Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Tuuli Jürgenson

Retrospektiivsete ja prospektiivsete andmete kombineerimine ülegenoomsetes seoseuuringutes

Matemaatika ja statistika õppekava

Matemaatilise statistika eriala

Magistritöö (30 EAP)

Juhendajad: Anastassia Kolde, MSc

Prof. Krista Fischer, PhD

Prof. Reedik Mägi, PhD

Tartu 2021

Retrospektiivsete ja prospektiivsete andmete kombineerimine ülegenoomsetes seoseuuringutes

Magistritöö

Tuuli Jürgenson

Lühikokkuvõte: Magistritöö eesmärk on leida jälgimiseelsete (retrospektiivsete) ja jälgimisaegsete (prospektiivsete) haigusjuhtude analüüsimiseks sobiv meetod, mis oleks rakendatav suuremahulistes geneetilistes seoseuuringutes. Jälgimiseelseteks juhtudeks nimetatakse neid inimesi, kes on uuritava haiguse saanud enne uuringuga liitumist, jälgimisaegseteks juhtudeks aga neid, kes esimest korda haigestuvad uuritavasse haigusesse pärast uuringuga liitumist. Huvi pakub see, kas jälgimiseelseid ja jälgimisaegseid haigusjuhtusid on parem analüüsida eraldi, kasutades vastavalt kas binaarse uuritava tunnuse mudelit või Coxi võrdeliste riskide mudelit, ning leitud hinnangud seejärel kombineerida, või analüüsida neid andmeid koos, tegemata vahet jälgimiseelsetel ja jälgimisaegsetel juhtudel.

Töö teoreetilises osas antakse ülevaade kasutatavatest meetoditest: elukestusanalüüsiks mõeldud Coxi võrdeliste riskide mudelist ning kahest binaarse tunnuse modelleerimise meetodist: logistilisest ja täiend-log-log regressioonist. Meetodite võrdlemiseks viiakse läbi simulatsiooniuuring, mille tarvis kirjeldatakse esmalt, kuidas simuleerida Weibulli jaotusega võrdeliste riskide mudelile vastavaid elukestusandmeid. Simulatsioonide põhjal on erinevaid haigusjuhtusid kõige parem analüüsida koos, kasutades selleks täiend-log-log mudelit. Võrreldud meetodeid rakendatakse Tartu Ülikooli Eesti Geenivaramu andmestikul, uurimaks teist tüüpi diabeedi ja geenivariantide vahelisi seoseid.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: geneetilised assotsiatsiooniuuringud, elukestusanalüüs, metaanalüüs, regressioonanalüüs, simulatsioon

Combining retrospective and prospective data for genome-wide association studies

Master's thesis

Tuuli Jürgenson

Abstract: The aim of this master's thesis is to find a method for combined analysis of prevalent (retrospective) and incident (prospective) cases that could be used when conducting genome-wide association studies. We define prevalent cases as individuals who have the disease of interest before being recruited into a study, and incident cases as individuals who develop the disease only after study recruitment. We are interested in whether it is better to analyse prevalent and incident cases separately, using either a binary response model or a Cox proportional hazards model respectively, and then combine the results using meta-analysis, or to analyse these data together without making any distinction between prevalent and incident cases.

The theoretical part of the thesis gives an overview of the methods used, namely Cox proportional hazards model and two binary regression models, logistic and complementary log-log regression. To compare different methods a simulation study is conducted, before which the thesis shows how to simulate survival data from Weibull proportional hazards model. Based on the results of the simulations, the best method is to analyse prevalent and incident cases together using a complementary log-log model. Lastly, a practical analysis is carried out to study associations between type 2 diabetes and genetic variants using data from Estonian Genome Centre at the University of Tartu.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Keywords: genetic association studies, survival analysis, meta-analysis, regression analysis, simulation

Sisukord

Kasutatud lühendid

cloglog	täiend-log-log (*complementary log-log*)
DNA	desoksüribonukleiinhape (*deoxyribonucleic acid*)
EHR	elektroonilised terviseandmed (*electronic health records*)
GWAS	ülegenoomne seoseanalüüs (*genome-wide association study*)
HR	riskimäärade suhe (*hazard ratio*)
ICD	rahvusvaheline haiguste klassifikatsioon (*The International Classification of Diseases*)
MAF	harvema alleeli sagedus (*minor allele frequency*)
MR	martingaalijäägid (*martingale residuals*)
OR	šansside suhe (*odds ratio*)
RMSE	ruutjuur keskmisest ruutveast (*root mean square error*)
SNP	üksiknukleotiidne polümorfism (*single nucleotide polymorphism*)
T2D	teist tüüpi diabeet (*type 2 diabetes*)
TÜ EGV	Tartu Ülikooli Eesti Geenivaramu

Sissejuhatus

Magistritöö eesmärk on välja selgitada, kuidas kõige efektiivsemalt analüüsida seoseid (geneetilise) riskiteguri ja haiguse vahel, kui osal uuritavatest on vastav haigus diagnoositud enne uuringuga liitumist ja osal uuritavatest pärast liitumist. Töö on motiveeritud Tartu Ülikooli Eesti Geenivaramus (TÜ EGV, edaspidi ka geenivaramu) tehtavatest uuringutest. Geenivaramuga liitunud inimeste geeniandmed on seotud nende terviseandmetega – teada on geenidoonorite diagnoosid nii geenivaramuga liitumisele eelnenud kui järgnenud ajast. Jälgimiseelseteks juhtudeks (*prevalent cases*) nimetame neid geenidoonoreid, kes on meile huvipakkuva haiguse saanud enne geenivaramuga liitumist. Jälgimisaegseteks juhtudeks (*incident cases*) aga neid geenidoonoreid, kes on haigestunud pärast geenivaramuga liitumist.

Et jälgimisaegsete haigusjuhtude puhul on teada ka diagnoosi saamise aeg, siis rakendatakse nende juhtude analüüsimisel enamasti Coxi võrdeliste riskide mudelit. Jälgimiseelsete juhtude puhul kasutatakse tavaliselt binaarse tunnuse analüüsimiseks mõeldud meetodeid, näiteks logistilist regressiooni. Töö eesmärk on leida lihtne meetod, mille abil saab neid haigusjuhtusid analüüsida koos ja mis oleks kasutatav ka suurte ülegenoomsete uuringute puhul, kus huvi pakuvad seosed haiguse ja miljonite geenivariantide vahel.

Magistritöös võrreldakse erinevaid meetodeid jälgimiseelsete ja jälgimisaegsete haigusjuhtude ning geenivariantide vaheliste seoste analüüsimiseks: logistilist ja täiend-log-log mudelit binaarsele tunnusele ning Coxi regressiooni elukestusandmetele. Töös uuritakse, kas jälgimiseelseid ja jälgimisaegseid juhtusid on parem analüüsida eraldi, kasutades vastavalt binaarsete tunnuste analüüsimiseks mõeldud mudelit ja Coxi mudelit, ning saadud hinnangud seejärel metaanalüüsi kasutades kombineerida, või analüüsida kõiki haigusjuhtusid koos, tegemata vahet, kas haigus saadi enne või pärast geenivaramuga liitumist.

Meetodite võrdlemiseks viiakse läbi simulatsiooniuuring; enne seda kirjeldatakse algoritmi sobivate elukestusandmete genereerimiseks. Uuritud meetodeid rakendatakse geenivaramu andmetel, selleks et analüüsida seoseid geenivariantide ja teist tüüpi diabeedi vahel.

Töö esimeses peatükis antakse lühiülevaade statistilise geneetika põhimõistest ning jälgimiseelsete ja jälgimisaegsete juhtude analüüsi eripäradest. Teises peatükis kirjeldatakse kasutatavaid analüüsimeetodeid. Põhjalikumalt tutvustatakse elukestusanalüüsi ja selle tegemiseks rakendatavaid Coxi ning Weibulli jaotusega võrdeliste riskide mudeleid. Samuti kirjutatakse kahest binaarsete tunnuste uurimise meetodist: logistilisest ja täiend-log-log regressioonist. Kolmandas peatükis kirjeldatakse, kuidas simuleerida võrdeliste riskide mudelile vastavaid haigestumisandmeid Weibulli jaotusest, tutvustatakse simulatsiooniplaani ja simulatsiooniuuringust saadud tulemusi. Neljandas peatükis antakse ülevaade geenivaramu andmetest ja nende põhjal tehtud analüüsi tulemustest.

Andmete simuleerimiseks ja analüüsimiseks kasutati statistikatarkvara R 3.6.1 (R Core Team, 2019) ning kõik arvutused tehti Tartu Ülikooli teadusarvutuste keskuse arvutusklastris (Teadusarvutuste keskus, www.hpc.ut.ee).

1 Ülegenoomne seoseuuring

1.1 Geneetilised variandid ja seoseuuringud

Kõik rakud sisaldavad geneetilist materjali, millest valdav osa asub raku tuumas ja on organiseerunud kromosoomidesse, mille moodustavad üks katkematu DNA-kaksikahel ja sellega seotud valgud. DNA on polümeer, mis kannab rakkudes edasi pärilikku informatsiooni ja koosneb kahest omavahel ühendatud nukleotiidide ahelast. Täielikku DNA järjestust nimetatakse genoomiks. Inimeste genoom on diploidne, mis tähendab, et iga kromosoom esineb kahes koopias.

Neid DNA osi, mis erinevate inimeste vahel varieeruvad, nimetatakse geneetilisteks variantideks. Üksiknukleotiidsed polümorfismid (SNP, *single nucleotide polymorphism*) on DNA järjestuse variatsioonid, mis on tekkinud ühe nukleotiidi asendumisel teisega. SNP-d on kõige sagedasemateks variantideks inimese genoomis. SNP-del on enamasti kaks alleeli ehk kaks erinevat võimalust, millised neljast nukleotiidist DNA-ahelas sellel kohal paikneda saavad. SNP esinemissagedust kirjeldatakse harvema alleeli sageduse (MAF, *minor allele frequency*) kaudu. MAF on konkreetse SNP vähem esineva alleeli suhteline sagedus sellesama SNP kõigi ülejäänud alleelide suhtes populatsioonis.

Selleks, et välja selgitada, millised geneetilised variandid on seotud mingi huvipakkuva tunnuse või haigusega, viiakse läbi geneetilisi seoseuuringuid. Enamasti jagatakse geneetilise seoseuuringu tegemiseks geenidoonorid huvipakkuva haiguse põhjal juhtudeks ja kontrollideks. Kõige enam uuritavateks geenivariantideks ongi just SNP-d. Huvi pakub see, kas geneetilise variandi üks alleelidest esineb juhtude seas sagedamini kui kontrollide seas, sest see viitab selle variandi seosele vastava fenotüübiga. Enamasti SNP-d ise haigusi ei põhjusta, vaid annavad pigem aimu, millised kromosoomid või genoomipiirkonnad võivad uuritava haigusega seotud olla. Uuringuid, kus huvi pakuvad väga paljud geenivariandid üle kogu genoomi, nimetatakse ülegenoomseteks seoseuuringuteks (GWAS, *genome-wide association study*).

Tavaliselt eeldatakse geneetiliste seoseuuringute tegemisel, et SNP mõju on aditiivne (Hayes, 2013). See tähendab, et kui SNP-l on kaks alleeli A ja B ning kolm võimalikku genotüüpi

AA, AB ja BB, siis harilikult on SNP kodeeritud arvudega $0, 1, 2$, mis tähistavad efektialleeli, näiteks alleeli B sagedust. Seega vastavad arvud $0, 1$ ja 2 genotüüpidele AA, AB ja BB.

Geenidoonori genotüübi määramiseks kasutatakse kindlaid genotüpiseerimiskiipe, mille abil määratakse vaid väike osa inimese genoomis asuvatest SNP-dest. Ülejäänud SNP-d määratakse imputeerimise abil – kasutatakse referentsgenoomi ja teadmist, et kromosoomil lähestikku paiknevad SNP-d on omavahel korreleeritud. TÜ EGV imputeerimiseks kasutatav referentspaneel on koostatud rohkem kui 2000 geenidoonori põhjal, kellele on tehtud täisgenoomi sekveneerimine ehk on teada kogu nende DNA järjestus. Imputeerimise abil ei saa alati kindlalt öelda, milline genotüüp indiviidil on. Seega tavaliselt väljastavad imputeerimisprogrammid igale indiviidile iga imputeeritud SNP kohta alleelidoosi – see on reaalarv 0 ja 2 vahel, mis kirjeldab eeldatavat efektialleeli sagedust.

Kui uuritavaks tunnuseks on haiguse olemasolu, siis kasutatakse geneetilistes seoseuuringutes tavaliselt logistilist regressiooni, uuritav tunnus võib olla ka pidev (näiteks inimese pikkus või vererõhk), sel juhul kasutatakse lineaarset regressiooni. Geneetilisi seoseuuringuid saab rakendada ka elukestusandmetele, sel juhul on enamasti kasutusel Coxi võrdeliste riskide mudel.

1.2 Retrospektiivsed ja prospektiivsed andmed

Tavalised epidemioloogilised uuringud saab jagada retrospektiivseteks ja prospektiivseteks olenevalt sellest, millal andmeid koguma hakatakse. Retrospektiivsete ehk tagasivaatavate uuringute korral kasutatakse andmeid nende sündmuste kohta, mis on toimunud enne uuringu algust. Prospektiivsete ehk ettesuunatud uuringute korral hakatakse andmeid koguma nende sündmuste kohta, mis toimuvad pärast uuringu algust.

Suurte biopankade andmed on tihti seotud elektrooniliste terviseandmetega (EHR, *electronic health records*). Ka TÜ EGV geenidoonorite geeniandmetega on ühendatud terviseandmed erinevatest meditsiiniallikatest nagu Haigekassa, Tartu Ülikooli Kliinikum, Põhja-Eesti Regionaalhaigla, E-tervis ning surma- ja vähiregister. Seega saab ka biopankade andmed jagada retrospektiivseteks ja prospektiivseteks olenevalt sellest, kas inimene sai huvipakkuva haiguse

enne või pärast geenidoonoriks hakkamist. Sellisel juhul loetakse uuringu alguseks iga indiviidi puhul tema geenivaramuga liitumise kuupäeva.

Retrospektiivsete ehk jälgimiseelsete juhtude uurimisel tuleb arvestada sellega, et mõne haiguse puhul mõjutab haigestumine uuringusse kaasamise tõenäosust. Kui mingil haigusel on kiire ja kõrge suremus, siis on vähe tõenäoline, et selle haiguse saanud inimene liitub geenivaramuga. Samuti võib haigestumine tunduvalt halvendada inimese elukvaliteeti, mistõttu ta ei pruugi olla motiveeritud geenivaramuga liituma.

Sellise haigusega inimesed, kes siiski geenivaramuga liituvad, võivad seega olla eriline alamrühm, kes kõnealuse haiguse kergelt läbi põdesid. Kui nüüd analüüsida neid inimesi ja leida seos mõne geenivariandi ja vastava haiguse vahel, võib see geenivariant olla hoopis kaitsva mõjuga ning olla seotud selle haiguse kerge läbipõdemisega. Geenivariante, mis suurendavad nii haigestumise riski kui selle haigusega seotud letaalsust, on aga sellisel juhul keeruline tuvastada, ja nende efekti haigusele on oht alahinnata. Seda probleemi kirjeldatakse ka mõistega ellujääjate efekt (*survival bias*) või *prevalence-incidence bias*. (Oleckno, 2008)

Teine jälgimiseelsete juhtude eripära on see, et nende puhul ei ole sageli teada diagnoosi saamise aeg. Infot geenidoonorite varasemate haiguste kohta saadakse tihti nende enda täidetud küsimustikest ja kui inimene ise haigestumise kuupäeva ei mäleta, siis märgitakse see andmestikku lihtsalt kui enne liitumist saadud haigus. See on ka üks põhjuseid, miks ei saa jälgimiseelsete haigusjuhtude analüüsimiseks kasutada elukestusanalüüsi meetodeid.

Jälgimisaegsete juhtude puhul selliseid probleeme ei esine, nende analüüsimisel on korralikult esindatud ka fataalse ja raske haiguskuluga haiged. Samuti on jälgimisaegsete juhtude puhul alati teada ka diagnoosi saamise kuupäev ja seega on nende uurimisel standardiks elukestusanalüüsi meetodite rakendamine. Kui kaasata analüüsi ainult jälgimisaegsed juhud, võib aga oluliselt väheneda analüüsi võimsus, eriti just haruldaste haiguste korral, mille puhul on juhtude arv niigi väike ja jälgimiseelsete juhtude väljajätmine ei ole soovitav. Seega on kogu infot efektiivselt kasutav analüüs ikkagi selline, kuhu on kaasatud nii jälgimiseelsed kui jälgimisaegsed andmed.

Prospektiivsetes ehk edasivaatavates epidemioloogistes uuringutes on kestusandmete analüüsimiseks tavapärane kasutada Coxi võrdeliste riskide mudelit. Coxi mudeli hindamine on aga tunduvalt arvutusmahukam kui binaarsete tunnuste analüüsimiseks mõeldud mudelite kasutamine. Seega on suuremahuliste genotüüp-fenotüüp seoseuuringute (näiteks ülegenoomsete seoseuuringute) puhul siiani kasutatud peamiselt logistilist regressiooni, seda ka selliste andmete puhul, kus tegelikult oleks võimalik rakendada Coxi mudelit. Paljudes GWAS-i läbiviimiseks kasutatavates populaarsetes tarkvarapakettides nagu PLINK (Purcell *et al*., 2007), BOLT-LMM (Loh *et al*., 2015), SAIGE (Zhou *et al*., 2018) ja REGENIE (Mbatchou *et al*., 2020), ei ole Coxi mudelit üldse implementeeritud.

Arvutusmahu vähendamiseks Coxi võrdeliste riskide mudeli kasutamisel GWAS-i läbiviimiseks on välja pakutud ja kasutusele võetud erinevaid meetodeid, nagu näiteks kaheastmeline meetod, kus esmalt tehakse ülegenoomne analüüs logistilise regressiooniga, filtreeritakse välja need SNP-d, mille p-väärtus on väiksem mingist kindlaksmääratud piirist, ning hinnatakse siis Coxi mudelid ainult nendele SNP-dele (Staley *et al*., 2017), või martingaalijääkide kasutamine Coxi mudeli lähendamiseks (Joshi *et al*., 2016; Pilling *et al*., 2017; Timmers *et al*., 2020). Martingaalijääkide meetodit tutvustatakse ja katsetatakse ka selles töös.

Jälgimiseelsete juhtude eraldi analüüsimisel on tavapärane kasutada logistilist regressiooni, kus sõltuvaks tunnuseks on binaarne tunnus, mis vastab haiguse esinemisele. Logistilist ja Coxi regressiooni on palju uuritud ja omavahel võrreldud selliste uuringutüüpide puhul nagu läbilõikeline uuring (van der Net *et al*., 2008), kohortuuring (Callas *et al*., 1998), sobitatud juht-kontrolluuring (Leffondré *et al*., 2003) ja juht-kohortuuring (Staley *et al*., 2017). Logistilise regressiooniga hinnatakse šansside suhet (OR, *odds ratio*), Coxi võrdeliste riskide mudeliga aga riskimäärade suhet (HR, *hazard ratio*). Erinevates analüüsides on leitud, et Coxi mudelitest saadud hinnangud on täpsemad, Coxi regressioon on teatud juhtudel võimsam, Coxi mudeli hindamine võtab rohkem aega (Staley *et al*., 2017) ning et šansside suhte ja riskimäärade suhte omavaheline erinevus on seda suurem, mida pikem on jälgimisaeg ning mida suurem on haigestumusrisk või haiguse ja riskiteguri vahelise seose tugevus (Symons ja Moore, 2002).

