktionaalisia annotaatioita hyödyntäen, mutta olisi ollut mahdollista testata menetelmää myös ilman funktionaalisia annotaatioita. SBayesRC-menetelmää käyttäessä valittiin ρ-parametrin (osuus LD-matriisin varianssista joka halutaan vähintään säilyttää matala-asteista mallia muodostettaessa) arvoksi 0.995. Tässäkin olisi ollut mahdollista kokeilla eri parametrin arvoja, esimerkiksi arvoja 0.99 tai 0.999. Tällaisten tekijöiden huomioonottaminen olisi kuitenkin nostanut tarkasteltavien PRS-muuttujien määrän moninkertaiseksi, joten tutkielman tavoitelaajuus huomioiden rajaus oli perusteltua. Suuremman LD-referenssin ja funktionaalisten annotaatioiden hyödyntäminen oli perusteltua myös siksi, että Zhengin ja kumppanien tutkimuksessa suuremman LD-referenssin ja funktionaalisten annotaatioiden käyttö johti systemaattisesti parempiin ennustetarkkuuksiin. Lisäksi funktionaalisten annotiaatioiden käyttö on yksi SBayesRC-menetelmän pääpiirteistä, joten niiden hyödyntämättä jättäminen olisi ollut kyseenalaista.

Neljäs rajoite koskee sovelluksen tapausryhmien otoskokoja. Vastetapahtumina suolistonsisäinen verenvuoto ja varsinkin kallonsisäinen verenvuoto ovat hyvin harvinaisia. Tästä syystä, vaikka sovelluksen tutkimusyksilöiden määrä 83 186 oli varsin suuri, oli kallonsisäisen verenvuodon elinaika-aineistossa vain 970 ja suolistonsisäisessä verenvuotoaineistossa 2074 vastetapahtuman kohdannutta yksilöä. Pienet tapausryhmien otoskoot kasvattavat tilastollista epävarmuutta muuttujien välisiä yhteyksiä estimoitaessa.

Vielä yksi rajoite koskee tutkimusyksilöiden verenohennuslääkityksen käyttöä. Sovelluksen analyyseissä haluttiin seurata verenohennuslääkitystä käyttäviä yksilöitä, ja tämä toteutettiin tarkastelemalla yksilöiden verenohennuslääkityksen ostotapahtumia. Ei ole kuitenkaan varmuutta siitä, että verenohennuslääkitystä ostanut yksilö myös käyttää sitä tarkoitetulla aikataululla, ja siksi tutkimusyksilöiden verenohennuslääkityksen käyttöön liittyy myös epävarmuutta. On kuitenkin luultavaa, että yleisesti ottaen verenohennuslääkitystä ostanut yksilö myös käyttää sitä.

Viitteet

[1] Albert, J. H., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association, 88(422), 669-679.

[2] Bastarache, L. (2021). Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annual Review of Biomedical Data Science, 4(1), 1-19.

[3] Breiman, L. (1996). Out-of-bag estimation. Technical Report, UC Berkeley, Department of Statistics.

[4] Bright, L. A., Burgess, S. C., Chowdhary, B. ym. (2009, October). Structural and functional-annotation of an equine whole genome oligoarray. In BMC Bioinformatics (Vol. 10, pp. 1-8). BioMed Central.

[5] Brown, T. A. (2023). Genomes 5. CRC Press.

[6] Cai, X., Teng, J., Ren, D. ym. (2022). Model Comparison of Heritability Enrichment Analysis in Livestock Population. Genes, 13(9), 1644.

[7] Choi, S. W., Mak, T. S. H., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nature Protocols, 15(9), 2759-2772.

[8] Cook, N. R. (2007). Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. Circulation, 115(7), 928-935.

[9] FinnGen. *Kohortti*. https://www.finngen.fi/fi/node/1987

[10] FinnGen. *Risteys*. https://risteys.finngen.fi/

[11] FinnGen. *Sairauspäätepisteet*. https://www.finngen.fi/fi/tutkijat/sairauspäätepisteet

[12] Finucane, H. K., Bulik-Sullivan, B., Gusev, A. ym. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics, 47(11), 1228-1235.

[13] Gazal, S., Finucane, H. K., Furlotte, N. A. ym. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics, 49(10), 1421-1427.

[14] Ge, T., Chen, C. Y., Ni, Y. ym. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature Communications, 10(1), 1776.

[15] GWAS Catalog. *GWAS-katalogi*. https://www.ebi.ac.uk/gwas/home

[16] Harrell, F. E., Califf, R. M., Pryor, D. B. ym. (1982). Evaluating the Yield of Medical Tests. JAMA, 247(18), 2543-2546.

[17] Hu, Y., Lu, Q., Powles, R. ym. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Computational Biology, 13(6), e1005589.

[18] International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature, 467(7311), 52-58.

[19] Jiang, L., Zheng, Z., Fang, H. ym. (2021). A generalized linear mixed model association tool for biobank-scale data. Nature Genetics, 53(11), 1616-1621.

[20] Krebs, J. E., Goldstein, E. S., & Kilpatrick, S. T. (2017). Lewin's Genes XII. Jones & Bartlett Learning.

[21] Lloyd-Jones, L. R., Zeng, J., Sidorenko, J. ym. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nature Communications, 10(1), 5086.

[22] Longato, E., Vettoretti, M., & Di Camillo, B. (2020). A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. Journal of Biomedical Informatics, 108, 103496.

[23] Mayerhofer, E., Parodi, L., Prapiadou, S. ym. (2023). Genetic Risk Score Improves Risk Stratification for Anticoagulation-Related Intracerebral Hemorrhage. Stroke, 54(3), 791-799.

[24] Palotie, A., Kaunisto, M., Harju, J. ym. (2019). FinnGen-tutkimuksen lupaukset. Duodecim, 135(10), 987-996.

[25] Privé, F., Arbel, J., & Vilhjálmsson, B. J. (2020). LDpred2: better, faster, stronger. Bioinformatics, 36(22-23), 5424-5431.

[26] Reed, J. L., Famili, I., Thiele, I. ym. (2006). Towards multidimensional genome annotation. Nature Reviews Genetics, 7(2), 130-141.

[27] Reich, D. E., Cargill, M., Bolk, S. ym. (2001). Linkage disequilibrium in the human genome. Nature, 411(6834), 199-204.

[28] Shoeb, M., & Fang, M. C. (2013). Assessing bleeding risk in patients taking anticoagulants. Journal of Thrombosis and Thrombolysis, 35, 312-319.

[29] Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics, 9(6), 477-485.

[30] Stein, L. (2001). Genome annotation: from sequence to biology. Nature Reviews Genetics, 2(7), 493-503

[31] Strandén, I. & Christensen, O. F. (2011). Allele coding in genomic evaluation. Genetics Selection Evolution, 43(1), 25.

[32] Uffelmann, E., Huang, Q. Q., Munung, N. S. ym. (2021). Genome-wide association studies. Nature Reviews Methods Primers, 1(1), 59.

[33] UK Biobank Whole-Genome Sequencing Consortium. (2025). Whole-genome sequencing of 490,640 UK Biobank participants. Nature, 645(8081), 692-701.

[34] International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921.

[35] Visscher, P. M., Hill, W. G. & Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. Nature Reviews Genetics, 9(4), 255-266.

[36] Woo, D., Falcone, G. J., Devan, W. J. ym. (2014). Meta-analysis of Genome-wide Association Studies Identifies 1q22 as a Susceptibility Locus for Intracerebral Hemorrhage. The American Journal of Human Genetics, 94(4), 511-521.

[37] Yang, J., Lee, S. H., Goddard, M. E. ym. (2011). GCTA: a Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics, 88(1), 76-82.

[38] Zheng, Z., Liu, S., Sidorenko, J. ym. (2024). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nature Genetics, 56(5), 767-777.

[39] Zhou, W., Nielsen, J. B., Fritsche, L. G. ym. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics, 50(9), 1335-1341.

Liite A Analyysi PRS-muuttujille joissa käytettiin Jiangin ja kumppanien GWAS-tunnuslukuja

A.1 Kallonsisäinen verenvuoto

A.1.1 SBayesRC, jatkuva PRS

Coxin regressiomallissa käytettiin ensin PRS-muuttujana SBayesRC-menetelmällä laskettua jatkuvaa PRS-muuttujaa, mallin tulokset on esitetty taulukossa 15. PRS-muuttujan uhkasuhde oli 0.998, siis hyvin lähellä arvoa 1. Lisäksi 90% luottamusväli sisälsi arvon 1, mikä merkitsee, että PRS-muuttuja ei ollut tilastollisesti merkitsevä valitulla kriteerillä. Koska log-uhkasuhde oli negatiivinen – tai yhtäpitävästi uhkasuhde oli alle arvon 1 – oli PRS-muuttujan yhteys vastemuuttujaan lisäksi negatiivinen: suurempi PRS oli yhteydessä pienempään uhkaan kohdata kallonsisäinen verenvuoto. Kun muuten saman mallin, mutta ilman yksilökohtaista satunnaisvaikutusta, C-indeksi laskettiin samasta aineistosta, sen arvoksi saatiin 0.660. Koska PRS-muuttuja ei ollut tilastollisesti merkitsevä, ei mallin ennustetarkkuuden jatkotarkastelua suoritettu bootstrap-validoiduilla C-indekseillä.

Muiden kovariaattien uhkasuhteet poikkesivat hyvin vähän malleista joissa PRS-muuttujat perustuivat Zhoun ja kumppanien GWAS-tunnuslukuihin, tilastollisesti merkitsevät kovariaatit olivat yhä yli 75 vuoden ikä (US 3.0), dementia (US 2.1), 65-75 vuoden ikä (US 1.7), aivohalvaus (US 1.7), alkoholiongelma (US 1.5), miessukupuoli (US 1.2) ja sydämen vajaatoiminta (US 1.1). Kovariaattien uhkasuhteet muuttuvat hyvin vähän muissa tämän luvun malleissa joissa vastemuuttuja oli aika kallonsisäiseen verenvuotoon, eikä niitä enää kommentoida tämän jälkeen.

A.1.2 SBayesRC, kategorinen PRS

Seuraavaksi testattiin SBayesRC-menetelmällä lasketun PRS-muuttujan kategorisoitua versiota, mallin tulokset on esitetty taulukossa 16. Kaikissa PRS-muuttujan luokissa uhkasuhteen 90% luottamusväli sisälsi arvon 1, mikään luokista ei siis ollut tilastollisesti merkitsevä valitulla kriteerillä. Lisäksi alimmalla tasolla log-uhkasuhde oli positiivinen ja ylimmällä negatiivinen, kun toimivalla riskisummamuuttujalla asian pitäisi olla toisin päin. Koska kategorisen PRS-muuttujan tasojen ja vastemuuttujan välillä ei ollut tilastollisesti merkitsevää yhteyttä, ei mallin ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla suoritettu. Muuten saman mallin mutta ilman satunnaisvaikutusta C-indeksi samalla aineistolla laskettuna oli 0.660.

A.1.3 PRS-CS, jatkuva PRS

Sama tarkastelu suoritettiin PRS-muuttujalla jonka laskemisessa käytettiin PRS-CS menetelmää. Coxin regressiomallin jossa käytettiin jatkuvaa PRS-muuttujaa tulokset on esitetty taulukossa 17. Kuten SBayesRC-menetelmällä lasketussa versiossa, PRS-muuttujan uhkasuhde oli 0.998 ja sen 90% luottamusväli sisälsi arvon 1. PRS-muuttuja ei siis ollut tilastollisesti merkitsevässä yhteydessä vastemuuttujaan valitulla kriteeritasolla, ja estimoitu yhteys on väärän suuntainen (suurempi PRS oli yhteydessä pienempään uhkaan kohdata kallonsisäinen verenvuoto). Kun muuten

Muuttuja	Log-US	US	US 90% LV	P-arvo
PRS (Jiang IC SBayesRC)	-0.002	0.998	(0.943, 1.055)	0.943
Ikä 65–75	0.551	1.734	(1.460, 2.060)	<0.001
Ikä yli 75	1.094	2.986	(2.516, 3.543)	<0.001
Mies	0.173	1.189	(1.053, 1.342)	0.019
Sukupuoli tuntematon	0.249	1.283	(0.906, 1.816)	0.238
Verisuonitauti	0.055	1.057	(0.930, 1.201)	0.478
Kohonnut verenpaine	-0.036	0.965	(0.758, 1.228)	0.807
Dyslipidemia	0.036	1.037	(0.902, 1.191)	0.668
Sydämen vajaatoiminta	0.138	1.148	(1.008, 1.309)	0.082
Diabetes	0.046	1.047	(0.924, 1.187)	0.545
Aivohalvaus	0.539	1.715	(1.489, 1.975)	<0.001
Alkoholiongelma	0.434	1.544	(1.145, 2.083)	0.017
Munuaisen vajaatoiminta tai dialyysi	0.191	1.210	(0.917, 1.597)	0.258
Maksan vajaatoiminta tai kirroosi	0.334	1.397	(0.719, 2.715)	0.408
Dementia	0.753	2.124	(1.721, 2.621)	<0.001
Psykiatrinen häiriö	0.126	1.134	(0.965, 1.333)	0.200

Taulukko 15: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, SBayesRC-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

saman mallin, mutta ilman yksilökohtaista satunnaisvaikutusta, C-indeksi laskettiin samasta aineistosta, sen arvoksi saatiin 0.660. Koska PRS-muuttuja ei ollut tilastollisesti merkitsevä, ei mallin ennustetarkkuuden jatkotarkastelua suoritettu bootstrap-validoiduilla C-indekseillä.