Märgime veel, et jälgimiseelsete ja jälgimisaegsete juhtude koos analüüsimiseks on küll välja töötatud spetsiaalseid meetodeid, näiteks eksponentsiaalse kalde meetod (Maziarz *et al*., 2019), kuid siiani ei ole need skaleeritavad ülegenoomsete analüüside jaoks.

2 Analüüsimetoodika

2.1 Elukestusanalüüs

Elukestusanalüüsi (*survival analysis*) kasutatakse selliste andmete puhul, kus huvi pakub aeg algmomendist mingi kindla sündmuse toimumiseni. Algmomendiks võib olla näiteks inimese sünd või uuringusse kaasamise aeg, huvipakkuvaks sündmuseks näiteks haigestumine, mingi sümptomi esinemine või surm. Ajavahemikku algmomendist lõppsündmuseni nimetatakse elukestuseks (*survival time*).

Tsenseerimine

Sageli ei ole kõikide subjektide elukestus teada, sellisel juhul nimetatakse nende elukestust tsenseerituks. Tavaliselt on tsenseerimine tingitud sellest, et uuritaval ei ole sündmus enne katse lõppu või analüüsi tegemist toimunud. Inimese elukestus tsenseeritakse ka siis, kui inimene sureb (kui surm ei ole huvipakkuvaks sündmuseks) või lahkub uuringust enne selle lõppemist. Tsenseerimise korral on peamiselt tegemist paremalt poolt tsenseerimisega, mis tähendab seda, et uuritav sündmus ei toimunud enne uuringu lõppu.

Tähtsaks eelduseks on see, et tsenseerimine oleks mitteinformatiivne ehk toimuks huvipakkuvast sündmusest sõltumatult. Paremalt tsenseeritud andmete korral tähendab see, et subjekti tsenseerimine mingil ajahetkel ei anna sündmuse toimumise aja kohta muud infot peale selle, et tsenseerimisaeg on väiksem sündmuse toimumise ajast. Haigestumisandmete uurimisel ei tohi seega inimese tsenseerimine oleneda tema tervislikust seisundist – kui näiteks kontrollgruppi kuuluvad inimesed lahkuvad uuringust seetõttu, et nad jäid väga haigeks, ja ravigruppi kuuluvad inimesed seetõttu, et nad tunnevad end juba tervelt, siis on tegemist informatiivse tsenseerimisega – ning kõiki klassikalisi tsenseeritud andmete uurimiseks mõeldud analüüsimeetodeid rakendada ei saa (Kalbfleisch ja Prentice, 2002: 13).

Selles töös kasutatavate geenivaramu andmete puhul on tegemist mitteinformatiivse paremalt tsenseerimisega – teada on viimane ajahetk, mil subjektil ei olnud huvipakkuvat sündmust toi-

munud – ning tsenseerimine on tingitud sellest, et geenidoonor ei ole analüüsi tegemise ajaks haigust saanud. Sel juhul võetakse tsenseerimisajaks kuupäev, mil toimus viimane ühendamine (*linking*) diagnoosiandmetega. Väike osa tsenseerimistest toimub aga ka seetõttu, et geenidoonor sureb enne huvipakkuva haiguse saamist. Kuigi me sel juhul teame, et inimene ei saa kunagi huvipakkuvat haigust, ja seega ei ole tegelikult tegemist mitteinformatiivse tsenseerimisega, siis eeldusel, et suremus ei mõjuta seost riskiteguri ja haiguse vahel, on Coxi võrdeliste riskide mudeli parameetreid ka sel juhul võimalik õigesti hinnata (Therneau *et al*., 2021).

Ajaskaala valik

Kestusandmete analüüsimisel on oluline valida sobiv algmoment ja ajaskaala. Sageli on algmomendi valikuks kas inimese sünniaeg või uuringuga liitumise aeg.

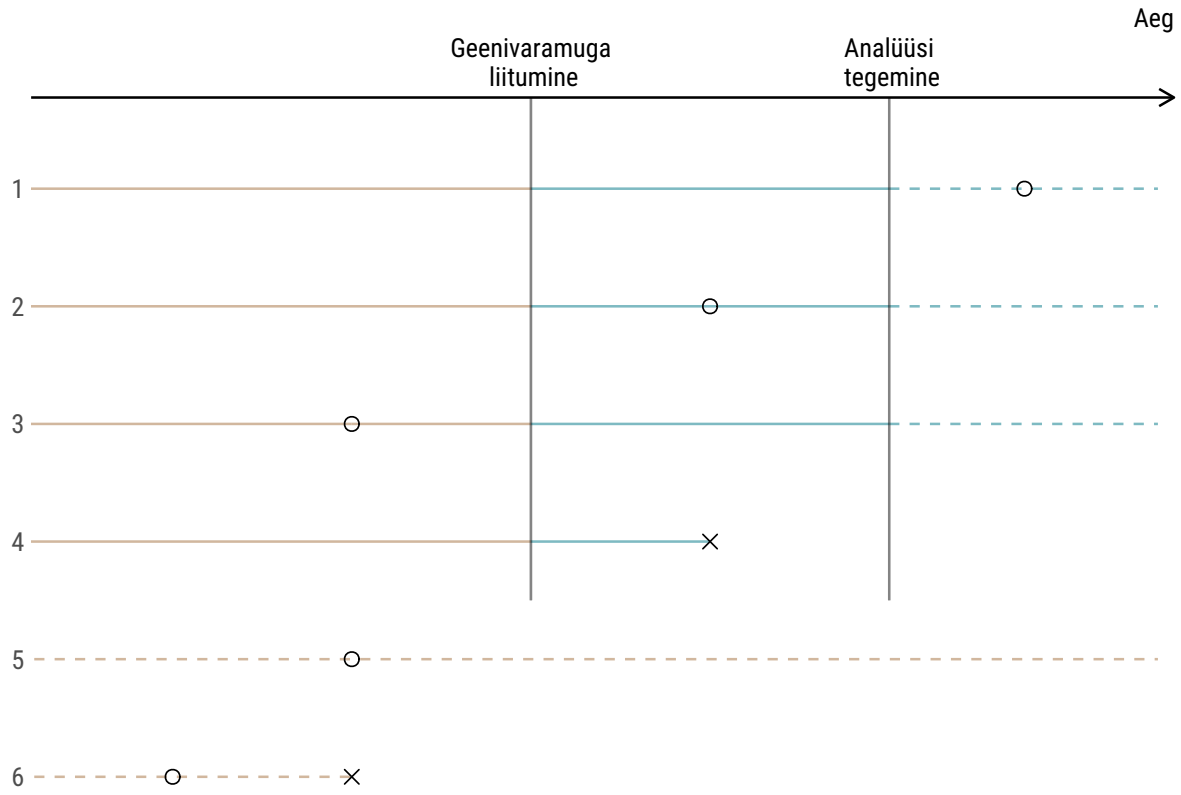
Kliiniliste uuringute puhul hakatakse inimesi tihti jälgima alates näiteks haigestumisest või mingi ravi alustamisest ja huvi pakub näiteks inimese tervenemine, surm või sümptomite taastekе. Sel juhul on loomulikuks algmomendiks uuringusse kaasamise aeg ja ajaskaalaks uuringus oldud aeg (*time-on-study*). Siis võrreldakse uuritavat teiste uuritavatega, kes on uuringus olnud sama kaua. Inimese vanuse arvesse võtmiseks lisatakse sel juhul kovariaadina mudelisse ka uuringuga liitumise vanus. (Thiébaut ja Bénichou, 2004)

Epidemioloogiliste uuringute puhul, kus huvi pakub aeg mingi haiguse tekkeni, kasutatakse ajaskaalana enamasti inimese vanust ja algmomendiks on seega sünniaeg (Thiébaut ja Bénichou, 2004). Sellisel juhul ei ole uuringuga liitumise aeg seotud huvipakkuva haigusega ja seega ei ole aeg uuringuga liitumisest haigestumiseni väga oluline. Arvestades, et enamiku haiguste puhul on vanus olulisim riskitegur, siis on tähtis just inimese vanuse arvesse võtmine, ja kui ajaskaala on inimese vanus, siis võrreldaksegi uuritavaid teiste samaealistega. Arvestada tuleb aga vasakult tõkestatusega (*left truncation*), see tähendab, et uuringusse saavad jõuda vaid need inimesed, kes on elus uuringu alguses. Vasakult tõkestatuse mittearvestamine võib viia nihkega hinnanguteni (Klein ja Moeschberger, 2003).

Samal ajal on näidatud, et kui epidemioloogiliste uuringute puhul on huvipakkuv argument-

tunnus ajas muutumatu ega ole seotud liitumisaegse vanusega, siis annab nende erinevate ajaskaalade kasutamine sarnaseid tulemusi (Korn *et al*., 1997; Ingram *et al*., 1997; Canchola *et al*., 2003; Thiébaut ja Bénichou, 2004). Selles töös kasutame elukestusandmete analüüsimisel ajaskaalana aega alates geenivaramuga liitumisest ja kovariaadina lisame vanuse liitumise ajal.

Eespool tutvustatud mõisteid on illustreeritud joonisel 1. Sellel on numbritega 1-6 tähistatud erinevad vaatlusalused subjektid, kes on kas paremalt tsenseeritud, vasakult tõkestatud või jälgimiseelsed või jälgimisaegsed juhud.



Joonis 1. Subjektid 1-4 on geenivaramuga liitunud. Subjektid 1 ja 4 on paremalt tsenseeritud, subjekt 2 on jälgimisaegne juht ja subjekt 3 on jälgimiseelne juht. Subjektid 5 ja 6 on vasakult tõkestatud.

Sellel joonisel on beeži joonega tähistatud aeg, mil indiviid ei olnud geenidoonor, ja sinise joonega on märgitud geenidoonoriks olemise aeg. Pidevjoonega on tähistatud see aeg, mil toimunud sündmused on analüüsi läbiviimise ajaks teada, kriipsjoonega see aeg, mille kohta info analüüsi tegemise ajal puudub. Risti ja ringjoonega on tähistatud vastavalt surm ja haigestumine. Subjektid 5 ja 6 kujutavad vasakult tõkestatust, nemad ei ole geenivaramuga liitunud

ja nende haigestumise kohta meil info puudub. Subjekt 3 on saanud huvipakkuva haiguse enne geenivaramuga liitumist ja on seega uuritava haiguse suhtes jälgimiseelne juht. Subjekt 2 on haigestunud pärast liitumist ja on seega uuritava haiguse suhtes jälgimisaegne juht. Subjekt 4 on paremalt tsenseeritud surma tõttu ja subjekt 1 on paremalt tsenseeritud seetõttu, et huvipakkuv sündmus ei olnud toimunud enne analüüsi tegemist.

2.2 Võrdeliste riskide mudel

See peatükk koos alapeatükkidega põhineb raamatul (Collett, 2015), kui ei ole viidatud teisiti.

Anname enne võrdeliste riskide mudeli tutvustamist ülevaate elukestusanalüüsi tähtsamatest mõistetest ja definitsioonidest.

Olgu T mittenegatiivne elukestust tähistav juhuslik suurus tihedusfunktsiooniga $f(t)$ ja jaotusfunktsiooniga $F(t)$. Üleelamisfunktsioon $S(t)$ on tõenäosus, et huvipakkuv sündmus toimub pärast ajamomenti $t \geqslant 0$:

$$S(t) = P(T > t) = 1 - F(t).$$

Riskifunktsioon $h(t)$ iseloomustab hetkelist sündmuse toimumise riski ajamomendil $t \geqslant 0$, kui on teada, et see ei toimunud enne ajamomenti t. Pideva aja korral on riskifunktsioon kujul

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t | \, T \geqslant t)}{\Delta t} = \frac{f(t)}{S(t)}$$

ja selle väärtused on vahemikus $[0, \infty)$.

Eelmise võrduse põhjal saame, et $t \geqslant 0$ korral kehtivad järgmised seosed:

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t),$$

millest

$$S(t) = \exp(-H(t)) \tag{1}$$

ja

$$H(t) = -\log S(t), \tag{2}$$

kus $H(t) = \int_0^t h(u)du$ on kumulatiivne riskifunktsioon, mis kirjeldab kumulatiivset sündmuse toimumise riski ajaks t.

Selleks, et leida elukestust prognoosivaid tunnuseid ja nende mõju riskifunktsioonile, kasutatakse võrdeliste riskide mudeleid (*proportional hazards models*).

Olgu meil n vaatlust, p kirjeldavat tunnust ja olgu i-nda ($i = 1, \ldots, n$) vaatluse riskifunktsioon $h_i(t)$ ning $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})^T$ tema argumenttunnuste väärtuste vektor. Olgu $h_0(t)$ baasriskifunktsioon ehk riskifunktsioon juhul, kui kõigi seletavate tunnuste väärtusteks on 0. Siis võrdeliste riskide mudeli korral saab i-ndale vaatlusele vastava riskifunktsiooni kirjutada kujul

$$h_i(t) = h_0(t)\psi(\boldsymbol{x}_i),$$

kus $\psi(\boldsymbol{x}_i)$ on ajast t mittesõltuv $\boldsymbol{x}_i$ funktsioon, mis kirjeldab riskifunktsioonide suhet ehk riskimäärade suhet (HR, *hazard ratio*). Et riskimäärade suhe ei saa olla negatiivne, kasutatakse tihti $\psi(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$, kus $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ on argumenttunnustele vastavate parameetrite vektor. Eelnevat kokku võttes on võrdeliste riskide mudeli kujuks

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{3}$$

Seega kahe indiviidi a ja b korral, kelle argumenttunnuste vektorid ning riskifunktsioonid on

vastavalt $\boldsymbol{x}_a$ ja $\boldsymbol{x}_b$ ning $h_a(t)$ ja $h_b(t)$, kehtib võrdeliste riskide eeldus:

$$\frac{h_a(t)}{h_b(t)} = \frac{h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}_a)}{h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}_b)} = \exp\{\boldsymbol{\beta}^T(\boldsymbol{x}_a - \boldsymbol{x}_b)\}$$

ehk riskimäärade suhe ei sõltu ajast t.

Viies mudeli (3) kujule

$$\log\left(\frac{h_i(t)}{h_0(t)}\right) = \boldsymbol{\beta}^T\boldsymbol{x}_i,$$

näeme, et tegemist on lineaarse mudeliga riskimäärade suhte logaritmile.

Parameetriliste võrdeliste riskide mudelite korral on baasriskifunktsioon $h_0(t)$ määratud mingi kindla jaotuse parameetritega. Poolparameetrilisel juhul hinnatakse ainult parameetrid $\boldsymbol{\beta}$ ja baasriskifunktsiooni ei määrata.

2.2.1 Coxi võrdeliste riskide mudel

Coxi võrdeliste riskide mudel, mis on üks populaarsemaid meetodeid elukestusanalüüsis, on poolparameetriline mudel, sest baasriskifunktsiooni $h_0(t)$ mudelis (3) ei määrata, küll aga leitakse hinnangud regressioonikordajatele β_j $(j = 1, \ldots, p)$ ja huvipakkuvatele riskimäärade suhetele.

Parameetrite hindamine

Coxi mudeli parameetrid saab hinnata suurima tõepära meetodil. Kasutusel on niinimetatud osaline tõepärafunktsioon, mis ei sõltu baasriskist $h_0(t)$ ega kasuta täpseid vaadeldud sündmuse toimumise aegu, vaid ainult nende aegade omavahelist järjestust.

Olgu meil vaatluse all n subjekti, kellest r-il toimub sündmus, ja ülejäänud $n - r$ subjekti on paremalt tsenseeritud. Järjestame sündmuse toimumiste ajad $t_{(1)} < \ldots < t_{(r)}$. Tähistagu iga $j = 1, \ldots, r$ korral $R_{(j)}$ nende inimeste hulka, kellel ei ole toimunud sündmust ja kes

ei ole tsenseeritud enne hetke $t_{(j)}$, ning olgu $\boldsymbol{x}_{(j)}$ argumenttunnuste vektor indiviidil, kellel esineb sündmus hetkel $t_{(j)}$. Selliste tähistuste korral on Coxi (1972; 1975) tuletatud mudelile (3) vastava osalise tõepärafunktsiooni kujuks

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{(j)})}{\sum_{l \in R_{(j)}} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)}.$$

Olgu $\boldsymbol{U}(\boldsymbol{\beta})$ logaritmitud osalise tõepärafunktsiooni esimene tuletis $\boldsymbol{\beta}$ järgi:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

Suurima osalise tõepära hinnangud leitakse siis võrrandist $\boldsymbol{U}(\boldsymbol{\beta}) = 0$, mille lahendamiseks kasutatakse tavaliselt iteratiivseid meetodeid, näiteks Newton-Raphsoni meetodit (Therneau ja Grambsch, 2000: 41).

Coxi võrdeliste riskide mudeli eelduseks on see, et riskimäärade suhe oleks ajas muutumatu, ehk et riskimäärad oleksid võrdelised. Selle eelduse täidetust saab kontrollida näiteks Schoenfeldi jääkide abil (vt lisa 1).

Võrdsed sündmuste toimumiste ajad

Siiani oleme teinud eelduse, et kõik sündmuste toimumiste ajad on erinevad. Praktikas võib aga juhtuda, et näiteks mõõtmiste ebatäpsuse tõttu esineb andmetes ka võrdseid sündmuse toimumiste aegu. Väga tihti on sündmuse toimumise aeg teada päeva täpsusega, kuid samal päeval saab sündmus toimuda mitmel indiviidil; seda näeme eelkõige suurtes valimites, nagu ka geenivaramu andmebaasis. Kui andmetes esineb võrdseid sündmuse toimumiste aegu, siis muutub vastav osaline tõepärafunktsioon arvutuslikult keeruliseks ja kasutatakse erinevaid lähendeid.

Enamasti kasutatakse lihtsat Breslow' lähendit (Breslow, 1972; Peto, 1972), kus samadel aegadel toimuvate sündmuste puhul eeldatakse, et need on erinevad ja toimuvad üksteise järel:

$$L(\boldsymbol{\beta}) \approx \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{s}_j)}{\left\{\sum_{l \in R_{(j)}} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)\right\}^{d_j}},$$

kus iga $j = 1, \ldots, r$ korral on $D_{(j)}$ nende indiviidide hulk, kellel toimus sündmus hetkel $t_{(j)}$, $\boldsymbol{s}_j = \sum_{i \in D_{(j)}} \boldsymbol{x}_i$ on nende indiviidide argumenttunnuste vektorite summa, kellel toimub sündmus hetkel $t_{(j)}$, ja d_j on ajahetkel $t_{(j)}$ toimunud sündmuste arv.