A.1.4 PRS-CS, kategorinen PRS

Seuraavaksi Coxin regressiomallissa käytettiin PRS-CS-menetelmällä lasketun PRS-muuttujan kategorista versiota, mallin tulokset on esitetty taulukossa 18. Kaikissa PRS-muuttujan luokissa uhkasuhteen 90% luottamusväli sisälsi arvon 1, eli mikään luokista ei ollut tilastollisesti merkitsevässä yhteydessä vastemuuttujaan valitulla kriteeritasolla. Log-uhkasuhteet olivat itseisarvoltaan pieniä eivätkä käyttäytyneet toivotulla tavalla, kahdessa korkeimmassa PRS-luokassa log-uhkasuhteet olivat negatiivisia kun niiden olisi pitänyt olla positiivisia ja luokassa 0.025–0.2 log-uhkasuhde oli positiivinen kun sen olisi pitänyt olla negatiivinen. Koska yhteys vastemuuttujaan ei ollut tilastollisesti merkitsevä missään kategorisen PRS-muuttujan luokassa, ei ennustetarkkuuden jatkotarkastelua C-indeksien bootstrap-validoinnilla suoritettu. Muuten saman mallin mutta ilman yksilökohtaista satunnaisvaikutusta samasta aineistosta laskettu C-indeksi oli 0.66.

Muuttuja	Log-US	US	US 90% LV	P-arvo
PRS (0 - 0.025)	0.068	1.070	(0.761, 1.504)	0.744
PRS (0.025 - 0.2)	-0.022	0.978	(0.838, 1.142)	0.815
PRS (0.8 - 0.975)	0.031	1.032	(0.888, 1.200)	0.732
PRS (0.975 - 1)	-0.079	0.924	(0.638, 1.339)	0.726
Ikä 65–75	0.550	1.733	(1.459, 2.059)	<0.001
Ikä yli 75	1.092	2.982	(2.513, 3.537)	<0.001
Mies	0.173	1.188	(1.053, 1.341)	0.019
Sukupuoli tuntematon	0.249	1.283	(0.907, 1.814)	0.237
Verisuonitauti	0.054	1.056	(0.930, 1.199)	0.482
Kohonnut verenpaine	-0.037	0.964	(0.757, 1.226)	0.800
Dyslipidemia	0.036	1.037	(0.903, 1.191)	0.669
Sydämen vajaatoiminta	0.137	1.147	(1.006, 1.306)	0.085
Diabetes	0.046	1.047	(0.924, 1.186)	0.548
Aivohalvaus	0.537	1.711	(1.486, 1.969)	<0.001
Alkoholiongelma	0.433	1.542	(1.145, 2.079)	0.017
Munuaisen vajaatoiminta tai dialyysi	0.191	1.210	(0.917, 1.596)	0.258
Maksan vajaatoiminta tai kirroosi	0.333	1.395	(0.719, 2.707)	0.409
Dementia	0.749	2.115	(1.715, 2.608)	<0.001
Psykiatrinen häiriö	0.125	1.134	(0.965, 1.332)	0.201

Taulukko 16: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, SBayesRC-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

Muuttuja	Log-US	US	US 90% LV	P-arvo
PRS (Jiang IC PRS-CS)	-0.002	0.998	(0.944, 1.056)	0.954
Ikä 65–75	0.551	1.734	(1.460, 2.060)	<0.001
Ikä yli 75	1.094	2.986	(2.516, 3.543)	<0.001
Mies	0.173	1.189	(1.053, 1.342)	0.019
Sukupuoli tuntematon	0.249	1.283	(0.907, 1.816)	0.238
Verisuonitauti	0.055	1.057	(0.930, 1.201)	0.478
Kohonnut verenpaine	-0.036	0.965	(0.758, 1.228)	0.806
Dyslipidemia	0.036	1.037	(0.902, 1.191)	0.669
Sydämen vajaatoiminta	0.138	1.148	(1.008, 1.309)	0.082
Diabetes	0.046	1.047	(0.924, 1.187)	0.545
Aivohalvaus	0.539	1.715	(1.489, 1.975)	<0.001
Alkoholiongelma	0.434	1.544	(1.145, 2.083)	0.017
Munuaisen vajaatoiminta tai dialyysi	0.191	1.210	(0.917, 1.597)	0.258
Maksan vajaatoiminta tai kirroosi	0.334	1.397	(0.718, 2.715)	0.408
Dementia	0.753	2.124	(1.721, 2.621)	<0.001
Psykiatrinen häiriö	0.126	1.134	(0.965, 1.333)	0.200

Taulukko 17: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, PRS-CS-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (0 - 0.025)	-0.039	0.961	(0.668, 1.384)	0.859
PRS (0.025 - 0.2)	0.033	1.033	(0.887, 1.204)	0.726
PRS (0.8 - 0.975)	-0.005	0.995	(0.854, 1.160)	0.960
PRS (0.975 - 1)	-0.022	0.978	(0.678, 1.410)	0.920
Ikä 65–75	0.551	1.735	(1.460, 2.061)	<0.001
Ikä yli 75	1.094	2.987	(2.517, 3.545)	<0.001
Mies	0.173	1.189	(1.053, 1.342)	0.019
Sukupuoli tuntematon	0.248	1.282	(0.905, 1.815)	0.240
Verisuonitauti	0.055	1.057	(0.930, 1.201)	0.478
Kohonnut verenpaine	-0.035	0.965	(0.758, 1.228)	0.809
Dyslipidemia	0.036	1.037	(0.902, 1.191)	0.669
Sydämen vajaatoiminta	0.139	1.149	(1.008, 1.309)	0.081
Diabetes	0.046	1.047	(0.924, 1.187)	0.544
Aivohalvaus	0.540	1.716	(1.490, 1.976)	<0.001
Alkoholiongelma	0.435	1.546	(1.146, 2.085)	0.017
Munuaisen vajaatoiminta tai dialyysi	0.191	1.210	(0.917, 1.598)	0.258
Maksan vajaatoiminta tai kirroosi	0.335	1.398	(0.719, 2.719)	0.407
Dementia	0.754	2.126	(1.723, 2.624)	<0.001
Psykiatrinen häiriö	0.126	1.135	(0.965, 1.334)	0.199

Taulukko 18: Coxin regressiomallin tulokset, kallonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, PRS-CS-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

A.2 Suolistonsisäinen verenvuoto

A.2.1 SBayesRC, jatkuva PRS

Aikaa suolistonsisäiseen verenvuototapahtumaan ennustettiin ensin Coxin regressiomallilla, jossa jatkuva PRS-muuttuja oli laskettu SBayesRC-menetelmällä, mallin tulokset on esitetty taulukossa 19. PRS-muuttujan uhkasuhde 1.04 on oikean suuntainen, suurempi PRS on yhteydessä lisääntyneeseen uhkaan kohdata suolistonsisäinen verenvuoto. Uhkasuhteen 90% luottamusväli kuitenkin sisälsi arvon 1, eli yhteys ei ollut tilastollisesti merkitsevä asetetulla kriteerillä $p < 0.1$. Kun muuten saman mallin mutta ilman yksilökohtaista satunnaisvaikutusta C-indeksi laskettiin samasta aineistosta, saatiin arvo 0.658. Koska PRS-muuttuja ei ollut tilastollisesti merkitsevä, ei mallin ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla suoritettu.

Muiden kovariaattien uhkasuhteet eivät poikenneet suuresti siitä mitä ne olivat malleissa joissa PRS-muuttujat oli laskettu käyttäen Zhoun ja kumppanien GWAS-tunnuslukuja, ainoat kovariaatit joiden yhteys ei ollut tilastollisesti merkitsevä olivat tuntematon sukupuoli ja dyslipidemia. Korkeimmat uhkasuhteet olivat kovariaateilla yli 75 vuoden ikä (US 2.2), alkoholiongelma (US 2.2), munuaisen vajaatoiminta (US 2.0), maksan vajaatoiminta (US 1.9) ja sydämen vajaatoiminta (US 1.6). Muilla tilastollisesti merkitsevillä kovariaateilla uhkasuhteet vaihtelivat välillä 1.1–1.4. Näiden kovariaattien uhkasuhteet eivät muuttuneet suuresti vaihtaessa PRS-muuttujaa toisiin Jiangin ja kumppanien GWAS-tunnuslukuihin perustuviin PRS-muuttujiin, eikä niitä kommentoida enää tämän jälkeen.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (Jiang GI SBayesRC PRS)	0.037	1.038	(0.996, 1.080)	0.135
Ikä 65–75	0.243	1.276	(1.139, 1.428)	<0.001
Ikä yli 75	0.811	2.249	(2.009, 2.518)	<0.001
Mies	0.230	1.259	(1.153, 1.375)	<0.001
Sukupuoli tuntematon	0.058	1.060	(0.809, 1.389)	0.724
Verisuonitauti	0.238	1.268	(1.158, 1.390)	<0.001
Kohonnut verenpaine	0.323	1.381	(1.137, 1.677)	0.006
Dyslipidemia	-0.060	0.942	(0.853, 1.039)	0.317
Sydämen vajaatoiminta	0.499	1.647	(1.508, 1.799)	<0.001
Diabetes	0.231	1.260	(1.155, 1.375)	<0.001
Aivohalvaus	0.135	1.144	(1.024, 1.279)	0.047
Alkoholiongelma	0.796	2.216	(1.817, 2.703)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.669	1.953	(1.650, 2.311)	<0.001
Maksan vajaatoiminta tai kirroosi	0.638	1.893	(1.266, 2.832)	0.009
Dementia	0.195	1.216	(1.013, 1.459)	0.078
Psykiatrinen häiriö	0.153	1.165	(1.037, 1.309)	0.031

Taulukko 19: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, SBayesRC-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

A.2.2 SBayesRC, kategorinen PRS

Seuraavaksi SBayesRC-menetelmällä laskettu PRS-muuttuja korvattiin sen kategorisella versiolla, mallin tulokset on esitetty taulukossa 20. Kaikissa kategorisen PRS-muuttujan luokissa uhkasuhteen 90% luottamusväli sisälsi arvon 1, luokat eivät siis olleet tilastollisesti merkitseviä. Tästä syystä mallin ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla ei suoritettu. Muuten saman mallin, mutta ilman yksilökohtaista satunnaisvaikutusta, C-indeksi laskettuna samalle aineistolle oli 0.658.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (0 - 0.025)	-0.106	0.900	(0.689, 1.175)	0.515
PRS (0.025 - 0.2)	0.011	1.011	(0.905, 1.129)	0.872
PRS (0.8 - 0.975)	0.049	1.051	(0.942, 1.171)	0.455
PRS (0.975 - 1)	0.013	1.013	(0.779, 1.318)	0.934
Ikä 65–75	0.243	1.275	(1.139, 1.427)	<0.001
Ikä yli 75	0.810	2.249	(2.008, 2.517)	<0.001
Mies	0.230	1.259	(1.153, 1.375)	<0.001
Sukupuoli tuntematon	0.060	1.062	(0.810, 1.391)	0.716
Verisuonitauti	0.238	1.269	(1.158, 1.391)	<0.001
Kohonnut verenpaine	0.324	1.383	(1.139, 1.679)	0.006
Dyslipidemia	-0.060	0.942	(0.853, 1.040)	0.319
Sydämen vajaatoiminta	0.500	1.649	(1.510, 1.801)	<0.001
Diabetes	0.232	1.261	(1.155, 1.375)	<0.001
Aivohalvaus	0.134	1.144	(1.023, 1.279)	0.048
Alkoholiongelma	0.797	2.219	(1.819, 2.708)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.671	1.956	(1.653, 2.315)	<0.001
Maksan vajaatoiminta tai kirroosi	0.637	1.890	(1.263, 2.828)	0.009
Dementia	0.195	1.215	(1.012, 1.458)	0.079
Psykiatrinen häiriö	0.154	1.166	(1.038, 1.310)	0.030

Taulukko 20: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, SBayesRC-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

A.2.3 PRS-CS, jatkuva PRS

Seuraavaksi testattiin PRS-CS-menetelmällä laskettuja PRS-muuttujia. Jatkuvan PRS-muuttujan sisältävän Coxin regressiomallin tulokset on esitetty taulukossa 21. PRS-muuttujan uhkasuhde oli 1.02, eli mallin mukaan uhka kohdata suolistonsisäinen verenvuototapahtuma on suurempi niillä yksilöillä, joilla on suurempi PRS-arvo. Uhkasuhteen 90% luottamusväli sisälsi kuitenkin arvon 1, eli yhteys ei ollut tilastollisesti merkitsevä kriteerillä $p < 0.1$. Koska PRS-muuttuja ei ollut tilastollisesti merkitsevä, ei ennustetarkkuuden jatkotarkastelua bootstrap-validaatiolla suoritettu. Muuten saman mallin, mutta ilman yksilökohtaista satunnaisvaikutusta, C-indeksi laskettuna samalle aineistolle oli 0.658.

Muuttuja	Log-US	US	US 90% LV	P-arvo
PRS (Jiang GI PRS-CS PRS)	0.020	1.021	(0.980, 1.063)	0.409
Ikä 65–75	0.243	1.275	(1.139, 1.428)	<0.001
Ikä yli 75	0.810	2.248	(2.008, 2.516)	<0.001
Mies	0.230	1.259	(1.153, 1.375)	<0.001
Sukupuoli tuntematon	0.061	1.063	(0.811, 1.392)	0.711
Verisuonitauti	0.237	1.268	(1.157, 1.389)	<0.001
Kohonnut verenpaine	0.323	1.382	(1.138, 1.678)	0.006
Dyslipidemia	-0.060	0.942	(0.853, 1.040)	0.318
Sydämen vajaatoiminta	0.499	1.647	(1.508, 1.799)	<0.001
Diabetes	0.231	1.260	(1.155, 1.374)	<0.001
Aivohalvaus	0.134	1.143	(1.023, 1.278)	0.048
Alkoholiongelma	0.795	2.215	(1.816, 2.701)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.669	1.951	(1.649, 2.309)	<0.001
Maksan vajaatoiminta tai kirroosi	0.638	1.893	(1.266, 2.830)	0.009
Dementia	0.194	1.214	(1.012, 1.456)	0.080
Psykiatrinen häiriö	0.153	1.166	(1.038, 1.309)	0.030

Taulukko 21: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, PRS-CS-menetelmä, jatkuva PRS. US = uhkasuhde, LV = luottamusväli.