Veidi keerulisema ja täpsema lähendi pakkus Efron (1977), mis on vaikimisi kasutusel ka R-i paketi *survival* (Therneau, 2020) funktsioonides:

$$L(\boldsymbol{\beta}) \approx \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{s}_j)}{\prod_{k=1}^{d_j} \left\{\sum_{l \in R_{(j)}} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l) - (k-1) d_j^{-1} \sum_{l \in D_{(j)}} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)\right\}}.$$

Paneme tähele, et juhul kui kõik sündmuste toimumiste ajad on erinevad ehk $d_j = 1$ iga $j = 1, \ldots, r$ korral, siis on mõlemad lähendid täpselt samad ja võrdsed esialgse Coxi osalise tõepärafunktsiooniga. Kui võrdseid sündmuste toimumiste aegu on vähe, siis annavad mõlemad lähendid sarnaseid tulemusi. Selles töös kasutame võrdsete haigestumisaegade puhul Efroni tõepärafunktsiooni lähendit.

Martingaalijäägid

Coxi mudeli jaoks on defineeritud erinevaid jääke: näiteks Cox-Snelli jäägid mudeli üleüldise sobivuse hindamiseks, martingaalijäägid kovariaatide sobiva funktsionaalse kuju määramiseks, hälbimuse jäägid erindite leidmiseks, Schoenfeldi jäägid ning skoorijäägid võrdeliste riskide eelduse kontrollimiseks ja erindlike vaatluste leidmiseks. Siin peatükis kirjeldame lähemalt martingaalijääke.

Martingaalijääkide nimetus tuleb sellest, et need on võimalik tuletada, kasutades juhuslike protsesside, täpsemalt martingaalide teooriat. Neid jääke kasutatakse näiteks selgitavate tunnuste funktsionaalse kuju määramiseks. Martingaalijääkide arvutamiseks on vaja hinnangut kumula-

tiivsele riskifunktsioonile. Riski- ja üleelamisfunktsiooni hindamisest on kirjutatud lisas 2.

Olgu meil nagu ennegi vaatluse all n indiviidi, kellest r-il toimub sündmus ja ülejäänud $n-r$ on paremalt tsenseeritud, ning olgu meil hinnatud p argumenttunnusega Coxi mudel. Olgu $\delta_i = 0$, kui i-s $(i = 1, \ldots, n)$ vaatlus on tsenseeritud, ja $\delta_i = 1$ muidu. Olgu i-nda indiviidi vaatlusajaks t_i ja olgu talle hinnatud Nelson-Aaleni kumulatiivne riskifunktsioon $\hat{H}_i(t)$ (vt lisa 2 valem (8)).

Martingaalijääk i-nda indiviidi jaoks on defineeritud järgmiselt:

$$\hat{m}_i = \delta_i - \hat{H}_i(t_i) = \delta_i - \hat{H}_0(t_i)\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i).$$

Martingaalijääkide väärtused on $-\infty$ ja 1 vahel, kusjuures kõik tsenseeritud vaatluste martingaalijäägid on negatiivsed.

Paneme tähele, et suurus δ_i on i-nda indiviidi vaadeldud sündmuste arv ajavahemikus algmomendist hetkeni t_i (kas 0 või 1 vastavalt sellele, kas indiviid on tsenseeritud või mitte). Hinnatud kumulatiivset riskifunktsiooni $\hat{H}_i(t_i)$ aga saab interpreteerida kui oodatavat sündmuste arvu indiviidil i ajavahemikus algmomendist hetkeni t_i. See tähendab, et i-nda indiviidi martingaalijääki saab interpreteerida kui tema ajavahemikus algmomendist hetkeni t_i vaadeldud ja oodatud sündmuste arvu vahet. Seega saab martingaalijääkide abil leida need subjektid, kellel hinnatud mudeli põhjal toimus sündmus liiga vara või liiga hilja – absoluutväärtuselt suured negatiivsed jäägid vastavad nendele subjektidele, kelle elukestus on pikk, kuid kellel mudeli põhjal pidanuks sündmus toimuma varem, ühelähedased jäägid vastavad neile, kelle elukestus oli lühem, kui võiks arvata mudeli põhjal.

Martingaalijääke saab kasutada Coxi mudeli lähendamiseks lineaarse regressiooniga. Selles töös kasutatakse martingaalijääke, hindamaks SNP-de mõju haigestumisele. Koosnegu i-nda indiviidi mudelis kasutatavate argumenttunnuste vektor kahte tüüpi tunnustest $\boldsymbol{x}_i = (g_i, z_{1i}, \ldots, z_{pi})^T$, kus g tähistab SNP-d ja z_j $(j = 1, \ldots, p)$ on ülejäänud kovariaadid. Olgu nende argumenttunnuste parameetriteks vastavalt β_g ja $\beta_1, \ldots, \beta_p$. SNP mõju hindamiseks sobitatakse esmalt Coxi mudel nii, et kaasatakse ainult mittegeneetilised argumenttunnused ehk

hinnatakse nii-öelda nullmudel SNP suhtes:

$$h_i(t) = h_0(t)\exp(\beta_1 z_{1i} + \ldots + \beta_p z_{pi}).$$

Kui on leitud selle mudeli parameetrite hinnangud $\hat{\beta}_j$ $(j = 1, \ldots, p)$ ja Nelson-Aaleni hinnang kumulatiivsele baasriskifunktsioonile, siis saab arvutada vastavad martingaalijäägid:

$$\hat{m}_i = \delta_i - \hat{H}_0(t_i)\exp(\hat{\beta}_1 z_{1i} + \ldots + \hat{\beta}_p z_{pi}).$$

Therneau *et al*. (1990) on näidanud, et selliselt arvutatud martingaalijäägid on seotud mudelist välja jäänud argumenttunnusega (antud juhul on selleks genotüüp g). Kui genotüüp g tuleks sarnaselt mudelis olevate tunnustega lisada mudeli lineaarsesse osasse ilma teisenduseta, siis on ka tema seos martingaalijäägiga lineaarne, ehk

$$\frac{\hat{m}_i}{c} = \beta_0 + \tilde{\beta}_g g_i + \varepsilon_i,$$

kus $c = d/n$ on sündmuste arvu ja kõigi vaatluste arvu suhe, β_0 on lineaarse mudeli vabaliige, $\tilde{\beta}_g \approx \beta_g$ ja $\varepsilon_i \sim N(0, \sigma^2)$ on sõltumatud vead. Seega on selle asemel, et iga huvipakkuva SNP-ga hinnata Coxi mudel, mis on ülegenoomsete seoseuuringute puhul väga aja- ja arvutusmahukas, võimalik hinnata üks Coxi nullmudel, kus ühegi SNP mõju ei ole arvesse võetud, ja siis hinnata iga SNP-ga lineaarne mudel, kus uuritavaks tunnuseks on nullmudeli põhjal arvutatud skaleeritud martingaalijääk.

Teoreetiliselt on teada, et selline lähendamine töötab ainult kindlates piirides. Selles töös uurime, millal ja kui hästi sobib martingaalijääkide meetodi rakendamine Coxi mudelite hindamise asemel.

2.2.2 Weibulli jaotusega võrdeliste riskide mudel

Coxi võrdeliste riskide mudel on poolparameetriline mudel, sest baasriskifunktsiooni kuju ei määrata. Seetõttu ei saa Coxi mudelit kasutada elukestusandmete simuleerimiseks. Seega, kuigi elukestusandmete analüüsimiseks kasutatakse enamasti Coxi mudelit, siis elukestusandmete simuleerimiseks kasutatakse tavaliselt parameetrilisi mudeleid.

Parameetriliste võrdeliste riskide mudeli puhul, kus baasriskifunktsiooni kuju on määratud, kasutatakse kõige enam Weibulli jaotust (Kleinbaum ja Klein, 2012: 304). Kirjeldame siin Weibulli jaotuse parametrisatsiooni, mida kasutatakse R-i funktsioonides *dweibull*, *pweibull*, *rweibull* ja *qweibull*. Weibulli jaotusel on kaks parameetrit: kujuparameeter $a > 0$ ja skaalaparameeter $b > 0$.

Olgu juhuslik suurus T Weibulli jaotusega $T \sim W(a, b)$. Sel juhul on T tihedusfunktsiooni ja jaotusfunktsiooni kujud $t \geqslant 0$ korral vastavalt

$$f(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} \exp\left\{-\left(\frac{t}{b}\right)^{a}\right\}$$

ja

$$F(t) = 1 - \exp\left\{-\left(\frac{t}{b}\right)^{a}\right\}.$$

Weibulli jaotuse riskifunktsioon on kujul

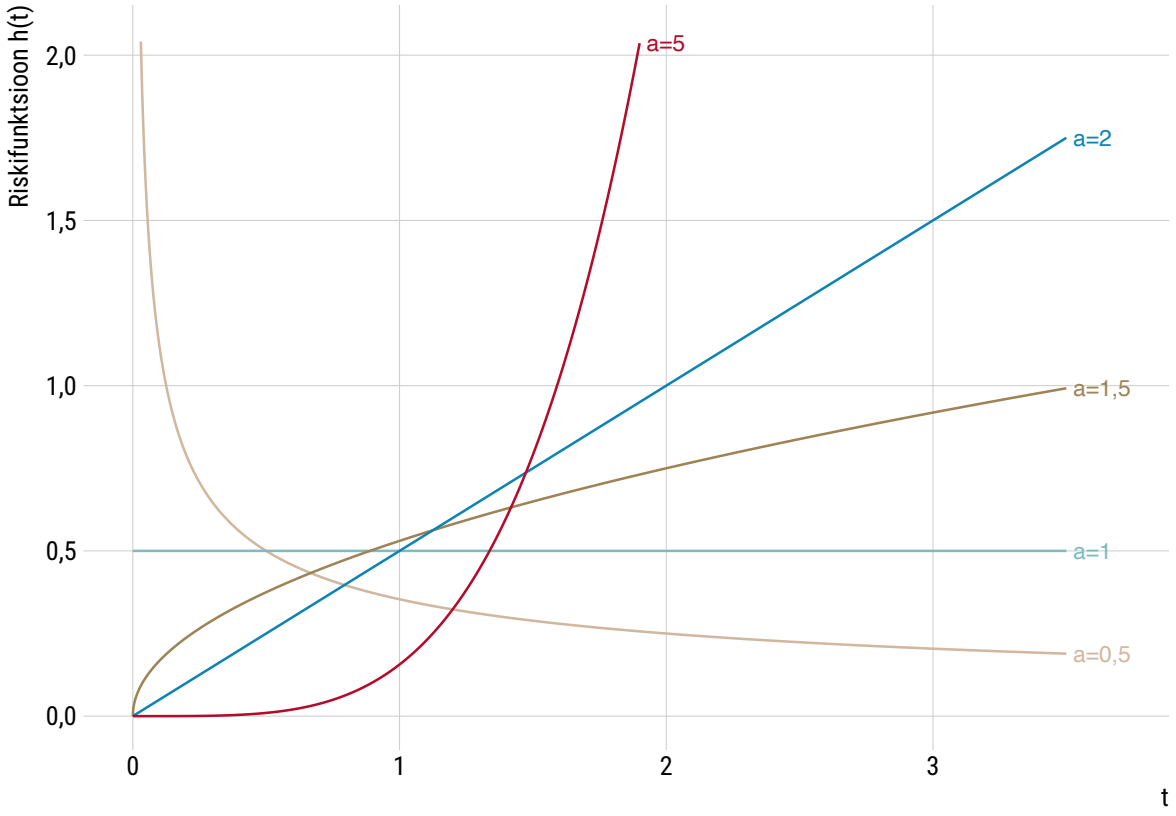
$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} = ab^{-a}t^{a-1}, \quad t \geqslant 0.$$

Seega on $a > 1$ korral tegemist kasvava riskiga, $a < 1$ korral kahaneva riskiga ja $a = 1$ korral

on risk konstantne. Weibulli jaotuse kumulatiivne riskifunktsioon on

$$H(t) = \int_0^t h(u)du = \\ = \int_0^t ab^{-a}u^{a-1}du = \left(\frac{t}{b}\right)^a, \quad t \geqslant 0. \tag{4}$$

Joonisel 2 on näited Weibulli jaotuse riskifunktsiooni kujude kohta erinevate kujuparameetrite a korral.



Joonis 2. Weibulli jaotuse $W(a, b = 2)$ riskifunktsioon $h(t)$ erinevate kujuparameetrite a korral

Tänu lihtsale riski- ja üleelamisfunktsioonile ning erinevatele võimalikele riskifunktsiooni kujudele on Weibulli jaotus parameetrilises elukestusanalüüsis laialdaselt kasutusel.

Vaatame nüüd võrdeliste riskide mudelit juhul, kui elukestused T on Weibulli jaotusega. Kui baaselukestus on Weibulli jaotusega $W(a, b)$, siis i-nda indiviidi, kelle argumenttunnuste vektor

on $\boldsymbol{x}_i$, riskifunktsioon avaldub kujul

$$\begin{aligned} h_i(t) &= h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_i) = \\ &= ab^{-a}t^{a-1}\exp(\boldsymbol{\beta}^T\mathbf{x}_i) = \\ &= a\left(b\left\{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)\right\}^{-\frac{1}{a}}\right)^{-a}t^{a-1} = ab_i^{-a}t^{a-1}, \end{aligned}$$

kus

$$b_i = b\left\{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)\right\}^{-\frac{1}{a}}. \tag{5}$$

Seega on i-nda indiviidi elukestus Weibulli jaotusega $W(a, b_i)$. Weibulli jaotusega võrdeliste riskide mudelite puhul eeldataksegi, et kujuparameeter a on igal indiviidil samasugune, aga skaalaparameeter b_i sõltub baasparameetritest a ja b, parameetrite vektorist $\boldsymbol{\beta}$ ning argumenttunnuste vektorist $\mathbf{x}_i$.

2.3 Binaarse tunnuse modelleerimine

Jälgimiseelsete haigusjuhtude uurimisel vaadeldakse haigusseisundit binaarse tunnusena, mille väärtuseks on kas 1 või 0 vastavalt diagnoosi esinemisele või puudumisele. Kirjeldame siin kahte binaarse tunnuse modelleerimise meetodit (nimetame neid edaspidi lihtsuse mõttes binaarseteks mudeliteks) – logistilist regressiooni ja täiend-log-log regressiooni – ning nende mudelite efektisuuruste seost võrdeliste riskide mudelite efektisuurusega ehk riskimäärade suhtega.

Olgu meil valimis n vaatlust ja p seletavat tunnust. Olgu Y_i $(i = 1, \ldots, n)$ binaarne juhuslik suurus, mis kirjeldab meile huvipakkuva sündmuse toimumist ehk diagnoosi esinemist i-ndal indiviidil, ja olgu $p_i = P(Y_i = 1)$ selle sündmuse toimumise tõenäosus.

2.3.1 Logistiline mudel

Logistilise regressiooni korral kasutatakse uuritava sündmuse tõenäosuse kirjeldamiseks sõltumatute argumenttunnuste lineaarkombinatsiooni kaudu logit-seosefunktsiooni:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i,$$

kus β_0 ja $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ on mudeli tundmatud parameetrid, $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})^T$ on i-nda indiviidi argumenttunnuste vektor ja $p_i/(1-p_i)$ on sündmuse toimumise šanss. Sündmuse toimumise šanss näitab, mitu korda on uuritava sündmuse toimumine tõenäolisem kui sündmuse mittetoimumine.

Tundmatud parameetrid β_j $(j = 0, \ldots, p)$ hinnatakse tavaliselt suurima tõepära meetodil ja hinnangute arvutamiseks kasutatakse iteratiivseid meetodeid, näiteks R-i funktsioonis *glm* on kasutusel Fisheri skoorimeetod.

Logistilise regressioonimudeli parameetreid interpreteeritakse tavaliselt šansside suhte (OR, *odds ratio*) kaudu. Erinegu indiviidid a ja b vaid j-nda $(j = 1, \ldots, p)$ argumenttunnuse väärtuse poolest: $x_{ja} = x_{jb} + 1$. Siis on nende indiviidide šansside suhte ja j-nda argumenttunnuse parameetri β_j seos järgmine:

$$\frac{p_a/(1-p_a)}{p_b/(1-p_b)} = \frac{\exp(\beta_0 + \beta_1 x_{1a} + \ldots + \beta_j(x_{jb}+1) + \ldots + \beta_p x_{pa})}{\exp(\beta_0 + \beta_1 x_{1a} + \ldots + \beta_j x_{jb} + \ldots + \beta_p x_{pa})} = \exp(\beta_j).$$

See tähendab, et kui j-nda argumenttunnuse väärtus muutub ühe ühiku võrra (ja teised väärtused jäävad samaks), siis šansid muutuvad $\exp(\beta_j)$ korda.

Šansside suhe OR ja riskimäärade suhe HR

Eespool nägime, et võrdeliste riskide mudeli puhul saab mudeli parameetreid interpreteerida kui logaritme riskimäärade suhetest, logistilise mudeli puhul aga kui logaritme šansside suhetest. Riskimäärade suhe ja šansside suhe on kaks sagedasti raporteeritavat suurust, kuid nende

interpreteerimisel ollakse sageli hooletu, kusjuures mõlemaid kasutatakse tihti hoopis kolmanda näitaja – suhtelise riski (*relative risk*, *risk ratio*) – iseloomustamiseks. Teatud juhtudel – näiteks, kui huvipakkuva sündmuse toimumise tõenäosus on väike või jälgimisaeg lühike – ongi need näitajad sarnased. Samuti on kõigi nende kolme näitaja suund alati sama: kui teame, et üks neist on suurem kui 1, on seda ka teised kaks, ja vastupidi. (Davies *et al.*, 1998; Bangdiwala, 2010; Sutradhar ja Austin, 2018)

Vaatame lähemalt, kuidas on omavahel seotud šansside suhe (OR) ja riskimäärade suhe (HR). Olgu meil lihtsuse mõttes võrdluses kaks gruppi: kontrollgrupp, kus argumenttunnus $x = 0$ ja riskigrupp, kus $x = 1$, riskifunktsioonidega vastavalt h_0 ja h_1. Olgu β tunnusele x vastav Coxi mudeli regressioonikordaja ja olgu jälgimisaja pikkuseks $t \geqslant 0$. Võrdeliste riskide mudeli kehtimise korral avaldub riskigrupi riskifunktsioon h_1 kontrollgrupi riskifunktsiooni h_0 kaudu:

$$h_1(t) = h_0(t) \exp(\beta x) = h_0(t) \exp(\beta)$$

ja kahe grupi riskimäärade suhe on

$$\mathrm{HR} = \frac{h_1(t)}{h_0(t)} = \frac{h_0(t) \exp(\beta)}{h_0(t)} = \exp(\beta).$$

Tähistame nüüd huvipakkuva sündmuse toimumise tõenäosuse ajavahemikus algmomendist hetkeni t kontrollgrupis p_0 ja riskigrupis p_1. Kumulatiivse baasriskifunktsiooni H_0 kaudu saame need tõenäosused kirja panna järgmiselt:

$$\begin{aligned} p_0 &= 1 - S_0(t) = 1 - \mathrm{e}^{-H_0(t)}, \\ p_1 &= 1 - S_1(t) = 1 - \mathrm{e}^{-H_1(t)} = 1 - \mathrm{e}^{-H_0(t) \exp(\beta)}. \end{aligned}$$

Kahe grupi šansside suhte saame nende tõenäosuste kaudu arvutada järgmiselt:

$$\begin{aligned}
\mathrm{OR} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} &= \frac{p_1(1-p_0)}{p_0(1-p_1)} = \\
&= \frac{\left(1-\mathrm{e}^{-H_0(t)\exp(\beta)}\right)\mathrm{e}^{-H_0(t)}}{\left(1-\mathrm{e}^{-H_0(t)}\right)\mathrm{e}^{-H_0(t)\exp(\beta)}} = \\
&= \frac{1-\mathrm{e}^{-H_0(t)\exp(\beta)}}{\left(\mathrm{e}^{H_0(t)}-1\right)\mathrm{e}^{-H_0(t)\exp(\beta)}} = \\
&= \frac{\mathrm{e}^{H_0(t)\exp(\beta)}-1}{\mathrm{e}^{H_0(t)}-1} = \\
&= \frac{(1-p_0)^{-\exp(\beta)}-1}{(1-p_0)^{-1}-1} = \frac{(1-p_0)^{-\mathrm{HR}}-1}{(1-p_0)^{-1}-1},
\end{aligned}$$

seega avaldub šansside suhe baasgrupi levimuse p_0 ja riskimäärade suhte $\mathrm{HR} = \exp(\beta)$ kaudu ning kumulatiivset baasriskifunktsiooni ei ole vaja teada (Green ja Symons, 1983).