A.2.4 PRS-CS, kategorinen PRS

Viimeisenä PRS-muuttujana käytettiin PRS-CS-menetelmällä lasketun PRS-muuttujan kategorista versiota, mallin tulokset on esitetty taulukossa 22. PRS-muuttujan luokkien log-uhkasuhteet ovat oikean suuntaiset (mitä matalampi PRS-taso sitä pienempi log-uhkasuhde ja mitä korkeampi PRS-taso sitä suurempi log-uhkasuhde), mutta kaikissa luokissa uhkasuhteen 90% luottamusväli sisälsi arvon 1, eli luokat eivät olleet tilastollisesti merkitseviä kriteerillä $p < 0.1$. Koska mikään kategorisen PRS-muuttujan luokista ei ollut tilastollisesti merkitsevä, ei ennustetarkkuuden jatkotarkastelua bootstrap-validoinnilla suoritettu. Muuten saman mallin, mutta ilman yksilökohtaista satunnaisvaikutusta, C-indeksi laskettuna elinaika-aineistolle oli 0.658.

Muuttuja	**Log-US**	**US**	**US 90% LV**	**P-arvo**
PRS (0 - 0.025)	-0.132	0.877	(0.665, 1.156)	0.434
PRS (0.025 - 0.2)	-0.023	0.977	(0.875, 1.091)	0.733
PRS (0.8 - 0.975)	0.043	1.044	(0.936, 1.164)	0.517
PRS (0.975 - 1)	0.064	1.066	(0.818, 1.388)	0.692
Ikä 65–75	0.243	1.275	(1.139, 1.428)	<0.001
Ikä yli 75	0.810	2.247	(2.008, 2.516)	<0.001
Mies	0.230	1.259	(1.153, 1.374)	<0.001
Sukupuoli tuntematon	0.061	1.063	(0.811, 1.392)	0.711
Verisuonitauti	0.237	1.268	(1.157, 1.389)	<0.001
Kohonnut verenpaine	0.323	1.382	(1.138, 1.678)	0.006
Dyslipidemia	-0.060	0.942	(0.853, 1.039)	0.317
Sydämen vajaatoiminta	0.499	1.647	(1.508, 1.799)	<0.001
Diabetes	0.231	1.259	(1.154, 1.374)	<0.001
Aivohalvaus	0.134	1.143	(1.023, 1.278)	0.048
Alkoholiongelma	0.796	2.216	(1.817, 2.702)	<0.001
Munuaisen vajaatoiminta tai dialyysi	0.669	1.952	(1.650, 2.309)	<0.001
Maksan vajaatoiminta tai kirroosi	0.637	1.890	(1.264, 2.825)	0.009
Dementia	0.193	1.213	(1.012, 1.456)	0.080
Psykiatrinen häiriö	0.153	1.165	(1.037, 1.309)	0.030

Taulukko 22: Coxin regressiomallin tulokset, suolistonsisäinen verenvuoto -aineisto, Jiang ja kumppanit GWAS, PRS-CS-menetelmä, kategorinen PRS. US = uhkasuhde, LV = luottamusväli.

Liite B SNP-pohjaisen heritabiliteetin estimointi

Heritabiliteetti on määritelmällisesti se osuus jonkin ominaisuuden kokonaisvarianssista populaatiossa, joka on selitettävissä geneettisellä variaatiolla [35]. Heritabiliteetilla voidaan tarkoittaa joko laajaa heritabiliteettia (*broad-sense heritability*) H^2 tai kapeaa heritabiliteettia (*narrow-sense heritability*) h^2. Laajassa heritabiliteetissa käsitellään koko geneettistä varianssia populaatiossa ja kapeassa heritabiliteetissa vain additiivisten geneettisten arvojen varianssia populaatiossa. Kun puhutaan vain heritabiliteetista tarkoitetaan yleensä kapeaa heritabiliteettia. Heritabiliteetin rikaste (*heritability enrichment*) taas määritellään yleensä snippien osajoukon selittämän heritabiliteetin ja snippien osajoukon suhteellisen osuuden osamääränä [6].

Koko geneettinen varianssi σ_g^2 on

$$\begin{aligned}\sigma_g^2 &= Var(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}' Var(\mathbf{X})\, \boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{R}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'\boldsymbol{\beta} \\ &= \boldsymbol{\beta}'\mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}'\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{Q}'\mathbf{Q}\boldsymbol{\beta} = \widehat{\mathbf{w}}'\widehat{\mathbf{w}},\end{aligned}$$

josta $\Lambda^{\frac{1}{2}}\mathbf{U}' = \mathbf{Q}$ ja $\mathbf{w} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{e}$ määriteltiin aiemmin matala-asteista mallia käsittelevässä luvussa 3.1 [38]. Tämä suure lasketaan otanta-algoritmissa 1 jokaisella iteraatiolla snippien yhteisvaikutusten $\boldsymbol{\beta}$ sen hetkisten arvojen perusteella. Kun fenotyypin varianssin oletetaan olevan arvoltaan 1, on SNP-pohjainen heritabiliteetti $h^2_{SNP} = \sigma_g^2$. Tämä estimoidaan Gibbsin otannan posteriorikeskiarvona kun sisäänajojakso on poistettu.

SBayesRC estimoi myös annotaatiokohtaisen suureen θ_c, eli per-SNP-heritabiliteettirikasteen (*per-SNP heritability enrichment*) [38]. Kokonaisvarianssi

$$\sigma_c^2 = \sum_{j=1}^{m_c} \beta_{jc}^2$$

binääriselle annotaatiolle c on varianssi, jonka tähän annotaatioluokkaan kuuluvat m_c snippiä selittävät. Tällöin θ_c on

$$\theta_c = \left(\frac{\sigma_c^2}{m_c} \Big/ \frac{\sigma_g^2}{m}\right) = \left(\frac{\sigma_c^2}{\sigma_g^2} \Big/ \frac{m_c}{m}\right),$$

joka vastaa aiemmin esitettyä määritelmää. Kvantitaavisessa annotaatiossa θ_c lasketaan sellaisen regressiomallin kulmakertoimena, jossa selitettävänä muuttujana on β_{jc}^2 ja selittävänä muuttujana on annotaatioarvo A_{jc}:

$$E[\boldsymbol{\beta}_c^2] = \mathbf{1}_{\mu_c} + \mathbf{A}_c\omega_c.$$

Nyt $\theta_c = 1 + \omega_c$. Merkinnällä $\boldsymbol{\beta}_c^2$ tarkoitetaan alkiokohtaista neliöintiä. Kuten SNP-pohjaisen heritabiliteetin tapauksessa, θ_c lasketaan jokaisella Gibbsin otannan iteraatiolla ja lopullinen estimaatti on otannan posteriorikeskiarvo kun sisäänajojakso on poistettu.

Liite C Polygeenisten riskisummien laskeminen tutkimusyksilöille

Tässä luvussa käydään tarkemmin läpi, kuinka polygeeniset riskisummat laskettiin FinnGen-yksilöille. SBayesRC-menetelmän tapauksessa tähän käytettiin R-ohjelmaa ja sbayesrc-paketin funktioita [38], PRS-CS-menetelmän tapauksessa myös FinnGenin virtuaaliympäristön sisäänrakennettuja työkaluja. Prosessi, jossa polygeeniset riskisummat yleisesti lasketaan SBayesRC-menetelmällä käyttäen R-ohjelmaa ja sbayesrc-pakettia [38] on esitetty taulukossa 23. Prosessin keskeiset vaiheet ovat GWAS-tunnuslukujen muokkaus COJO-formaatin mukaiseksi, GWAS-tunnuslukujen yhteenmukaistaminen LD-referenssin kanssa funktioilla tidy ja impute, snippien yhteisvaikutusten estimointi menetelmän pääfunktiolla sbayesrc ja lopulta polygeenisten riskisummien laskeminen tutkimusyksilöille funktiolla prs. Seuraavissa luvuissa käsitellään yksityiskohtaisemmin, kuinka nämä askeleet toteutettiin tutkielman sovelluksen tapauksessa.

C.1 GWAS-tunnuslukujen muokkaus

SBayesRC-menetelmän pääfunktio sbayesrc() – joka estimoi snippien yhteisvaikutukset edellä esitetyllä Gibbsin otannalla – ottaa syötteenä GWAS-tunnusluvut tekstitiedostona COJO-formaatissa joka sisältää seuraavat muuttujat: SNP (snippitunniste), A1 (vaikutusalleeli), A2 (vaihtoehtoinen alleeli), freq (vaikutusalleelin suhteellinen osuus populaatiossa), b (marginaalivaikutus), se (keskivirhe), p (p-arvo) ja N (per-SNP otoskoko). Zhoun ja kumppanien [39] sekä Jiangin ja kumppanien [19] GWAS-tunnusluvuista valittiin vastaavat muuttujat, ja näiden nimet muutettiin COJO-formaattia vastaaviksi. Muuttujien nimet alkuperäisissä GWAS-tunnuslukutiedostoissa on esitetty taulukossa 24. Zhoun ja kumppanien GWAS-tunnusluvuissa ei ollut per-SNP otoskokoa (kuinka monelta yksilöltä kyseisen snipin arvo on mitattu eli snippikohtainen otoskoko), molempien GWAS-tunnuslukujen tapauksessa jokaiselle snipille asetettiin tämän muuttujan arvoksi GWAS-katalogissa [15] ilmoitettu kokonaisotoskoko (400 813 yksilöä kallonsisäisen verenvuodon ja 406 294 yksilöä suolistonsisäisen verenvuoton GWAS-tunnusluvuille).

Seuraavaksi GWAS-tunnuslukuja muokattiin jotta niissä olisi samat snipit kuin LD-referenssissä, tähän käytettiin kahta SBayesRC-paketin apufunktiota [38]. Ensinnäkin apufunktiota tidy, joka luo GWAS-tunnuslukutiedostosta siistityn version. Se ensinnäkin pitää mukana vain validit snipit: muuttujissa N, b, se ja freq ei saa olla puuttuvia arvoja ja arvojen täytyy olla äärellisiä. Lisäksi muuttujan p arvojen täytyy olla välillä [0, 1]. Funktio pudottaa GWAS-tunnusluvuista ne snipit joita ei löydy LD-referenssitiedostosta. Se suorittaa lisäksi seuraavia laadunvalvontaoperaatioita snipeille:

	Menetelmän askeleet
1	Tiedostot ja resurssit: GWAS-tunnusluvut tekstitiedostona GWAS.txt, LD-referenssi-tiedostokokonaisuus kansiossa LD, annotaatiodata tekstitiedostona annotaatio.txt ja tutkimusyksilöiden genotyyppidata tiedostokokonaisuutena kansiossa genotype.
2	Valitse tiedostosta GWAS.txt tarvittavat muuttujat snippitunnus, vaikutusalleeli, vaihtoehtoinen alleeli, vaikutusalleelin suhteellinen osuus, marginaalivaikutus, keskivirhe, p-arvo ja per-SNP-otoskoko.
3	Jos tiedostosta GWAS.txt puuttuu per-SNP-otoskoko, lisää tiedostoon tämä muuttuja ja aseta sen jokaiseksi arvoksi GWAS-tutkimuksen kokonaisotoskoko.
4	Nimeä tiedoston GWAS.txt sarakkeet uudelleen COJO-formaatin mukaisesti: SNP, A1, A2, freq, b, se, p ja N.
5	Aja sbayesrc-paketin funktio tidy, tämä ottaa syötteenä tiedoston GWAS.txt (parametrille mafile annetaan tiedostopolku tiedostoon GWAS.txt) ja LD-kansion (parametrille LDdir annetaan tiedostopolku kansioon LD). Funktiolle annetaan myös tiedostopolku parametrille output. Funktio tuottaa uuden GWAS-tunnuslukutiedoston, johon on jätetty tiedostosta GWAS.txt vain snipit jotka löytyvät LD-referenssistä ja jotka ovat läpäisseet tietyt laadunvalvontaoperaatiot. Olkoon tämä uusi tiedosto GWAS_tidy.txt.
6	Tämä askel suoritetaan jos tiedostossa GWAS_tidy.txt ei ole kaikkia LD-referenssin snippejä. Aja sbayesrc-paketin funktio impute, tämä ottaa syötteenä tiedoston GWAS_tidy.txt (parametrille mafile annetaan tiedostopolku tiedostoon GWAS_tidy.txt) ja LD-kansion (parametrille LDdir annetaan tiedostopolku kansioon LD). Funktiolle annetaan myös tiedostopolku parametrille output. Funktio tuottaa uuden GWAS-tunnuslukutiedoston, johon on imputoitu LD-referenssin snipit jotka puuttuvat tiedostosta GWAS_tidy.txt Olkoon tämä uusi tiedosto GWAS_tidyimputed.txt.
7	Aja sbayesrc-paketin pääfunktio sbayesrc, tämä ottaa syötteenä tiedoston GWAS_tidy.txt (parametrille mafile annetaan tiedostopolku tiedostoon GWAS_tidy_impute.txt), LD-kansion (parametrille LDdir annetaan tiedostopolku kansioon LD) ja annotaatiotiedoston (parametrille annot annetaan tiedostopolku tiedostoon annotaatio.txt). Funktiolle annetaan myös tiedostopolku outputille parametrille outPrefix. Funktio estimoi snippien yhteisvaikutukset. Olkoon tämä uusi tiedosto SNP_weights.txt.
8	Jos tiedoston SNP_weights.txt snippitunnukset ovat eri formaatissa kuin snippitunnukset tutkimusyksilöiden genotyyppidatassa, tulee snippitunnukset tiedostossa SNP_weights.txt muuntaa tähän formaattiin.
9	Aja sbayesrc-paketin funktio prs, tämä ottaa syötteenä tiedoston SNP_weights.txt (parametrille weights annetaan tiedostopolku tiedostoon SNP_weights.txt) ja tilastoyksilöiden genotyyppidatan (parametrille genoPrefix annetaan tiedostopolku kansioon genotype). Funktiolle annetaan myös output-tiedostopolku parametrille out. Funktio laskee polygeeniset riskisummat genotyyppidatan yksilöille ja antaa ne tekstitiedostona, olkoon tämä tiedosto PRS.txt.