Seda seost on illustreeritud joonisel 3. Sellel on näidatud, kuidas sõltub šansside suhe OR kontrollgrupi levimusest p_0 erinevate riskimäärade suhete HR korral. Paneme tähele, et

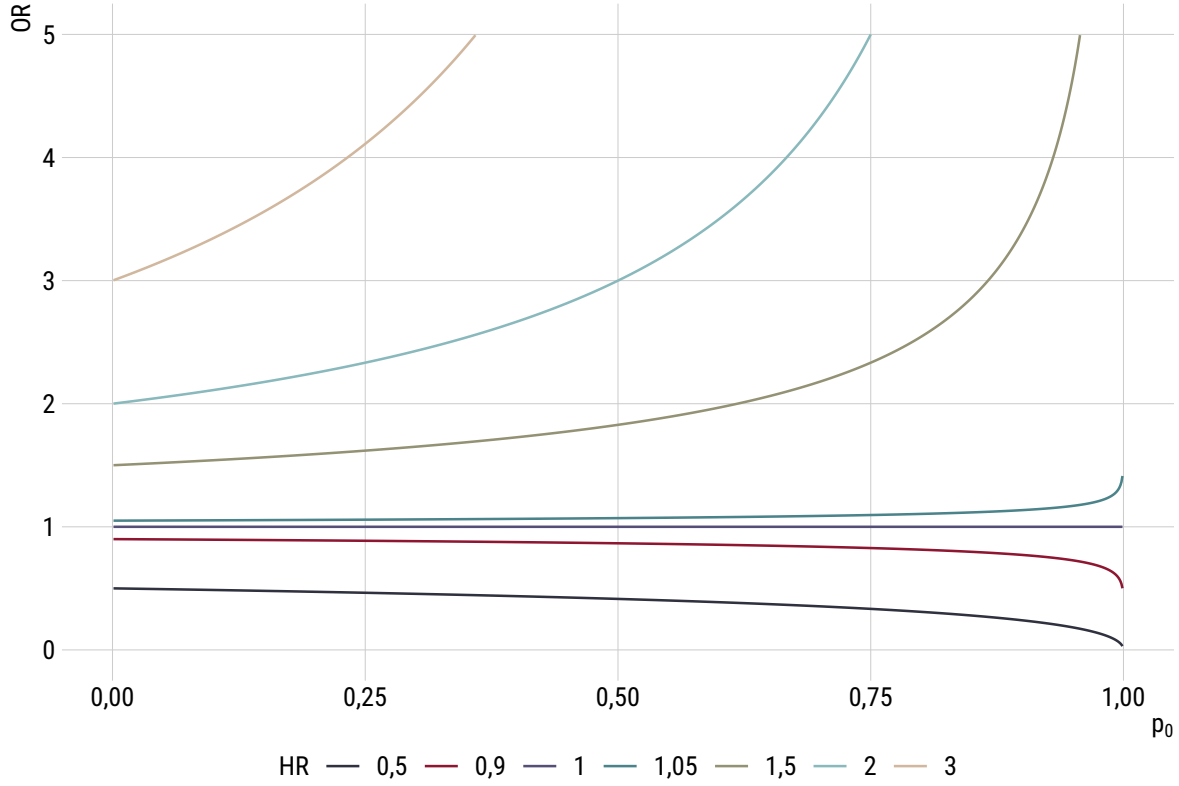
$$\begin{cases}
\mathrm{OR} > \mathrm{HR}, \text{ kui } \mathrm{HR} > 1 \text{ ehk } \beta > 0, \\
\mathrm{OR} < \mathrm{HR}, \text{ kui } \mathrm{HR} < 1 \text{ ehk } \beta < 0, \\
\mathrm{OR} = \mathrm{HR}, \text{ kui } \mathrm{HR} = 1 \text{ ehk } \beta = 0.
\end{cases}$$

Samuti näeme, et HR ja OR väärtused on sarnased, kui HR on ühe lähedal ning väärtused on seda erinevamad, mida suurem on p_0.

Et $\exp(x) \approx 1 + x$, kui x on absoluutväärtuselt väike, siis saame šansside suhet ka lähendada:

$$\begin{aligned}
\mathrm{OR} = \frac{\mathrm{e}^{H_0(t)\mathrm{HR}}-1}{\mathrm{e}^{H_0(t)}-1} &\approx \\
&\approx \frac{1+H_0(t)\mathrm{HR}-1}{1+H_0(t)-1} = \mathrm{HR},
\end{aligned}$$

juhul kui vaatlusaeg t on lühike ja seega kumulatiivne riskifunktsioon $H_0(t)$ väike, ning sündmuse esinemise tõenäosus vaatlusaja jooksul ega riskifaktori mõju ei ole suured (Green ja Symons, 1983).



Joonis 3. Šansside suhe OR erinevate riskimäärade suhete HR ja baasgrupi levimuste p_0 korral

2.3.2 Täiend-log-log mudel

Analoogiliselt logistilise regressiooniga kasutatakse ka täiend-log-log mudelit, kirjeldamaks uuritava sündmuse toimumise tõenäosust ja selle muutumist sõltuvalt argumenttunnuste väärtuste muutumisest. Selle mudeli puhul rakendatakse täiend-log-log (cloglog, *complementary log-log*) seosefunktsiooni:

$$\operatorname{cloglog}(p_i) = \log(-\log(1-p_i)) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i,$$

kus β_0 ja $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ on mudeli tundmatud parameetrid, $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})^T$ on i-nda indiviidi argumenttunnuste vektor, ja millest sündmuse toimumise tõenäosus p_i avaldub kujul

$$p_i = 1 - \exp(-\exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)).$$

Täiend-log-log mudeli saab R-is hinnata funktsiooniga *glm*, kus argumendina tuleb kasutada *family = binomial(link = "cloglog")*. Täiend-log-log mudeli parameetrid hinnatakse enamasti suurima tõepära meetodil, funktsioonis *glm* kasutatakse selleks Fisheri skoorimeetodit.

Vaatame, kuidas on seotud Coxi võrdeliste riskide elukestusmudel täiend-log-log mudeliga. Olgu meil vaatluse all n indiviidi ja vastaku haiguse diagnoosimise vanused T_i $(i = 1, \ldots, n)$ Coxi võrdeliste riskide mudelile ehk olgu haigestumise vanusele T_i vastav kumulatiivne riskifunktsioon

$$H_i(t) = H_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i),$$

kus $\boldsymbol{\beta}$ on võrdeliste riskide mudeli parameetrite vektor, $\boldsymbol{x}_i$ on i-nda indiviidi argumenttunnuste vektor ja H_0 on kumulatiivne baasriskifunktsioon, mille kuju jäetakse Coxi mudeli puhul määramata.

Olgu jälgimisaja pikkuseks i-ndal indiviidil $t_i \geqslant 0$ (see võib olla näiteks liitumisvanus). Haigestumise indikaator on siis $Y_i = I(T_i \leqslant t_i)$. Kasutades seost (1), saame tõenäosuse, et i-ndal indiviidil on haigus diagnoositud ajaks t_i, kirjutada kujul

$$\begin{aligned} p_i = P(Y_i = 1 |\, \boldsymbol{x}_i, t_i) &= P(T_i \leqslant t_i | \boldsymbol{x}_i) = \\ &= 1 - S_i(t_i) = \\ &= 1 - \exp(-H_i(t_i)) = \\ &= 1 - \exp(-H_0(t_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)) = \\ &= 1 - \exp(-\exp(\log(H_0(t_i)) + \boldsymbol{\beta}^T \boldsymbol{x}_i)). \end{aligned}$$

Juhul kui $H_0(t_i) = \exp(\beta_0 + \beta_t t_i)$, saame

$$p_i = 1 - \exp(-\exp(\beta_0 + \beta_t t_i + \boldsymbol{\beta}^T \boldsymbol{x}_i)).$$

See vastab täpselt täiend-log-log mudelile vabaliikmega β_0 ja parameetrite vektoriga $(\beta_t; \boldsymbol{\beta})$

(jälgimisaeg t_i on lisatud argumenttunnusena). Seega saab ka täiend-log-log mudeli parameetreid interpreteerida kui riskimäärade suhteid. See kehtib aga vaid siis, kui on täidetud eeldus, et kumulatiivne risk avaldub kui $H_0(t_i) = \exp(\beta_0 + \beta_t t_i)$. Viimane eeldus ei pruugi aga alati täidetud olla. Meid huvitaval juhul, kus haigestumine sõltub genotüübist, avaldub i-nda indiviidi haigestumise tõenäosus võrdeliste riskide mudeli kehtimise korral järgmiselt:

$$\begin{aligned} P(T_i \leqslant t_i |\, g_i) &= 1 - S_i(t_i) = \\ &= 1 - \exp(-H_0(t_i)\exp(\beta_g g_i)), \end{aligned}$$

kus $g_i \in \{0, 1, 2\}$ on i-nda indiviidi genotüüp, H_0 on kumulatiivne baasriskifunktsioon ja β_g on genotüübile vastav parameeter võrdeliste riskide mudelis.

Tähistades haigestumise indikaatori $Y_i = I(T_i \leqslant t_i)$, avaldub haigestumise tõenäosus täiend-log-log mudeli kehtimise korral järgmiselt:

$$P(Y_i = 1 |\, t_i, g_i) = 1 - \exp(-\exp(\beta_0 + \beta_g' g_i + \beta_t t_i)),$$

kus β_0, β_g' ja β_t on vastavalt täiend-log-log mudeli vabaliige, genotüübile vastav parameeter ja liitumisvanusele vastav parameeter.

Näeme, et mõlema mudeli korral saame üleelamistõenäosuse kirjutada liitumisvanusest sõltuva funktsiooni ja genotüübist sõltuva funktsiooni kaudu: võrdeliste riskide mudeli puhul saame selle kirjutada kujul

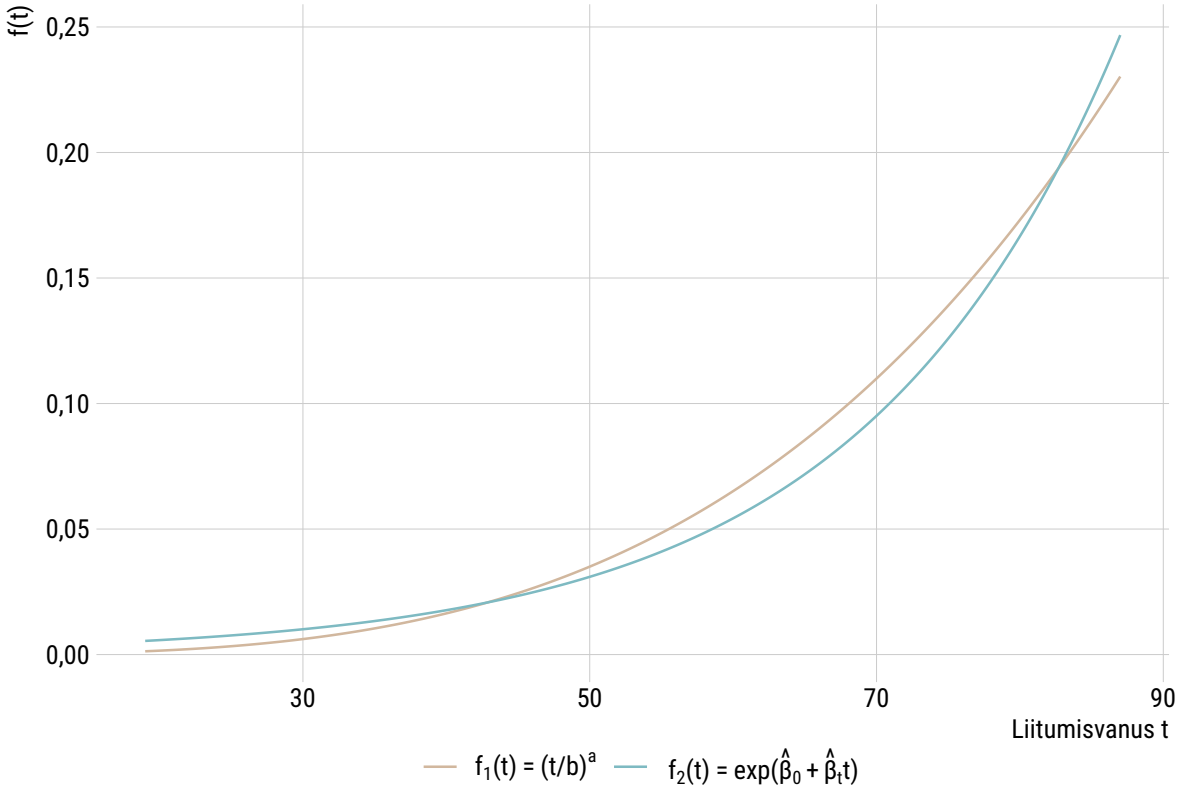
$$P(T_i \leqslant t_i |\, g_i) = 1 - \exp(-f_1(t_i)\exp(\beta_g g_i)),$$

kus $f_1(t_i) = H_0(t_i)$, ja täiend-log-log mudeli puhul kujul

$$\begin{aligned} P(Y_i = 1 |\, t_i, g_i) &= 1 - \exp(-\exp(\beta_0 + \beta_t t_i)\exp(\beta_g' g_i)) = \\ &= 1 - \exp(-f_2(t_i)\exp(\beta_g' g_i)), \end{aligned}$$

kus $f_2(t_i) = \exp(\beta_0 + \beta_t t_i)$. Kuna mõlema mudeli puhul saame hinnata ühtesama haigestumise tõenäosust, siis nüüd juhul, kui $f_1(t_i) = f_2(t_i)$ valimi liitumisvanuste t_i korral, siis ka $\beta_g = \beta'_g$. Et f_1 on kumulatiivne baasriskifunktsioon ja võib olla kuitahes keerulise kujuga, kuid f_2 on alati e aste lineaarsest liitumisvanuse funktsioonist, siis üldjuhul need funktsioonid võrdsed ei ole. Praktikas võivad need funktsioonid siiski sarnased olla.

Näitena on joonisel 4 kujutatud f_1 ja f_2 juhul, kui haigestumisandmed on simuleeritud vastavalt Weibulli jaotusega võrdeliste riskide mudelile.



Joonis 4. Liitumisvanusest t sõltuvad funktsioonid f_1 ja f_2 vastavalt Weibulli jaotusega võrdeliste riskide mudeli korral ja täiend-log-log mudeli korral. Parameetrite väärtusteks on $a = 3{,}4, b = 134, \hat{\beta}_0 = -6{,}28, \hat{\beta}_t = 0{,}0561$.

Selle näite jaoks on i-nda $(i = 1, \ldots, 1000)$ indiviidi haigestumise vanus T_i genereeritud Weibulli jaotusest

$$W\left(a, b\left\{\exp(\beta_g g_i)\right\}^{-\frac{1}{a}}\right),$$

kus $a = 3{,}4$ ja $b = 134$ on Weibulli jaotuse baasparameetrid, $g_i \in \{0, 1, 2\}$ on i-nda indiviidi

genotüüp ning $\beta_g = \log(1{,}5) \approx 0{,}405$ on genotüübile vastav parameeter. Sel juhul saame Weibulli jaotuse kumulatiivse riskifunktsiooni kuju (4) põhjal, et

$$f_1(t) = \left(\frac{t}{134}\right)^{3{,}4}.$$

Liitumisvanused t_i on valitud tagasipanekuta juhuvalikuga TÜ EGV geenidoonorite liitumisvanuste seast ja i-nda indiviidi haigestumise indikaatoriks on $Y_i = I(T_i \leqslant t_i)$. Binaarsele haigestumise tunnusele on sobitatud täiend-log-log mudel, kus parameetrite väärtusteks hinnati $\hat{\beta}_0 = -6{,}28$, $\hat{\beta}_t = 0{,}0561$ ja $\hat{\beta}'_g = 0{,}404$. Seega

$$f_2(t) = \exp(-6{,}28 + 0{,}0561t).$$

Näeme, et funktsioonide f_1 ja f_2 väärtused on genereeritud liitumisvanuste korral sarnased ja ka genotüübile vastava parameetri hinnang täiend-log-log mudelist on sarnane õigele parameetrile. Seega, praktikas võime öelda, et kui $f_1 \approx f_2$ vaatluse all olevate jälgimisaegade korral, siis ka $\beta_g \approx \beta'_g$. Sedasama seost uurime aga ka simulatsiooniuuringus.

2.4 Meta-analüüs

Meta-analüüsiks ehk analüüside analüüsiks nimetatakse enamasti erinevate uuringute tulemuste statistilist analüüsi eesmärgiga neid tulemusi koondada ja üldistada (Glass, 1976). Meta-analüüsi kasutatakse peamiselt selliste uuringute koondamiseks, kus on käsitletud samu teemasid, näiteks uuritud mingi kindla ravi efektiivsust või geenivariantide mõju teatud haigusele.

Üksikutel geneetilistel seoseuuringutel ei ole tihti piisavalt võimsust, et tuvastada väikese sageduse ja efektiga geenivariante. Tänu mitme uuringu kombineerimisele on võimalik saavutada suuremat statistilist võimsust ja tuvastada selliseid seoseid, mis üksikute analüüside puhul jääksid leidmata.

Selles töös kasutatakse meta-analüüsi, kombineerimaks binaarsete mudelite ja võrdeliste riskide

mudelite hinnanguid vastavalt jälgimiseelsete ja jälgimisaegsete juhtude analüüsist.

Lihtsaim meta-analüüsi tüüp on fikseeritud mõjudega meta-analüüs, mille puhul eeldatakse, et igas kaasatavas uuringus on mingi uuritava töötluse (näiteks mõne ravimeetodi või geenivariandi) tegelik mõju β samasugune.