Taulukko 23: Polygeenisten riskisummien laskemisen askeleet SBayesRC-menetelmällä R-ohjelmalla.

COJO	Zhou ym. GWAS	Jiang ym. GWAS
SNP	rsid	hm_rsid
A1	effect_allele	hm_effect_allele
A2	other_allele	hm_other_allele
freq	effect_allele_frequency	hm_effect_allele_frequency
b	beta	hm_beta
se	standard_error	standard_error
p	p_value	p_value
N	-	n

Taulukko 24: Käytettävien muuttujien nimet COJO formaatissa sekä Zhoun ja Jiangin GWAS-tunnusluvuissa.

- Tunnuslukujen alleelien tulee olla yhteensopivat LD-referenssin alleelien kanssa.
- Ero alleelien suhteellisissa osuuksissa tunnuslukujen ja LD-referenssin välillä saa olla enintään annetun kynnysarvon verran, oletusarvona kynnysarvo on 0.2.
- Per-SNP otoskoko saa poiketa keskiarvosta enintään x kertaa keskihajonnan verran, oletusarvona x on 3.
- Kahdella eri tavalla tunnusluvuista estimoidun genotyyppivarianssin suhde saa poiketa arvosta 1 enintään annetun kynnysarvon verran, oletusarvona kynnysarvo on 0.5.

Kaikkien GWAS-tunnuslukujen kohdalla tidy-funktion ajamisen jälkeen tunnusluvuista puuttui jonkin verran LD-referenssistä löytyviä snippejä. Nämä puuttuvat snipit imputoitiin tunnuslukuihin apufunktiolla impute, jonka jälkeen tunnusluvuissa oli samat snipit kuin LD-referenssissä. Funktio impute myös tarvittaessa kääntää ympäri (*flip*) alleelit sellaisissa GWAS-tunnuslukujen snipeissä, missä LD-referenssin vastaavassa snipissä on muuten samat alleelit mutta vaikutus- ja vaihtoehtoinen alleeli ovat toisin päin, esimerkiksi

GWAS-tunnuslukujen snippi	**LD-referenssin snippi**
vaikutusalleeli: A	vaikutusalleeli: T
vaihtoehtoinen alleeli: T	vaihtoehtoinen alleeli: A.

Tällaisessa tilanteessa impute vaihtaa GWAS-tunnuslukujen snipistä vaikutusalleelin vaihtoehtoiseksi alleeliksi ja vaihtoehtoisen alleelin vaikutusalleeliksi. Luonnollisesti marginaalivaikutusta ja vaikutusalleelin suhteellista osuutta täytyy myös muuttaa, impute muuttaa marginaalivaikutuksen vastaluvukseen ja suhteellisen osuuden p luvuksi $1-p$. Kaikki LD-referenssin snipit löytyvät BaselineLD v2.2 annotaatiotiedostosta, koska Zheng ja kumppanit valitsivat tutkimuksessaan käyttämät (ja LD-referenssissä olevat) 7356518 snippiä yhteensovittamalla UK biopankin, BaselineLD v2.2 -mallin ja LifeLines-kohortin snipit keskenään [38].

Kun tidy-funktiota käytettiin Zhoun ja kumppanien kallonsisäisen verenvuodon GWAS-tunnusluvuille [39], oli tunnusluvuissa 27 982 788 validia snippiä joista 7 344 489 löytyi LD-referenssin snipeistä. Näistä snipeistä 7 344 465 läpäisi laadunvalvontaoperaatiot. Tämän jälkeen impute-funktiolla imputoitiin 12 053 snippiä, eli noin 0.16% lopullisista snipeistä. Funktio myös käänsi 12 416 snippiä ympäri. Zhoun ja kumppanien suolistonsisäisen verenvuodon GWAS-tunnuslukujen tapauksessa tunnusluvuissa oli 28 030 926 validia snippiä joista 7 344 489 löytyi LD-referenssin snipeistä ja 7 344 465 läpäisi laadunvalvontaoperaatiot, selvästi samat snipit kuin kallonsisäisen verenvuodon tunnuslukujen tapauksessa. Imputointi tapahtui samoin kuin kallonsisäisen verenvuodon tunnusluvuille, 12 053 snippiä imputoitiin ja 12 416 snippiä käännettiin ympäri.

Kun tidy-funktiota käytettiin Jiangin ja kumppanien kallonsisäisen verenvuodon GWAS-tunnusluvuille, oli näissä 11 796 992 validia snippiä joista 7 288 423 löytyi LD-referenssin snipeistä. Näistä snipeistä 7 115 775 läpäisi laadunvalvontaoperaatiot. Tämän jälkeen impute-funktiolla imputoitiin 240 743 snippiä, eli noin 3.27% lopullisista snipeistä. Funktio myös käänsi 11 791 snippiä ympäri. Jiangin ja kumppanien hematemesis-piirteen GWAS-tunnuslukujen tapauksessa tunnusluvuissa oli 11 796 992 validia snippiä joista 7 288 423 löytyi LD-referenssin snipeistä ja 7 115 775 läpäisi laadunvalvontaoperaatiot, myös tässä tapauksessa selvästi samat snipit kuin kallonsisäisen verenvuodon tunnuslukujen tapauksessa. Imputointi tapahtui samoin kuin kallonsisäisen verenvuodon tunnusluvuille, 240 743 snippiä imputoitiin ja 11 791 snippiä käännettiin ympäri. Näiden toimien jälkeen kaikissa GWAS-tunnusluvuissa oli samat 7 356 518 snippiä kuin LD-referenssissä.