Olgu meil vaatluse all n sõltumatut uuringut, kus on leitud hinnangud $\hat{\beta}_i$ $(i = 1, \ldots, n)$ tegelikule mõjule β. Olgu σ_i nende hinnangute standardvigade hinnangud ja tähistame $w_i = 1/\sigma_i^2$. Siis fikseeritud mõjudega mudeli korral avaldub kombineeritud hinnang tegelikule mõjule üksikhinnangute kaalutud keskmisena:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} w_i \hat{\beta}_i}{\sum_{i=1}^{n} w_i},$$

kus i-nda $(i = 1, \ldots, n)$ uuringu kaaluks on $w_i / \sum_{j=1}^{n} w_j$. Seega, mida täpsem on uuringus leitud hinnang ehk mida väiksem on tema standardvea hinnang, seda suurema kaaluga ta kombineeritud hinnangusse panustab. Selline kaalude valik tagab, et meta-analüüsi hinnang on nihketa (eeldusel, et esialgsed hinnangud $\hat{\beta}_i$ $(i = 1, \ldots, n)$ on nihketa) ja vähima võimaliku dispersiooniga. Kombineeritud hinnangu $\hat{\beta}$ dispersioon on $D(\hat{\beta}) = (\sum_{i=1}^{n} w_i)^{-1}$.

Selles töös kasutamegi fikseeritud mõjudega meta-analüüsi ja selle läbiviimiseks kasutame R-i paketti *metafor* (Viechtbauer, 2010).

3 Simulatsiooniuuring

Erinevate analüüsimeetodite – Coxi ja martingaalijääkide mudeli ning logistilise ja täiend-log-log mudeli – võrdlemiseks viiakse läbi simulatsiooniuuring. Et võrdeliste riskide mudel on kirja pandud riskifunktsioonide kaudu, kuid elukestusanalüüsi rakendamiseks on vaja teada konkreetseid elukestuseid, siis kõigepealt kirjeldatakse, kuidas genereerida haigestumisandmeid nii, et need vastaksid etteantud võrdeliste riskide mudelile.

3.1 Haigestumisandmete simuleerimine

Simulatsiooniuuringu läbiviimiseks tuleb genereerida haigestumist kirjeldavad andmed. Haigestumise vanuse jaotusena kasutame Weibulli jaotust. Olgu n valimi suurus. Algoritm haigestumisandmete genereerimiseks nii, et efektialleel suurendaks või vähendaks haigestumise riskimäära vastavalt kindlaksmääratud riskimäärade suhtele, on järgmine: iga indiviidi $i = 1, \ldots, n$ korral

1. genereeritakse genotüüp G_i võimalike väärtustega $0, 1, 2$ (mis tähistavad efektialleelide arvu), kasutades binoomjaotust $\mathrm{B}(2, \mathrm{MAF})$, kus MAF on fikseeritud efektialleeli (harvema alleeli) sagedus;

2. genereeritakse elukestus T_i Weibulli jaotusest $W(\alpha, \beta)$;

3. genereeritakse haigestumise vanus H_i Weibulli jaotusest $W(a, b_g)$, kus b_g on b_0, b_1 või b_2 vastavalt sellele, kas inimese genotüüp on $0, 1$ või 2;

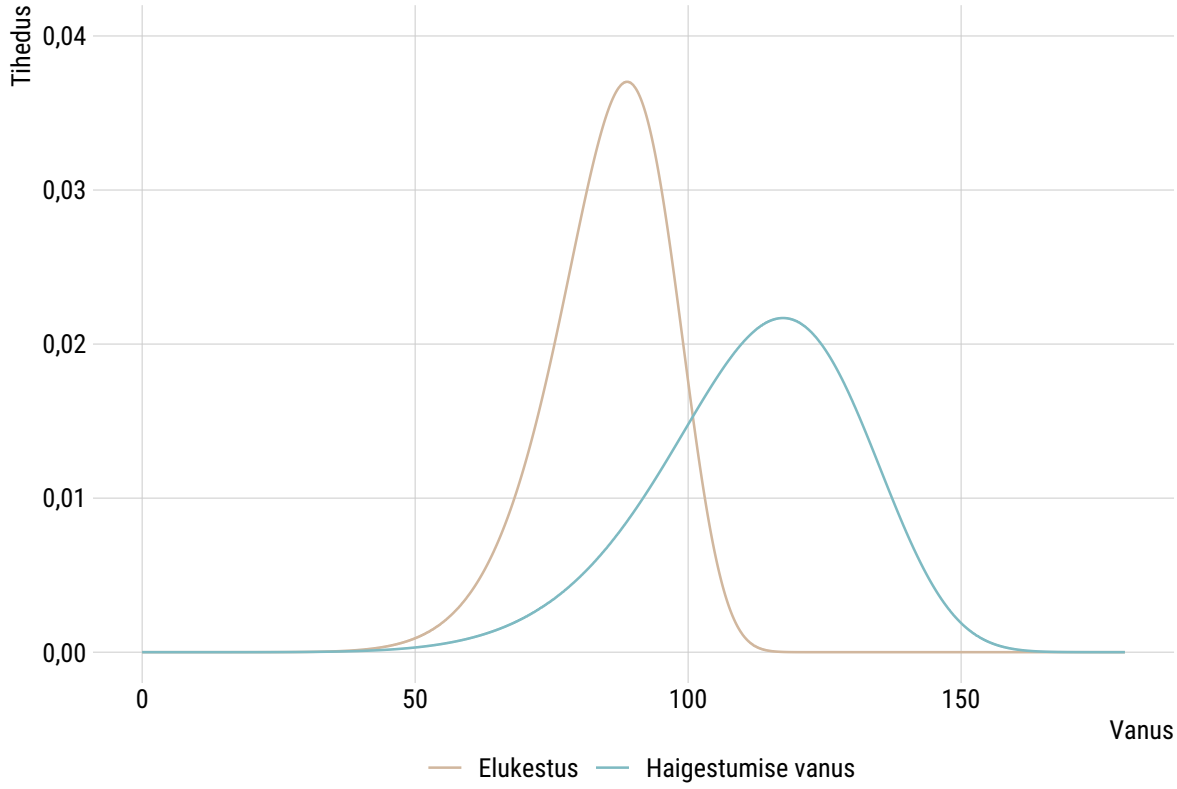
4. valitakse vaatlusvanuseks D_i minimaalne haigestumise vanusest ja elukestusest: $D_i = \min\{H_i, T_i\}$;

5. märgitakse ta haigeks, kui tema vaatlusvanus on haigestumise vanus, seega haigestumist kirjeldab indikaator $Y_i = I(H_i \leqslant T_i) = I(D_i = H_i)$.

Selliseks simuleerimiseks tuleb kõigepealt leida sobivad Weibulli jaotuse parameetrid α ja β elukestuse jaoks ning a, b_0, b_1 ja b_2 haigestumise vanuse jaoks.

Üks võimalus sobivate parameetrite leidmiseks on kasutada suurima tõepära meetodit ja hinnata need parameetrid näiteks TÜ EGV andmetelt. Selleks saab kasutada R-i paketi *survival* funktsiooni *survreg*. Selles töös kasutamegi elukestuse genereerimiseks geenivaramu andmetelt hinnatud parameetreid $\alpha = 9$ ja $\beta = 90$.

Joonisel 5 on näitena kujutatud elukestuse ja haigestumise vanuse tihedusfunktsioonid juhul, kui elukestuse parameetrid on $\alpha = 9$ ja $\beta = 90$ ning haigestumise vanuse parameetrid on $a = 7$ ja $b_0 = 120$. Selliste parameetrite korral on haigestumise mediaanvanus umbes 82 aastat ja haiguse levimus umbes 11%.



Joonis 5. Elukestuse ja haigestumise vanuse tihedusfunktsioonid, kui elukestus on Weibulli jaotusega $W(9, 90)$ ja haigestumise vanus on Weibulli jaotusega $W(7, 120)$

Haigestumise vanuse parameetrid on samuti võimalik hinnata huvipakkuva haiguse andmete põhjal. Et aga simulatsiooniuuringu käigus soovime näha ka seda, kuidas hinnangud käitu-

vad haiguse erinevate levimuste korral, siis arvutame need parameetrid ise. Selleks fikseerime esmalt simuleeritava haiguse mediaanvanuse m haigestumise hetkel ja soovitava haiguse levimuse p baasgrupis ehk nende inimeste seas, kellel $G_i = 0$.

Kuna haigestumise vanuse jaotuse parameetrid sõltuvad inimese genotüübist, siis leitakse esmalt Weibulli jaotuse baasparameetrid a ja b_0, millega genereeritakse haigestumise vanus neile, kellel ei ole ühtegi efektialleeli ehk kelle genotüüp on $G_i = 0$. Parameetrid b_1 ja b_2 saab seejärel arvutada fikseeritud riskimäärade suhte kaudu.

Olgu elukestust tähistav juhuslik suurus $T \sim W(\alpha, \beta)$ ja sellest sõltumatu haigestumise vanust baasgrupis tähistav juhuslik suurus $H_0 \sim W(a, b_0)$ jaotus- ning tihedusfunktsioonidega vastavalt F_T, f_T ja F_{H_0}, f_{H_0}.

Kuna haigestumise vanus ja elukestus genereeritakse igale vaatlusele ning haigeks märgitakse need inimesed, kelle puhul on simuleeritud haigestumise vanus väiksem kui simuleeritud elukestus, siis määrab baasgrupi levimuse p tõenäosus $P(H_0 < T)$. Avaldades selle võrduse elukestuse ja haigestumise vanuse jaotus- ning tihedusfunktsioonide kaudu, leiame esimese seose a ja b_0 vahel:

$$\begin{aligned}
p &= P(H_0 < T) = \\
&= P(H_0 - T < 0) = \\
&= \int_0^\infty \int_0^t f_{H_0}(h) f_T(t) \,\mathrm{d}h \,\mathrm{d}t = \\
&= \int_0^\infty F_{H_0}(t) f_T(t) \,\mathrm{d}t = \\
&= \int_0^\infty \left(1 - \exp\left\{-\left(\frac{t}{b_0}\right)^a\right\}\right) \frac{\alpha}{\beta} \left(\frac{t}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\} \mathrm{d}t.
\end{aligned}$$

Haigestumise mediaanvanuse puhul võtame arvesse ainult neid inimesi, kes said haiguse ehk kelle haigestumise vanus on väiksem kui elukestus. Seega, kui soovime, et haigestumise mediaanvanus baasgrupis oleks m, peab kehtima $P(H_0 < m | H_0 < T) = 0{,}5$.

Kasutades tingliku tõenäosuse valemit, saame nüüd avaldada ka teise seose parameetrite a ja b_0

vahel:

$$
\begin{aligned}
0{,}5 &= P(H_0 < m | H_0 < T) = \\
&= \frac{P(H_0 < m, H_0 < T)}{P(H_0 < T)} = \\
&= \frac{1}{p} \int_0^m \int_h^\infty f_T(t) f_{H_0}(h)\, \mathrm{d}t\, \mathrm{d}h = \\
&= \frac{1}{p} \int_0^m (1 - F_T(h)) f_{H_0}(h)\, \mathrm{d}h = \\
&= \frac{1}{p} \int_0^m \exp\left\{ -\left(\frac{h}{\beta}\right)^\alpha \right\} \frac{a}{b_0} \left(\frac{h}{b_0}\right)^{a-1} \exp\left\{ -\left(\frac{h}{b_0}\right)^a \right\} \mathrm{d}h.
\end{aligned}
$$

Oleme saanud kaks võrrandit a ja b_0 määramiseks. Arvestades seda, et tegu on keeruliste integraalidega, millel üldjuhul lihtsat analüütilist kuju ei leidu, kasutame a ja b_0 määramiseks numbrilist integreerimist. Selles töös on seda tehtud programmi MATLAB abil, kasutades mittelineaarsete võrrandisüsteemide numbriliseks lahendamiseks mõeldud *fsolve* funktsiooni (MATLAB Optimization Toolbox, 2021). Tabelis 1 on näitena toodud leitud lahendid haigestumise mediaanvanuse $m = 60$ korral ja viie erineva baasgrupi levimuse p korral.

Tabel 1. Haigestumise vanust kirjeldava Weibulli jaotuse kujuparameeter a ja skaalaparameeter b_0 erinevate baaslevimuste korral, kui haigestumise mediaanvanus baasgrupis on 60

Baaslevimus p	a	b_0
0,01	1,9254	936,7213
0,05	1,9562	391,9832
0,10	1,9974	264,8140
0,20	2,0897	175,7165
0,30	2,1989	136,7687

Selleks, et haigestumine sõltuks genotüübist – efektialleeliga isikute seas oleks haigestumise riskimäär kas suurem või väiksem kui baasgrupis – tuleb haigestumise vanust kirjeldava Weibulli jaotuse skaalaparameetrit muuta vastavalt efektialleelide arvule ja fikseeritud riskimäärade suhtele.

Selleks, et riskimäärade suhe ühe võrra suurema arvu efektialleelide suhtes oleks HR, peab

võrduse (5) põhjal kehtima

$$b_1 = b_0 \left\{\exp(\beta_g \cdot 1)\right\}^{-\frac{1}{a}}$$

ja analoogiliselt

$$b_2 = b_1 \left\{\exp(\beta_g \cdot 1)\right\}^{-\frac{1}{a}} = b_0 \left\{\exp(\beta_g \cdot 2)\right\}^{-\frac{1}{a}},$$

kus β_g on $\log(\mathrm{HR})$. Seega, üldjuhul genereeritakse i-ndale indiviidile haigestumise vanus Weibulli jaotusest

$$W\left(a, b_0 \left\{\exp(\beta_g G_i)\right\}^{-\frac{1}{a}}\right),$$

kus $G_i \in \{0, 1, 2\}$ on inimese genotüüp.

Kood selliste andmete simuleerimiseks R-is on toodud lisas 3.

3.2 Simulatsiooniplaan

Simulatsiooniuuringus käsitleme kolme olukorda. Esimesel juhul eeldame, et meil on olemas sünnikohort, kus kõik inimesed on vaatluse all sünnist kuni surmani. Seega teame täpselt iga inimese haigestumise vanust ja elukestust, ning iga haiguse saanud inimene on jälgimisaegne juht. See on nii-öelda ideaalne olukord, mida päriselus tegelikult ette ei tule. Teisel juhul eeldame, et inimesed liituvad geenivaramuga oma elu jooksul ja osa juhtudest on seega jälgimiseelsed, osa jälgimisaegsed. Kolmanda stsenaariumi puhul eeldame nagu teiseski stsenaariumis, et geenivaramuga liitutakse elu jooksul, kuid enam me jälgimiseelsetel ja jälgimisaegsetel juhtudel vahet ei tee.

Simulatsioon viiakse läbi erinevate efektisuuruste $\mathrm{HR} \in \{0{,}9; 0{,}95; 1; 1{,}05; 1{,}1; 1{,}2; 1{,}5\}$, harvema alleeli sageduste $\mathrm{MAF} \in \{0{,}05; 0{,}1; 0{,}2; 0{,}3; 0{,}4; 0{,}5\}$, haigestumise mediaanvanuse $m = 60$ ja haiguse baasgrupi levimuste $p \in \{0{,}05; 0{,}1; 0{,}2\}$ korral. Valimimahuna kasutati $N = 100\,000$, mis on ligikaudu võrdne TÜ EGV kohordi suurusega, kust on välja jäetud lähisuguluses olevad geenidoonorid. Erinevate mudelite tulemuste võrdlemiseks hinnati iga parameetrite kombinatsiooni korral iga stsenaariumi ja iga mudeli jaoks $n_{\text{sim}} = 1000$ simulatsiooni

keskmisena hinnangute empiiriline nihe:

$$\frac{1}{n_{\text{sim}}}\sum_{k=1}^{n_{\text{sim}}}(\hat{\beta}_k - \log(\text{HR})),$$

ruutjuur keskmisest ruutveast (RMSE, *root mean square error*):

$$\sqrt{\frac{1}{n_{\text{sim}}}\sum_{k=1}^{n_{\text{sim}}}(\hat{\beta}_k - \log(\text{HR}))^2}$$

ja võimsus, kasutades olulisuse nivood $0{,}05$:

$$\frac{1}{n_{\text{sim}}}\sum_{k=1}^{n_{\text{sim}}} I(P_k < 0{,}05),$$

kus $\hat{\beta}_k$ ja P_k on vastavalt k-nda ($k = 1, \ldots, n_{\text{sim}}$) valimi puhul leitud hinnang efektisuuruse logaritmile ja mudelist saadud statistilise testi p-väärtus. Juhul kui tegelik efektisuurus on $\text{HR} = 1$, saab viimase valemi järgi arvutada empiirilise I liiki vea tõenäosuse.

Kirjeldame nüüd täpsemalt, kuidas andmete genereerimine ja analüüs erinevate stsenaariumite puhul läbi viiakse. Olgu meil genereeritud andmed nii, nagu on kirjeldatud peatüki 3.1 alguses. Seega on meil iga indiviidi $i = 1, \ldots, n$ jaoks olemas genotüüp G_i, elukestus T_i ja genotüübist sõltuv haigestumise vanus H_i.

Esimene stsenaarium

Esimese stsenaariumi puhul hinnatakse Coxi võrdeliste riskide mudel, kasutades ajaskaalana inimese vanust haigestumise ajal ja ainsa argumenttunnusena kaasatakse genotüüp. Et inimesed liitusid sündides, siis inimese vanus haigestumise ajal on samal ajal ka aeg liitumisest haigestumiseni. Lisaks lähendatakse sama mudelit martingaalijääkide abil. Selleks sobitatakse ilma kovariaatideta Coxi mudel ja hinnatakse sellelt skaleeritud martingaalijäägid, millele sobitatakse seejärel lineaarne mudel, kus ainsaks seletavaks tunnuseks on genotüüp.

Teine stsenaarium

Teise ja kolmanda stsenaariumi puhul genereeritakse igale inimesele lisaks ka sünniaasta S_i ühtlasest jaotusest $U(1920, 1980)$ ja liitumisaasta L_i ühtlasest jaotusest $U(2000, 2020)$. Sellise genereerimise puhul jäetakse andmestikku alles vaid need inimesed, kes elasid vähemalt liitumiseni ehk kelle elukestus on suurem kui liitumisvanus: $T_i > L_i - S_i$. Samuti valitakse nii-öelda katkestusaeg K, millest alates on kõik inimesed tsenseeritud. Geenivaramu andmete analüüsimisel on katkestusajaks tavaliselt kuupäev, mil saabusid viimased andmed meditsiiniandmete allikatest. Simulatsiooniuuringus on katkestusajaks valitud aasta 2050 – inimesed, kes jäid haigeks või surid pärast seda aastat, on tsenseeritud, ja nende vaatlusvanus on vanus aastal 2050. Seega on inimese vaatlusvanuseks minimaalne haigestumise vanusest, elukestusest ja vanusest katkestusajal: $D_i = \min\{H_i, T_i, K - S_i\}$, ja haigeks märgitakse need, kelle vaatlusvanus on nende haigestumise vanus, seega haigestumist kirjeldab indikaator $Y_i = I(D_i = H_i)$.

Teise stsenaariumi puhul jagatakse andmestik enne analüüsi haigestumis- ja liitumisaja järgi kaheks. Ühte andmestikku jäävad kõik jälgimiseelsed juhud ja juhuslikult valitud terved kontrollid nii, et iga juhu kohta on andmestikus neli kontrolli. Terveteks kontrollideks on need, kes ei olnud haigestunud enne ega pärast liitumist. Kontrollide ja juhtude suhe 4:1 on juht-kontrolluuringutes laialt kasutusel, sest on näidatud, et sellest suurema suhte kasutamine testide statistilist võimsust oluliselt ei paranda (Hong ja Park, 2012). Teise andmestikku jäävad kõik jälgimisaegsed juhud ja kõik need terved kontrollid, keda ei valitud esimesse andmestikku.

Esimesele andmestikule hinnatakse nii logistilise regressiooni mudel kui täiend-log-log mudel, kus funktsioontunnuseks on haiguse olemasolu kirjeldav binaarne tunnus ja seletavaks tunnuseks genotüüp. Kovariaadina lisatakse mudelisse ka inimese liitumisvanus. Liitumisvanus on see vanus, mille korral me juhtude puhul teame, et nad olid selleks ajaks haigeks jäänud, kontrollide puhul teame, et selles vanuses olid nad terved.

Teisele ehk jälgimisaegsete juhtude andmestikule hinnatakse Coxi võrdeliste riskide mudel, kasutades ajaskaalana aega liitumisest: $D_i - (L_i - S_i)$. Seletavaks tunnuseks on genotüüp ja kovariaadina on arvesse võetud liitumisvanus. Seda mudelit lähendatakse ka martingaali-

jääkide abil, sobitades esmalt Coxi mudeli, kuhu kovariaadina on lisatud vaid liitumisvanus. Selle mudeli skaleeritud martingaalijääkidele hinnatakse seejärel lineaarne mudel, kus seletavaks tunnuseks on genotüüp.

Kahe andmestiku analüüsi tulemused meta-analüüsitakse kokku, kasutades fikseeritud mõjudega meta-analüüsi mudelit. Nii Coxi mudeli kui martingaalijääkide mudeli hinnangud kombineeritakse nii logistilise kui täiend-log-log mudeli hinnangutega, seega saadakse kokku neli meta-analüüsi hinnangut.

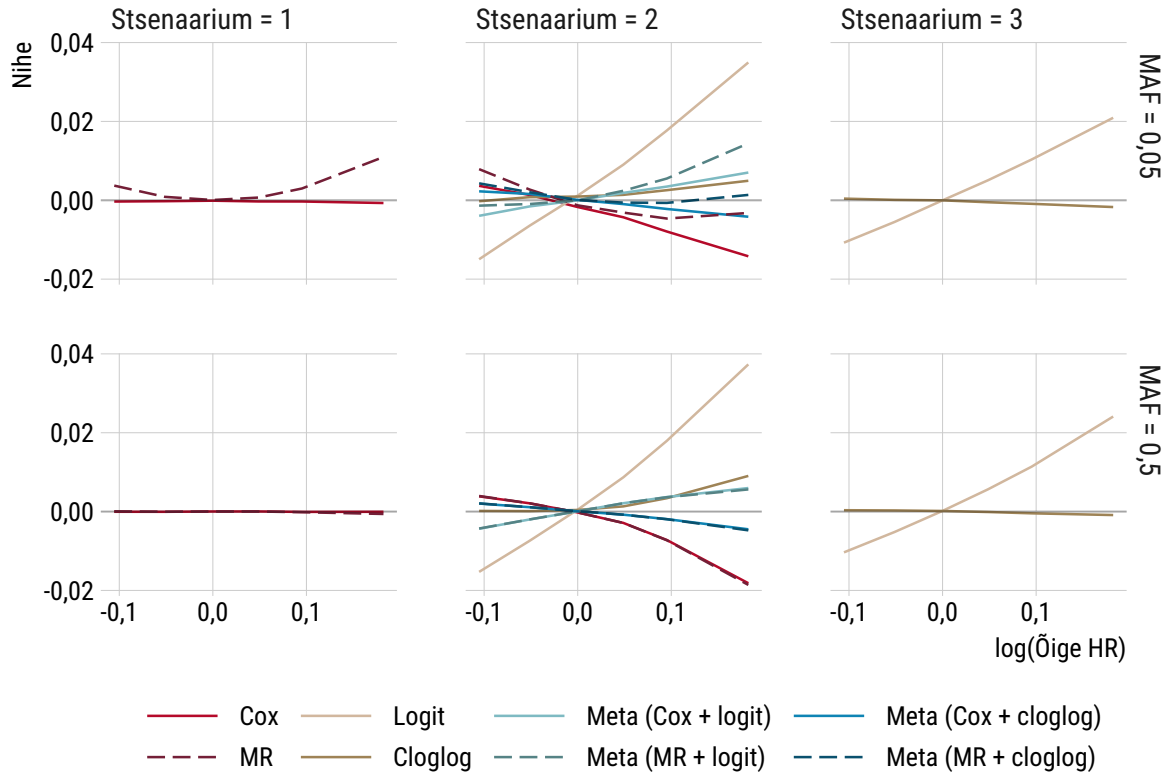
Kolmas stsenaarium

Kolmandas stsenaariumis eeldame, et osa inimesi on jälgimiseelsed, osa jälgimisaegsed juhud, kuid analüüsi läbi viies me neil juhtudel vahet ei tee. Selle stsenaariumi puhul on analoogiliselt teise stsenaariumiga alles jäetud vaid need inimesed, kes elasid vähemalt liitumiseni, ja arvesse on võetud katkestusaega. Enam aga indiviide liitumisaja järgi kaheks ei jagata. Andmetele hinnatakse logistiline mudel ja täiend-log-log mudel, kus uuritavaks tunnuseks on haiguse olemasolu kirjeldav binaarne tunnus ja kovariaatidena on mudelis genotüüp ja inimese sünniaeg.

3.3 Tulemused

Enne simulatsiooniuuringu tulemuste kirjeldamist tuletame meelde, et nagu on kirjeldatud peatükis 3.1, on andmete genereerimise aluseks riskimäärade suhe (HR). Samal ajal kasutatakse andmete analüüsimiseks ka logistilist regressiooni, mis hindab šansside suhet (OR) – see on riskimäärade suhtest erinev näitaja. Šansside suhe kirjeldab riskifaktori mõju haigestumise šansile, riskimäärade suhe aga võtab arvesse riskifaktori mõju ka haigestumise ajale. Seega tuleb meeles pidada, et oodatavalt on logistilist regressiooni kasutades leitud hinnangud erinevad tegelikust efektisuurusest, kuid see ei tähenda, et need hinnangud on klassikalises mõttes nihkega.

Joonisel 6 on kujutatud simulatsioonide tulemusel leitud hinnangute nihked kõigi kolme stsenaariumi korral.



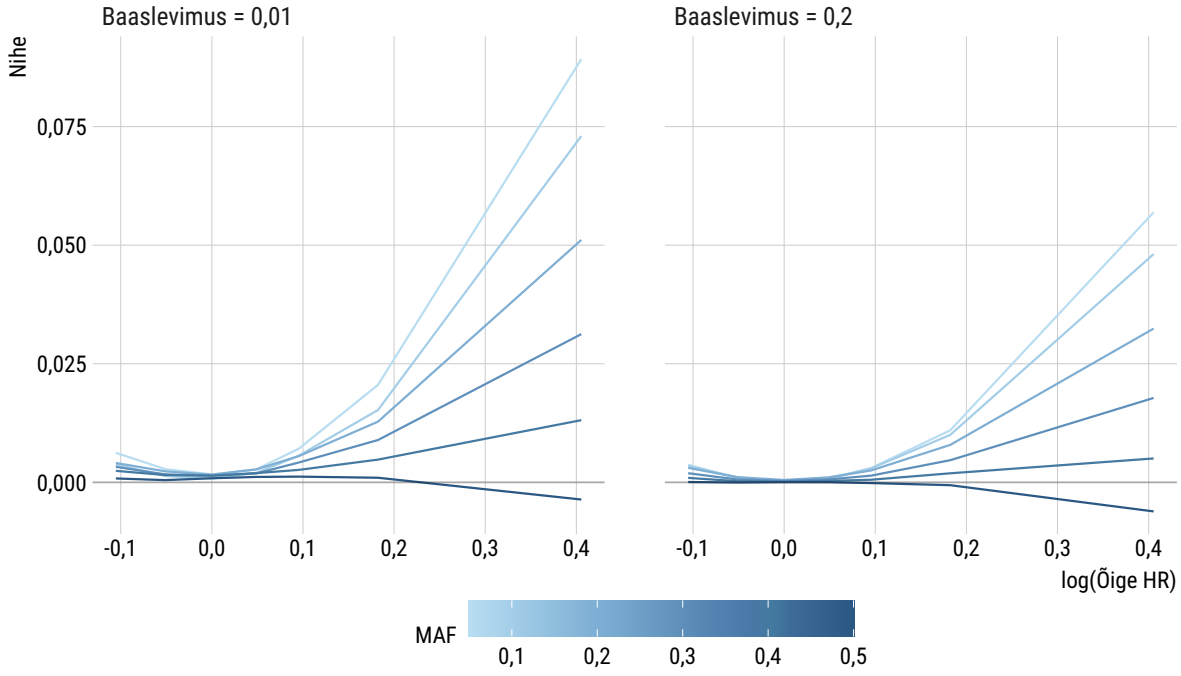
Joonis 6. Simulatsiooniuuringus leitud hinnangute nihked kõigi kolme stsenaariumi, baaslevimuse $p = 0{,}2$, harvema alleeli sageduste $\mathrm{MAF} \in \{0{,}05; 0{,}5\}$ ja riskimäärade suhete $\mathrm{HR} \in \{0{,}9; 0{,}95; 1; 1{,}05; 1{,}1; 1{,}2\}$ korral. Logit - logistiline mudel, cloglog - täiend-log-log mudel, MR - martingaalijääkide mudel

Oodatavalt olid tulemused parimad esimeses stsenaariumis hinnatud Coxi mudeli korral: õiget efektisuurust hinnatakse nihketa iga HR korral, see on nii ka ülejäänud vaadeldud baaslevimuste ja harvema alleeli sageduste puhul. Huvi pakub aga teise ja kolmanda stsenaariumi tulemuste omavaheline võrdlus. Kolmanda stsenaariumi puhul näeme, et täiend-log-log mudel hindab väga hästi õiget efektisuurust ehk riskimäärade suhet kõigi parameetrite kombinatsioonide korral. Oodatavalt on logistilise regressiooni hinnangud simuleerimise aluseks olevatest riskimäärade suhetest seda erinevamad, mida erinevam on vastav riskimäärade suhte logaritm nullist: kui õige riskimäärade suhe on 1, siis on hinnang nihketa, kui õige riskimäärade suhe on väiksem kui 1, siis on efektisuurus alahinnatud ja kui õige riskimäärade suhe on suurem kui 1, on efektisuurus ülehinnatud.

Teise stsenaariumi puhul on olulised just metahinnangud (joonisel sinistes toonides), ülejäänud

hinnangud on lihtsalt vahesammud metahinnangute arvutamiseks. Paneme tähele, et teise stsenaariumi puhul on ka Coxi (ja martingaalijääkide) mudeli hinnangud nihkega. Kui tegelik efektisuurus on ühest väiksem, on tegemist väikese ülehinnanguga, kui tegelik efekt on ühest suurem, on tegemist alahinnanguga. Metahinnangute puhul näeme analoogiliselt kolmanda stsenaariumiga, et logistilisel regressioonil põhinevad metahinnangud alahindavad ühest väiksema efektisuuruse puhul õiget efekti ja ühest suurema efektisuuruse puhul ülehindavad õiget efekti. Täiend-log-log regressioonil põhinevate metahinnangute puhul on see seos vastupidine.

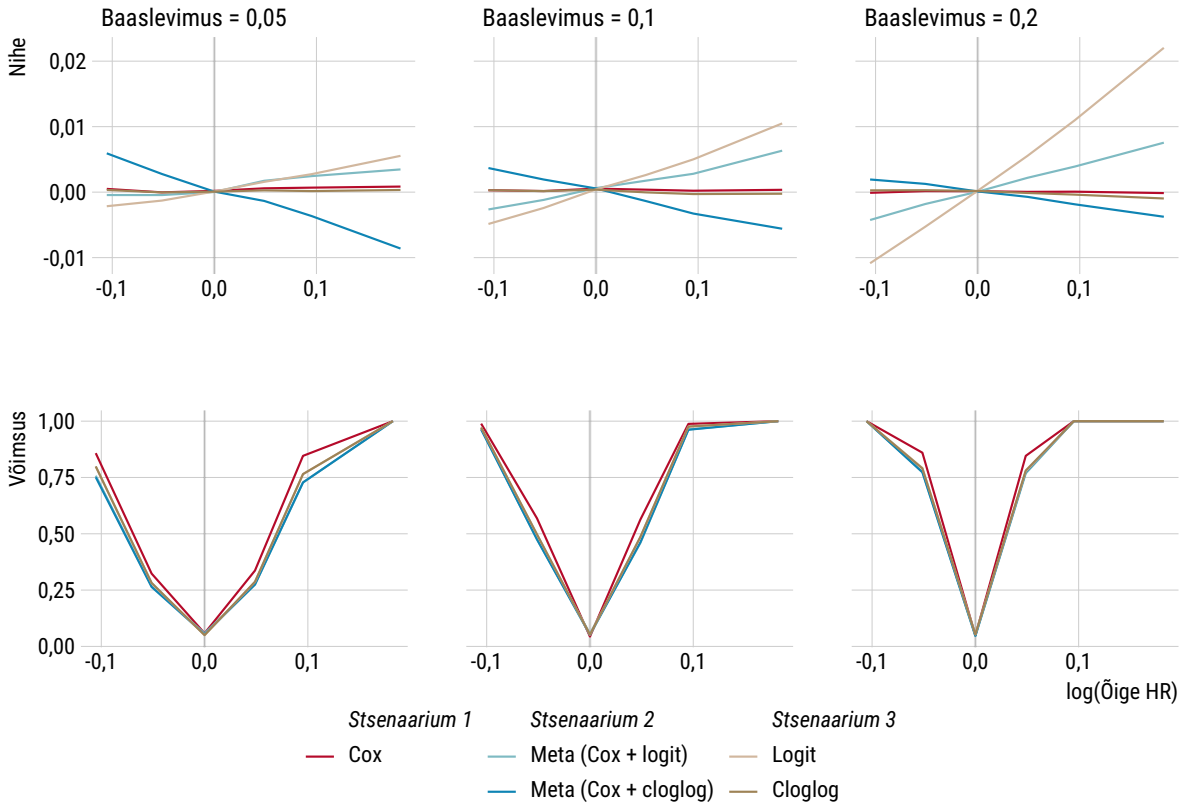
Jooniselt näeme ka, et erinevate MAF-ide puhul käituvad erinevate meetodite hinnangud sarnaselt, vaid martingaalijääkide hinnangu nihet mõjutab MAF oluliselt: suurema MAF-i korral on martingaalijääkide kaudu leidud hinnangud väga sarnased Coxi mudeli hinnangutega. Sedasama näeme tegelikult ka kõigi teiste uuritud levimuste ja vaadeldud riskimäärade suhete korral: suure MAF-i korral lähendab martingaalijääkide meetod väga hästi Coxi võrdeliste riskide mudelit. Joonisel 7 on kujutatud martingaalijääkide meetodi hinnangud esimese stsenaariumi korral.



Joonis 7. Simulatsiooniuuringus leitud martingaalijääkide meetodi hinnangute nihked esimese stsenaariumi, baaslevimuste $p \in \{0{,}01; 0{,}2\}$, harvema alleeli sageduste $\mathrm{MAF} \in \{0{,}05; 0{,}1; 0{,}2; 0{,}3; 0{,}4; 0{,}5\}$ ja riskimäärade suhete $\mathrm{HR} \in \{0{,}9; 0{,}95; 1; 1{,}05; 1{,}1; 1{,}2; 1{,}5\}$ korral.

Samasugust seost näeme ka teise stsenaariumi puhul: suurema MAF-i korral on martingaalijääkide hinnangud lähedased Coxi mudeli hinnangutele kõigi vaadeldud baaslevimuste väärtuste korral, ja sama kehtib ka vastavate metahinnangute korral. Samal ajal on ühelähedaste efektisuuruste puhul (mis on ülegenoomsetes seoseuuringutes tavapärased (Zemunik ja Borask, 2011; Scott *et al.*, 2017)) martingaalijääkide meetodi hinnangute erinevus Coxi mudelite hinnangutest väike ka madalate MAF-ide puhul. Kuna aga martingaalijääkide hinnangute uurimine ei ole selle töö peamine eesmärk, siis edaspidi keskendume Coxi mudeli hinnangutele ja nende kaudu arvutatud metahinnangutele.

Joonisel 8 on kujutatud, kuidas sõltuvad erinevate meetodite hinnangute nihe ja võimsus baaslevimusest ning riskimäärade suhtest.



Joonis 8. Simulatsiooniuuringus leitud hinnangute nihked ja võimsused harvema alleeli sageduse $\mathrm{MAF} = 0{,}1$, baaslevimuste $p \in \{0{,}05; 0{,}1; 0{,}2\}$ ja riskimäärade suhete $\mathrm{HR} \in \{0{,}9; 0{,}95; 1; 1{,}05; 1{,}1; 1{,}2\}$ korral. Logit - logistiline mudel, cloglog - täiend-log-log mudel

Jooniselt näeme, et üldiselt käituvad erinevate meetodite hinnangud kõigi vaadeldud levimuste

puhul sarnaselt. Esimese stsenaariumi Coxi mudel annab kõigi levimuste puhul nihketa hinnangu ja suurima võimsuse. Huvipakkuvatest meetoditest on ka teise stsenaariumi täiend-log-log mudeli hinnangud kõigi baaslevimuste puhul nihketa. Oodatult on võimsus iga meetodi puhul seda suurem, mida suurem on baaslevimus. Näeme ka, et võimsuse poolest on huvipakkuvad meetodid väga sarnased: teise stsenaariumi meta-analüüsi meetodite võimsus on küll iga levimuse puhul teistest madalam, kuid see erinevus on väike.

Kui õige riskimäärade suhe on 1, siis annavad kõik uuritud meetodid nihketa hinnangu ja empiiriline I liiki vea tõenäosus on ligikaudu $0{,}05$. Iga levimuse puhul näeme oodatult ka seda, et logistilisel regressioonil põhinevad meetodid (teise stsenaariumi *Meta (Cox + logit)* ja kolmanda stsenaariumi *Logit*) ala- või ülehindavad õiget riskimäärade suhet vastavalt sellele, kas õige HR on väiksem või suurem ühest, ning erinevus õigest riskimäärade suhtest on seda suurem, mida suurem on baaslevimus.

Tabelis 2 on ka simulatsiooniuuringu numbrilised tulemused kõigi stsenaariumite, baaslevimuse $p = 0{,}2$ ja harvema alleeli sageduse $\mathrm{MAF} = 0{,}05$ korral.

Taas näeme, et esimese stsenaariumi puhul on hinnangud oodatult nihketa, vähima RMSE-ga ja suurima võimsusega. Huvipakkuvatest meetoditest annab iga näitaja puhul parimaid tulemusi aga kolmas stsenaarium ja täiend-log-log mudel. Riskimäärade suhte $\mathrm{HR} = 1$ korral on kõigi meetodite hinnangud nihketa, sarnase RMSE-ga ning I liiki viga on $0{,}05$ lähedal. Ühelähedaste riskimäärade suhete puhul on nihked iga meetodi puhul väikesed. Oodatavalt näeme, et mida erinevam on riskimäärade suhe ühest, seda suurem on logistilist regressiooni kasutavate meetodite hinnangute nihe võrreldes ülejäänud meetoditega. Võimsuse poolest on kõik meetodid sarnased, siiski on kolmanda stsenaariumi täiend-log-log mudeli võimsus kõikide riskimäärade suhete korral teiste huvipakkuvate mudelite võimsusest veidi suurem.