C.2 Yhteisvaikutusten estimointi

Tämän jälkeen snipeille estimoitiin yhteisvaikutukset SBayesRC-paketin pääfunktiolla sbayesrc, kaikille GWAS-tunnusluvuille erikseen. Kuten aiemmin jo todettiin, ennen SBayesRC-menetelmän käyttöä tulee määrittää kuinka suuri osuus ρ jokaisen LD-lohkon $\mathbf{R}_k$ alkuperäisestä varianssista halutaan säilyttää dimensionvähennyksessä. Oletusasetuksilla sbayesrc-funktio määrittää parametrin ρ arvon annettuihin GWAS-tunnuslukuihin perustuvalla pseudovalidoinnilla [38]. Tämä pseudovalidointiaskel on kuitenkin mahdollista poistaa käytöstä, jolloin parametri ρ annetaan funktiolle suoraan syöteparametrina. Tässä tutkielmassa parametrille ρ annettiin kaikissa tapauksissa suoraan arvo 0.995. Tähän päädyttiin ensinnäkin siksi, että käytettäessä pseudovalidointia oletusasetuksilla parametrin arvoksi valikoitui usein matalin mahdollinen arvo 0.9, jolloin funktion ajo pysähtyy virheilmoitukseen. Funktion ajo on ohjelmoitu pysähtymään siksi, että kun päädytään testattavien arvojen reuna-arvoon, on hyvin mahdollista että optimaallinen arvo löytyy testattavien arvojen vaihteluvälin ulkopuolelta, jolloin voi olla viisasta laajentaa vaihteluväliä. Toiseksi Zheng ja kumppanit suorittivat tutkimuksessaan kalibrointianalyysiä mm. tälle parametrille käyttäen simuloitua dataa, analyysin tulosten visualisointi näkyy tutkimuksen lisäosan kuvaajassa 2 (*Supplementary Figure 2*) [38]. Kaikissa tilanteissa ennustetarkkuus kasvaa parametrin ρ arvon kasvaessa arvoon 0.995 asti, jonka jälkeen ennustetarkkuus ei enää juuri kasva parametria ρ kasvattamalla. Näin ollen valinta $\rho = 0.995$ oli perusteltu.

Kuten aiemmin todettiin, SBayesRC-menetelmän tilastollinen malli on Bayes-malli ja estimointi tapahtuu MCMC-otannalla. MCMC-otannan aloituspiste ei välttämättä ole lähellä parametriavaruuden sitä osaa missä on suurin osa sen todennäköisyysmassasta, ja vaatii jonkin verran iteraatioita ennen kuin MCMC-otanta hakeutuu tälle alueelle. Siksi MCMC-otoksen alusta poistetaan usein ns. sisäänajojakso (*burn-in period*). Näin on myös tässä tapauksessa, sbayesrc-funktio ajaa oletusasetuksilla 3000 iteraatiota, joista ensimmäiset 1000 poistetaan sisäänajojaksona. Tässä tutkielmassa sbayesrc-funktio ajettiin 10000:lla iteraatiolla, joista ensimmäiset 4000 poistettiin sisäänajojaksona.

C.3 PRS-CS menetetelmä verrokkina SBayesRC-menetelmälle

SBayesRC on melko uusi polygeenisen prediktion menetelmä, ja tutkielmassa haluttiin vertailun vuoksi käyttää myös toista menetelmää. PRS-CS on vanhempi menetelmä, joka kuten SBayesRC pohjautuu Bayes-malliin, mutta ei käytä funktionaalisia annotaatioita [14]. Menetelmä on laajalti käytetty ja myös FinnGenin virtuaaliympäristössä oli työkalut snippien yhteisvaikutusten estimointiin sitä käyttäen. Nämä yhteisvaikutukset estimoitiin PRS-CS-menetelmällä valituille GWAS-tunnusluvuille, jotta kahdella eri menetelmällä laskettujen polygeenisten riskisummien toimivuutta voitaisiin vertailla keskenään.

Käytetyt GWAS-tunnusluvut olivat samat kuin mitä käytettiin SBayesRC-menetelmässä. Menetelmälle syötetettiin muuttujat snippitunnus, kromosomitunnus, emäsparin fyysinen sijainti, vaikutusalleeli, vaihtoehtoinen alleeli, marginaalivaikutus sekä p-arvo. Muuttujien nimet GWAS-tunnusluvuissa on esitetty taulukossa 25. Muita menetelmälle annettuja tietoja olivat tieto syntyperästä (eurooppalainen syntyperä), käytetty genomireferenssi (hg38) sekä kokonaisotoskoko. Zhoun ja kumppanien GWAS-tunnusluvuissa kokonaisotoskoko oli 400 813 kallonsisäisessä verenvuodossa ja 406 294 suolistonsisäisessä verenvuodossa, Jiangin ja kumppanien molemmissa GWAS-tunnusluvuissa kokonaisotoskoko oli 456 348. Kaikissa neljässä tapauksessa yhteisvaikutusten estimoinnissa käytettiin FinnGenin eurooppalaisen syntyperän LD-referenssiä jossa oli 1 120 696 snippiä.

Muuttuja	**Zhou ym. GWAS**	**Jiang ym. GWAS**
Snippitunnus	rsid	hm_rsid
Kromosomitunnus	chromosome	chromosome
Fyysinen sijainti	base_pair_location	base_pair_location
Vaikutusalleeli	effect_allele	hm_effect_allele
Vaihtoehtoinen alleeli	other_allele	hm_other_allele
b	beta	hm_beta
p	p_value	p_value

Taulukko 25: PRS-CS-menetelmässä käytettyjen muuttujien nimet Zhoun ja Jiangin GWAS-tunnusluvuissa.

C.4 Polygeenisten riskisummien laskeminen

Viimeiseksi kaikille FinnGen-yksilöille laskettiin GWAS-tunnuslukuja vastaavat polygeeniset riskisummat käyttämällä estimoituja yhteisvaikutuksia. Tähän käytettiin sbayesrc-paketin funktiota prs, jolle annetaan syötteenä snippipainot eli snippien yhteisvaikutukset sekä yksilötason genotyyppidata [38]. Funktio ajettiin 8 kertaa, joka kerralla syötteenä olivat yhdet estimoidut SNP-yhteisvaiktutukset sekä FinnGen-yksilöiden genotyyppidata. Jokainen saatu polygeeninen riskisummamuuttuja skaalattiin vähentämällä keskiarvo ja jakamalla keskihajonnalla. Tämän jälkeen jokaiselle FinnGen-yksilölle oli olemassa 8 eri polygeenista riskisummaa: kahdelle eri vastetapahtumalle (kallon- ja suolistonsisäinen verenvuoto), käyttäen GWAS-tunnuslukuja kahdesta vaihtoehtoisesta tutkimuksesta (Zhou ja kumppanit sekä Jiang ja kumppanit) ja käyttäen kahta eri menetelmää (SBayesRC ja PRS-CS).

Tämän lisäksi jokaisesta polygeenisesta riskisummasta muodostettiin vielä viisitasoinen kategorinen versio. Jokaiselle riskisummamuuttujalle määritettiin seuraavat kvantiilivälit: 0–0.025, 0.025–0.2, 0.2–0.8, 0.8–0.975 ja 0.975–1. Kategoriset riskisummamuuttujat kertoivat mihin näistä kvantiiliväleistä FinnGen-yksilöt kuuluivat. Nämä kategoriset muuttujat mukaan laskettuna lopulta käytössä oli 16 erilaista polygeenista riskisummamuuttujaa analyysejä varten.