Tabel 2. Simulatsioonide tulemused: keskmine parameetri hinnang, hinnang nihkele, RMSE-le ja võimsusele, baaslevimuse $p = 0{,}2$ ja harvema alleeli sageduse $\mathrm{MAF} = 0{,}05$ korral

Õige HR (log(Õige HR))	Stsenaarium	Mudel	$\hat{\mathrm{E}}[\hat{\beta}]$	Nihe	RMSE	Võimsus
0,9 (−0,1054)	1	Cox	−0,106	0,000	0,024	0,997
	2	Meta (Cox + cloglog)	−0,103	0,002	0,026	0,982
		Meta (Cox + logit)	−0,109	−0,004	0,028	0,982
	3	Cloglog	−0,105	0,000	0,026	0,986
		Logit	−0,116	−0,011	0,031	0,986
1 (0)	1	Cox	0,000	0,000	0,023	0,052
	2	Meta (Cox + cloglog)	0,000	0,000	0,025	0,044
		Meta (Cox + logit)	0,000	0,000	0,027	0,048
	3	Cloglog	0,000	0,000	0,025	0,053
		Logit	0,000	0,000	0,028	0,053
1,05 (0,0488)	1	Cox	0,048	0,000	0,022	0,586
	2	Meta (Cox + cloglog)	0,048	−0,001	0,025	0,503
		Meta (Cox + logit)	0,051	0,002	0,026	0,494
	3	Cloglog	0,048	−0,001	0,025	0,527
		Logit	0,054	0,005	0,028	0,523
1,1 (0,0953)	1	Cox	0,095	0,000	0,022	0,986
	2	Meta (Cox + cloglog)	0,093	−0,002	0,024	0,967
		Meta (Cox + logit)	0,099	0,003	0,026	0,966
	3	Cloglog	0,094	−0,001	0,024	0,975
		Logit	0,106	0,010	0,029	0,975
2 (0,6931)	1	Cox	0,693	0,000	0,017	1,000
	2	Meta (Cox + cloglog)	0,675	−0,018	0,026	1,000
		Meta (Cox + logit)	0,722	0,029	0,036	1,000
	3	Cloglog	0,688	−0,005	0,020	1,000
		Logit	0,808	0,115	0,117	1,000

Logit - logistiline mudel, cloglog - täiend-log-log mudel, RMSE - ruutjuur keskmisest ruutveast

Eelnevat kokku võttes näeme, et keerulisem meetod – jälgimiseelsete ja jälgimisaegsete juhtude eraldi analüüsimine ning seejärel saadud hinnangute kombineerimine – ei ole parem lihtsast meetodist, kus kõiki juhtusid analüüsitakse koos binaarse mudeliga. Binaarsete mudelite vahel valides võiks eelistada täiend-log-log regressiooni: selle kaudu saame hinnata efektisuurused, mis on interpreteeritavad samamoodi nagu Coxi mudeli efektid ehk riskimäärade suhted.

4 Geenivaramu andmete analüüs T2D näitel

Selles peatükis rakendame uuritud meetodeid TÜ EGV geenidoonorite andmetel, et uurida seoseid teist tüüpi diabeedi ja geenivariantide vahel. Esmalt anname ülevaate analüüsis kasutatavatest geenivaramu andmetest ja seejärel kirjeldame analüüsi läbiviimist ning tulemusi.

4.1 Andmete kirjeldus

Diagnooside andmed

TÜ EGV geenidoonorite diagnooside andmestik on kokku pandud info põhjal, mis on pärit erinevatest meditsiiniallikatest – Eesti Haigekassast, Tartu Ülikooli Kliinikumist, Põhja-Eesti Regionaalhaiglast, E-tervise andmetest ja surma- ning vähiregistrist – ja geenidoonorite enda täidetud küsimustikest. Geenidoonorite viimased diagnooside andmed meditsiiniallikatest on pärit 2019. aasta detsembrist.

Diagnooside defineerimiseks kasutatakse ICD-10 koode. ICD-10 (*The International Classification of Diseases, Tenth Revision*) on rahvusvahelise haiguste klassifikatsiooni kümnes versioon, tänu millele on võimalik koondada rahvusvahelist statistikat haiguste ja surmapõhjuste kohta. Kõikidel haigustel on oma ICD-10 kood, mis koosneb ühest tähest ja kahest numbrist, millele võib järgneda ka täpsustav arv. Arstid märgivad patsiendi andmed ja talle määratud ICD-10 koodid elektroonilistesse andmebaasidesse, mis on omakorda ühendatud ka geenivaramu andmetega. Nii on iga geenidoonori puhul kättesaadav kogu tema haiguste ajalugu ICD-10 koodidena, mis sobivad erinevate fenotüüpide defineerimiseks geneetiliste seoseuuringute tarvis.

Selles töös rakendame vaadeldud meetodeid, uurimaks geenivariantide seoseid teist tüüpi diabeediga (T2D, *type 2 diabetes*). Esimest ja teist tüüpi diabeedi tähisteks ICD-10 süsteemis on vastavalt E10 ja E11.

Diabeet on energiaainevahetuse häire, mille korral ei tooda kõhunääre piisavalt insuliini, insuliini toime on nõrgenenud või selle eritumine puudulik. Insuliin on eluks hädavajalik hormoon,

mida toodetakse kõhunäärmes ning mis aitab keharakkudel omastada veresuhkrut. Häiritud energiaainevahetus väljendub vere suurenenud glükoosisisaldusena. Esimest tüüpi diabeeti põhjustab kõhunäärmes insuliini tootvate rakkude beetarakkude hävitamine inimese enda immuunsüsteemi poolt ja see haigus algab tavaliselt, aga mitte alati, lapseeas või noorena. Teist tüüpi diabeeti põhjustab kõhunäärme suutmatus toota piisavalt insuliini või insuliini toime nõrgenemine ehk insuliinresistentsus ja see haigus algab pigem täiskasvanueas. Diabeet ei ole ravitav, kuid seda on võimalik kontrolli all hoida tervisliku eluviisi ja ravimite abil. (Leik, 2016)

Kõigist diabeedi juhtudest moodustab teist tüüpi diabeet 90-95%. Peamisteks teist tüüpi diabeeti haigestumise riskiteguriteks on vanus, ülekaalulisus või rasvumine, tasakaalustamata toitumine ja vähene liikumine. (American Diabetes Association, 2021)

On teada, et teist tüüpi diabeet on päriliku eelsoodumusega, kuid pärilikkuse hinnangud varieeruvad erinevate uuringute kohaselt 25-80% (Prasad ja Groop, 2015). Seega ei ole geneetika roll teist tüüpi diabeedi riskifaktorina siiani hästi teada ja selle uurimine pakub praegusel personaalmeditsiini arendamise ajastul väga suurt huvi (Chung *et al*., 2020).

Eestis on teist tüüpi diabeedi levimuseks hinnatud 7-9%, samal ajal on teada, et see haigus on aladiagnoositud (Ambos *et al*., 2016). Teist tüüpi diabeet ei ole kiire suremusega, mistõttu ei tohiks jälgimiseelsed ning jälgimisaegsed juhud üksteisest oluliselt erineda.

Analüüsi tegemiseks märgiti juhtudeks kõik sellised geenidoonorid, kellel oli teist tüüpi diabeet diagnoositud vähemalt ühel korral, välja arvatud need, kellel oli diagnoositud ka esimest tüüpi diabeet. Kontrollideks valiti geenidoonorid, kellel ei olnud ühtegi esimest ega teist tüüpi diabeedi diagnoosi.

Genotüübiandmed

Uuritavateks geenivariantideks on valitud sellised SNP-d, mille kohta on teada, et need on seotud teist tüüpi diabeediga. Täpsemalt valiti analüüsimiseks kaheksa vähima p-väärtusega tulemust viimasest suurest transetnilisest teist tüüpi diabeedi meta-analüüsist (Mahajan *et al*., 2020). Iga geenidoonori puhul on teada nendele SNP-dele vastavad genotüübid ehk alleeli-

doosid (arvud 0 ja 2 vahel, mis kirjeldavad vastava SNP eeldatavat efektialleeli sagedust).

Geenidoonorite genotüübiandmetele on TÜ EGV bioinformaatika tuumiklaboris eelnevalt tehtud kvaliteedikontroll programmiga Plink 1.9 (Purcell *et al.*, 2007). Selle käigus jäeti alles need indiviidid, kelle puhul geneetiline sugu vastas geenidoonori ankeedis olevale soole, genotüüp oli määratud vähemalt 98% genotüpiseerimise kiibi peal olevatest positsioonidest ning heterosügootsete genotüüpide osakaal vastas ligikaudu kogu andmestiku keskmisele ehk jäi vahemikku keskmine $\pm$ 3 standardhälvet.

Sama programmiga on tehtud ka sugulusanalüüs. Nimelt on geenivaramu andmetest enne analüüsi välja jäetud need geenidoonorid, kes on omavahel kuni teise astme sugulased (esimese astme sugulased on omavahel vanemad ja lapsed, teise astme sugulased aga vanavanemad ja lapselapsed ning õed ja vennad). Selleks leitakse iga kahe indiviidi kohta nende ühispõlvnemise hinnang, mis näitab, kui kaugel (geneetilises mõttes) on nende viimane ühine esivanem. Seda hinnangut arvestades luuakse nimekiri indiviididest, kes tuleb andmestikust välja jätta. Sugulaste väljajätmine on optimeeritud teist tüüpi diabeedi juhtude suhtes ehk eelistatult jäetakse andmestikku alles see geenidoonor, kellel on teist tüüpi diabeedi diagnoos.

Programmi Plink 2.0 (Chang *et al.*, 2015) abil on genotüübiandmetega tehtud ka peakomponentanalüüs. Leitud peakomponendid lisatakse populatsiooni struktureerituse arvesse võtmiseks regressioonimudelitesse kovariaatidena.

Kirjeldav analüüs

Ilma sugulasteta TÜ EGV andmestikus on info 96 917 geenidoonori kohta, kellest 65% on naised ja 35% mehed. Nende vanus geenivaramuga liitumise ajal on olnud vahemikus 18-103 aastat, kusjuures keskmine liitumisvanus on 45 aastat.

Teist tüüpi diabeedi juhtude suhtes optimeeritud sugulasteta andmestikus on 11 239 teist tüüpi diabeedi diagnoosiga geenidoonorit, mis on 11,6% kogu andmestikust (esialgses sugulastega andmestikus on juhtude osakaaluks 12108/183058 $\approx$ 6%). Keskmiseks haigestumise vanuseks on 58 (standardhälbega 13) aastat. Teist tüüpi diabeedi juhtudest 7736 on jälgimiseelsed ja 3503

jälgimisaegsed.

Geenidoonorite andmete analüüs on läbi viidud analoogiliselt nii simulatsiooniuuringu teise kui kolmanda stsenaariumiga.

Iga doonori puhul on teada kuupäev, mil ta geenidoonoriks hakkas. Analoogiliselt simulatsiooniuuringu teises stsenaariumis tehtuga jagatakse ka geenidoonorite andmed kaheks. Esimesse andmestikku jäävad jälgimiseelsed juhud (ehk need juhud, kelle vanus haigestumise ajal ei ole suurem kui liitumisvanus) ja juhuslikult valitud kontrollid, kusjuures iga juhu kohta valitakse andmestikku neli kontrolli. Kokku jäi esimesse andmestikku 38 680 geenidoonorit. Kõik ülejäänud 58 237 geenidoonorit ehk jälgimisaegsed juhud ja allesjäänud kontrollid moodustavad teise andmestiku. Geenidoonorite vaatlusajaks märgitakse juhtudel esimene kuupäev, mil neil oli diagnoositud teist tüüpi diabeet, kontrollidel kas surmakuupäev või katkestusaeg 31.12.2019 vastavalt sellele, kas doonor oli surnud enne katkestusaega või mitte. Katkestusajaks on valitud kuupäev 31.12.2019, sest sellest päevast pärinevad viimased diagnoosiandmed Haigekassast.

Esimesele andmestikule hinnatakse nii logistiline kui täiend-log-log mudel, kus uuritavaks tunnuseks on teist tüüpi diabeedi diagnoosi olemasolu ja seletavaks tunnuseks SNP. Kovariaatidena lisatakse mudelisse ka liitumisvanus, sugu ja viis esimest peakomponenti. Teisele andmestikule hinnatakse Coxi võrdeliste riskide mudel ja martingaalijääkide mudel, ajaskaalana kasutatakse aega alates liitumisest ja kovariaatidena lisatakse mudelisse SNP, liitumisvanus, sugu ja viis esimest peakomponenti.

Kolmandale stsenaariumile vastava analüüsi puhul uuritakse jälgimiseelseid ja jälgimisaegseid haigusjuhtusid koos. Igale geenidoonorile määratakse binaarne haigestumise tunnus, mille väärtuseks on 1 või 0 vastavalt sellele, kas tegemist on teist tüüpi diabeedi juhu või kontrolliga. Andmetele hinnatakse nii logistiline kui täiend-log-log mudel, kus seletavateks tunnusteks on SNP, sugu, sünniaasta ja viis esimest peakomponenti.

4.2 Tulemused

Tabelis 3 on esitatud info analüüsis kasutatud SNP-de kohta.

Tabel 3. Info analüüsis kasutatud SNP-de kohta

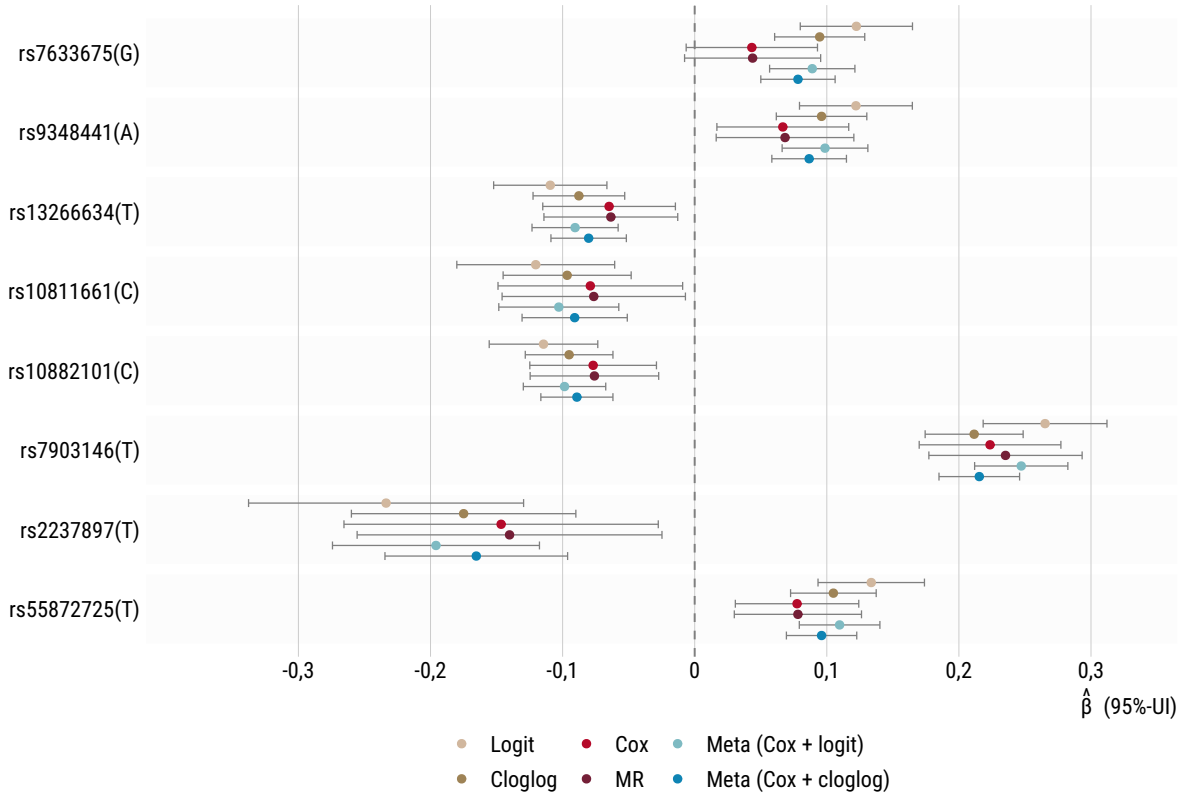
SNP nimi	Kromosoom	Efekti-alleel	Teine alleel	MAF	Üleeuroopalise meta-analüüsi $\hat{\beta}$	Transetnilise meta-analüüsi p
rs7633675	3	G	T	0,320	0,109	$5{,}8{\cdot}10^{-131}$
rs9348441	6	A	T	0,308	0,137	$6{,}2{\cdot}10^{-235}$
rs13266634	8	T	C	0,339	−0,108	$3{,}2{\cdot}10^{-115}$
rs10811661	9	C	T	0,138	−0,177	$1{,}1{\cdot}10^{-201}$
rs10882101	10	C	T	0,416	−0,110	$1{,}8{\cdot}10^{-125}$
rs7903146	10	T	C	0,219	0,298	≈ 0
rs2237897	11	T	C	0,044	−0,192	$5{,}5{\cdot}10^{-233}$
rs55872725	16	T	C	0,453	0,122	$4{,}7{\cdot}10^{-128}$

MAF - harvema alleeli sagedus geenivaramu kohordis

Näeme, et enamiku SNP-de puhul on harvema alleeli sagedus MAF (mis on samal ajal ka efektialleeli sagedus) kõrge. Tabelis on iga SNP puhul välja toodud ka üleeuroopalise meta-analüüsi hinnangud, mis on pärit samast teist tüüpi diabeedi transetnilisest meta-uuringust.

Joonisel 9 on kujutatud teisele stsenaariumile vastava analüüsi tulemused. Nende tulemuste puhul jagati andmestik esmalt kaheks – ühes on jälgimiseelsed juhud ja juhuslikult valitud terved kontrollid, teises jälgimisaegsed juhud ja ülejäänud terved kontrollid. Esimesele andmestikule on hinnatud logistiline ja täiend-log-log mudel. Teisele andmestikule on hinnatud Coxi võrdeliste riskide mudel, mida on lähendatud ka martingaalijääkide meetodiga. Coxi mudelite puhul oli iga SNP korral võrdeliste riskide eeldus Schoenfeldi jääkide testi põhjal täidetud. Meta-analüüsi hinnangud on saadud esimese ja teise andmestiku hinnangute kombineerimisel.

Esmalt märgime, et Coxi mudeli ja martingaalijääkide mudeli hinnangud on kõigi kaheksa SNP korral väga sarnased. See on kooskõlas simulatsioonide tulemustega: ühelähedaste efektisuuruste ja pigem suurte MAF-ide puhul sobib martingaalijääkide meetod Coxi mudeli lähendamiseks hästi. Sedasama seost näeme ka vastavate metahinnangute puhul (*Meta (MR + logit)* ja *Meta(MR + cloglog)*), kuid lihtsuse huvides ei ole neid joonisele märgitud. Kuna martingaalijääkide meetodi uurimine ei ole selle töö põhieesmärk, siis edaspidi keskendume Coxi mudeli

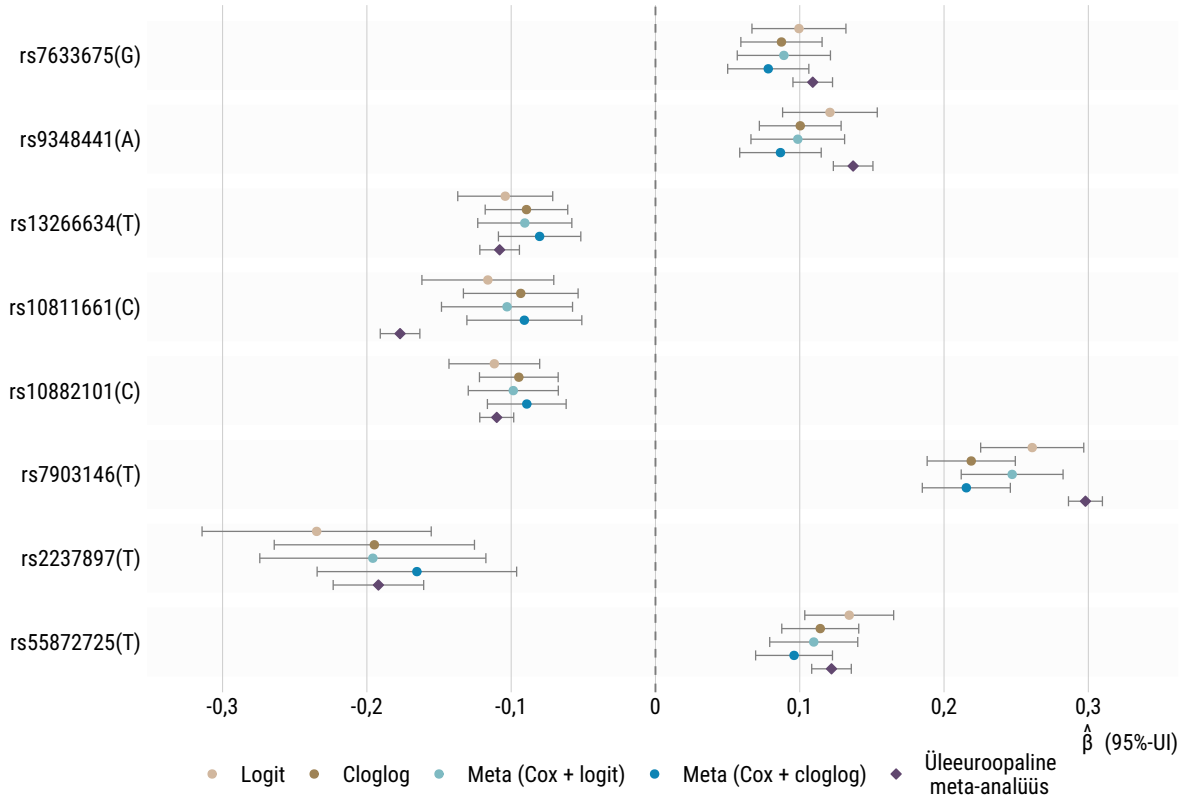


Joonis 9. T2D seosed kaheksa valitud SNP-ga (hinnang koos 95%-usaldusintervalliga) teise stsenaariumi puhul. Logit- ja cloglog-lingiga mudelid on hinnatud jälgimiseelsete juhtude andmetelt, Coxi mudel ja martingaalijääkide mudel on hinnatud jälgimisaegsete juhtude andmetelt. SNP nime järel on märgitud efektialleel. Logit - logistiline mudel, cloglog - täiend-log-log mudel, MR - martingaalijääkide mudel

ja binaarsete mudelite võrdlemisele.

Oodatult näeme, et logistilise regressiooni hinnangud on absoluutväärtuselt suuremad nii täiend-log-log mudeli hinnangutest kui Coxi mudeli hinnangutest. Nagu nägime ka teoreetiliselt ja simulatsiooniuuringu tulemusena, on täiend-log-log mudeli hinnangud Coxi mudeli hinnangutega sarnasemad kui logistilise mudeli hinnangud. Samasugune seos kehtib ka vastavate meta-hinnangute (*Meta (Cox + logit)* ja *Meta (Cox + cloglog)*) puhul.

Joonisel 10 on kujutatud lõplikud analüüsitulemused: metahinnangud teisest stsenaariumist ja kolmandale stsenaariumile vastavad binaarsete mudelite hinnangud.



Joonis 10. T2D seosed kaheksa valitud SNP-ga (hinnang koos 95%-usaldusintervalliga). Logit- ja cloglog-lingiga mudelid on hinnatud kõikidelt andmetelt, metahinnangute puhul on kombineeritud jälgimiseelsete juhtude analüüsi tulemused jälgimisaegsete juhtude analüüsi tulemustega. SNP nime järel on märgitud efektialleel. Logit - logistiline mudel, cloglog - täiend-log-log mudel, MR - martingaali-jääkide mudel

Täpsemalt on iga SNP puhul kujutatud nelja erineva meetodi hinnang ja vastav 95%-usaldusintervall. Võrdluseks on esitatud ka transetnilise meta-analüüsi Euroopa-põhised efekti-suuruste hinnangud koos 95%-usaldusintervalliga. Et selles metauuringus on analüüside läbi-viimisel kasutatud logistilist regressiooni, on vastavateks efektisuurusteks logaritmitud šansside suhted $(\log(\mathrm{OR}))$. Seega annavad need hinnangud aimu meie efektisuuruste ja -suundade õig-susest, kuid ei sobi selleks, et hinnata erinevate mudelite hinnangute headust: on oodatav, et enamiku SNP-de puhul on üleeuroopalise meta-analüüsi hinnangutele kõige lähemal just logis-tilise mudeli hinnangud.

Hinnangute suurused on kõigi SNP-de ja kõigi meetodite puhul sarnased ja efektisuuru-sed on ühe lähedal – minimaalne $\exp(\hat{\beta})$ on $0{,}79$ ja maksimaalne $\exp(\hat{\beta})$ on $1{,}30$. Kõige

suuremat erinevust hinnangute suuruse vahel näeme rs2237897 puhul: *Logit* hinnangu ja *Meta (Cox + cloglog)* hinnangu vahe on $-0{,}039$. Ka p-väärtuste poolest on hinnangud sarnased: kuigi logistilist regressiooni kasutavate meetodite hinnangute standardvead olid iga SNP puhul suuremad kui vastavad täiend-log-log regressiooni kasutavad meetodid, kompenseerisid seda logistilise mudeli absoluutväärtuselt suuremad hinnangud. Seega on p-väärtused pea iga SNP puhul kõige väiksemad just *Logit* meetodi puhul, kuid erinevused *Cloglog* meetodiga on väikesed (suurim *Logit* ja *Cloglog* meetodi p-väärtuste vahe on rs10882101 korral ja see on umbes $3{,}51 \cdot 10^{-6}$).

Seega ei näe me, et hinnangute meta-analüüsimine annaks eelise lihtsa binaarse mudeli kasutamise ees. Arvestada tuleb ka sellega, et ainult logistilise või täiend-log-log mudeli kasutamine on arvutuslikult palju kiirem (ainult kaheksa SNP seoseid uurides see muidugi välja tule) ja ka andmete ettevalmistamine on oluliselt lihtsam: vaja on teada vaid seda, kas inimesel oli haigus diagnoositud, ning ei ole vaja arvestada haigestumis- ja liitumiskuupäevade andmetega, leidmaks haigestumise vanust ning eraldamaks jälgimisaegseid ning jälgimiseelseid haigusjuhtusid erinevatesse andmestikesse. Meta-analüüsi meetod on seega oluliselt töömahukam, hõlmates endas juhtude eristamist, erinevate mudelite hindamist ja seejärel hinnangute kombineerimist. Kõike seda silmas pidades võib öelda, et jälgimiseelseid ja jälgimisaegseid juhtusid võiks analüüsida koos, kasutades selleks binaarset mudelit.

5 Arutelu

Selles töös nägime nii simulatsiooniuuringus kui geenivaramu andmete analüüsimisel, et Coxi võrdeliste riskide mudeli rakendamine ei pruugi alati anda olulist eelist lihtsa binaarse mudeli kasutamise ees. See teadmine teeb retrospektiivsete ja prospektiivsete andmete kombineerimise lihtsaks, sest mõlemat tüüpi andmeid on võimalik analüüsida koos. Näitasime, et kui hinnatud parameetrit soovitakse tõlgendada kui logaritmilist riskimäärade suhet (see tähendab, nii nagu Coxi mudeli parameetrit), siis tuleks traditsioonilise logistilise seosefunktsiooni asemel kasutada täiend-log-log seosefunktsiooni.

Suurte metauuringute puhul on erinevatel osalevatel kohortidel analüüside läbiviimiseks erinevad võimalused. Seetõttu võidakse ühtsuse mõttes iga kohordi puhul analüüs läbi viia mõne lihtsa meetodiga, näiteks logistilist regressiooni kasutades, olgugi et osal kohortidest võivad olla olemas nii andmed kui tarkvara Coxi võrdeliste riskide mudeli rakendamiseks. Teine võimalus, mida kasutatakse, on mudelite hindamine vastavalt iga kohordi võimalustele – kui on võimalik, kasutatakse Coxi mudelit, kuid kui on olemas vaid binaarsed haigestumise andmed, hinnatakse logistiline mudel – ja siis meta-analüüsitakse kokku nende erinevate mudelite hinnangud: riskimäärade suhted ja šansside suhted. Sellises olukorras tuleb hästi välja täiend-log-log mudeli kasutamise eelis: kui binaarse tunnuse analüüs teha täiend-log-log mudeli kaudu, on hinnangud interpreteeritavad samamoodi nagu Coxi mudeli hinnangud, ning nende hinnangute meta-analüüsimine on õigustatud.

Coxi mudeli paremus võib välja tulla näiteks siis, kui haigestumise vanusest sõltuvus on keeruline. Selles töös kasutasime haigestumisandmete genereerimiseks Weibulli jaotust, mille korral on haigestumise ja vanuse vaheline seos lihtne, ja see võib olla põhjuseks, miks töötas väga hästi ka binaarne täiend-log-log mudel, kus vanuse arvesse võtmiseks oli lihtsalt sünniaasta kovariaadina mudelisse lisatud. Samal ajal ei ole selge, kuidas on kõige õigem binaarse mudeli hindamisel vanust arvesse võtta. Ainult jälgimiseelsete juhtude analüüsimisel on loomulikuks valikuks lisada kovariaadina mudelisse geenivaramuga liitumise vanus – juhtude puhul teame, et selleks vanuseks olid nad juba haigestunud, kontrollide puhul teame, et selles vanuses nad

veel haiged ei olnud. Kui aga analüüsida jälgimiseelseid ja -aegseid haigusjuhtusid koos, siis pole liitumisvanus enam informatiivne. Selles töös on sel juhul mudelisse lisatud inimese sünniaeg: see on üksüheselt seotud inimese vanusega analüüsi tegemise hetkel, mis juhtude puhul on vanus, enne mida nad olid haigestunud, ja kontrollide puhul vanus, enne mida nad ei olnud haigestunud. Selle lähenemise puhul ei arvestata aga sellega, et kõik geenidoonorid ei ole analüüsi tegemise ajaks enam elus, ja samuti ei pruugi see juhtude puhul olla väga hästi seotud haigestumise vanusega, sest mõni juhtudest võis haigeks jääda vahetult enne analüüsi tegemist, mõni aga aastakümneid enne seda. Edaspidi võiks uurida, kas mõni muu meetod vanuse arvesse võtmiseks on sobivam, näiteks kasutades kontrollide puhul vanust analüüsi tegemise või surma ajal, jälgimiseelsete juhtude puhul liitumisvanust ja jälgimisaegsete juhtude puhul vanust haigestumise ajal.

Selles töös tehtud simulatsiooniuuringu puhul ei ole arvestatud sellega, et mõni geenivariant võib jälgimiseelsete juhtude puhul mõjutada lisaks haigestumisele ka nende geenivaramuga liitumise tõenäosust. Samas on tähtis teadvustada, et jälgimiseelsete juhtude hulka saavad sattuda ainult need inimesed, kes pärast mingi diagnoosi saamist olid piisavalt terved, et geenivaramuga liituda, kuid jälgimisaegsete juhtude seas näeme me kõikide erinevate raskusastmetega haigestunuid. Seetõttu on jälgimiseelsete juhtude analüüsil üle esindatud madala haigusjärgse suremusega juhud ja nii võib näiteks madala suremusega seotud geenivariant näida olevat seotud suurema haigestumisriskiga.

Seega on nii hinnangute kombineerimine kui ka jälgimiseelsete ja jälgimisaegsete juhtude koos analüüsimine õige ainult siis, kui ei ole alust arvata, et mingi SNP mõjutab lisaks haigestumisele ka näiteks haiguse kulgu või suremust. Vastasel juhul võib see meetod eelmainitud põhjuste tõttu anda nihkega hinnanguid. Kui uurida geenivariantide seosed mingi haigusega, millel on teadaolevalt näiteks kiire suremus, siis ei pruugi olla õige jälgimiseelseid ja -aegseid juhtusid kokku võtta. Ülegenoomse analüüsi puhul võiks neid juhtusid seega esmalt analüüsida eraldi ja seejärel huvipakkuvate geenivariantide puhul testida erinevate mudelite hinnangute võrdsust.

Kokkuvõte

Magistritöö eesmärk oli leida jälgimiseelsete ja jälgimisaegsete haigusjuhtude koos analüüsimiseks sobiv meetod, mis oleks rakendatav ülegenoomsetes seoseuuringutes. Peamiselt pakkus huvi see, kas jälgimiseelsete ja jälgimisaegsete juhtude analüüsimine eraldi vastavalt binaarse mudeli ja Coxi võrdeliste riskide mudeli abil ning seejärel hinnangute kombineerimine on parem, kui analüüsida kõiki haigusjuhtusid koos, kasutades ainult binaarset mudelit. Lisaks uuriti, kui hästi töötab Coxi mudeli lähendamiseks mõeldud martingaalijääkide meetod.

Esmalt anti ülevaade biopankade andmete ja nendega seotud elektrooniliste terviseandmete analüüsimise eripäradest. Samuti kirjeldati erinevusi jälgimiseelsete ja jälgimisaegsete haigusjuhtude vahel. Seejärel kirjutati elukestusandmete eripäradest ja nende analüüsimiseks rakendatavast Coxi võrdeliste riskide mudelist ning kahest binaarse tunnuse analüüsimiseks kasutatavast meetodist: logistilisest ja täiend-log-log mudelist. Lisaks kirjeldati seoseid nende mudelite efektisuuruste vahel.

Erinevate meetodite võrdlemiseks viidi läbi simulatsiooniuuring. Selle tarvis kirjeldati esmalt, kuidas genereerida Weibulli jaotusega võrdeliste riskide mudelile vastavaid elukestusandmeid. Samuti tuletati viis, leidmaks sobivad Weibulli jaotuse parameetrid kindlaksmääratud levimuse ja haigestumise mediaanvanusega haigestumisandmete simuleerimiseks. Neid teadmisi kasutati andmete simuleerimiseks ja analüüsimiseks kolme erineva stsenaariumi korral: 1. kõik haigusjuhud on jälgimisaegsed, 2. osa juhtudest on jälgimiseelsed, osa jälgimisaegsed, 3. jälgimiseelseid ja jälgimisaegseid juhtusid ei eristata.

Simulatsiooniuuringu tulemusel selgus, et jälgimiseelsete ja jälgimisaegsete juhtude eraldi analüüsimine, kasutades vastavalt binaarset mudelit ning Coxi mudelit, ja sellele järgnev saadud hinnangute kombineerimine ei ole parem kui kõigi haigusjuhtude koos analüüsimine binaarse mudeliga. Samuti selgus, et täiend-log-log mudel hindab väga hästi andmete genereerimise aluseks olevaid riskimäärade suhteid. Ka nägime, et martingaalijääkide meetodi kasutamine on õigustatud, kui harvema alleeli sagedus on kõrge või efektisuurus ei ole palju erinev ühest.

Meetodeid rakendati Tartu Ülikooli Eesti Geenivaramu andmetel, uurimaks teist tüüpi diabeedi seoseid SNP-dega, mis valiti välja viimase suurima teist tüüpi diabeedi transetnilise meta-uuringu tulemustest. Esmalt nägime, et martingaalijääkide meetod töötab Coxi regressiooni lähendamiseks vaadeldud SNP-de korral hästi – selle hinnangud on väga lähedased Coxi mudeli hinnangutega. Samuti nägime, et erinevate mudelite puhul on tulemused sarnased ja keerulisem meta-analüüsi meetod ei anna eelist lihtsa täiend-log-log mudeli kasutamise ees. Seega võiks ülegenoomsete seoseuuringute puhul jälgimiseelseid ja -aegseid haigusjuhtusid analüüsida koos, kasutades selleks täiend-log-log mudelit.

Viited

Ambos, A., Raie, E., Kiudma, T., Reppo, I., Rätsep, A., Tammiksaar, K., Toomsoo, T., ja Volke, V. (2016). 2. tüüpi diabeedi Eesti ravijuhend 2016. *Eesti Arst*, *95*(7), 465–473. doi: 10.15157/ea.v0i0.13016

American Diabetes Association. (2021). 2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2021. *Diabetes Care*, *44*(Supplement 1), S15–S33. doi: 10.2337/dc21-S002

Bangdiwala, S. I. (2010). At odds with ratios. *International Journal of Injury Control and Safety Promotion*, *17*(1), 73–76. doi: 10.1080/17457301003588229

Breslow, N. E. (1972). Contribution to Discussion on Professor Cox's Paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 216–217. doi: 10.1111/j.2517-6161.1972.tb00900.x

Callas, P. W., Pastides, H., ja Hosmer, D. W. (1998). Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *American Journal of Industrial Medicine*, *33*(1), 33–47. doi: 10.1002/(SICI)1097-0274(199801)33:1<33::AID-AJIM5>3.0.CO;2-X

Canchola, A., Stewart, S., Bernstein, L., West, D., Ross, R., Deapen, D., Pinder, R., Reynolds, P., Wright, W., Anton-Culver, H., Peel, D., Ziogas, A., ja Horn-Ross, P. (2003). Cox regression using different time-scales. Kättesaadav: https://www.lexjansen.com/wuss/2003/DataAnalysis/i-cox_time_scales.pdf

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., ja Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). doi: 10.1186/s13742-015-0047-8

Chung, W. K., Erion, K., Florez, J. C., Hattersley, A. T., Hivert, M. F., Lee, C. G., McCarthy, M. I., Nolan, J. J., Norris, J. M., Pearson, E. R., Philipson, L., McElvaine, A. T., Cefalu, W. T.,

Rich, S. S., ja Franks, P. W. (2020). Precision medicine in diabetes: A consensus report from the american diabetes association (ada) and the european association for the study of diabetes (easd). *Diabetes Care*, *43*(7), 1617–1635. doi: 10.2337/dci20-0022

Collett, D. (2015). *Modelling Survival Data in Medical Research*. New York, NY: Chapman and Hall/CRC. doi: 10.1201/b18041

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202. doi: 10.1111/j.2517-6161.1972.tb00899.x

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276. doi: 10.2307/2335362

Davies, H. T. O., Crombie, I. K., ja Tavakoli, M. (1998). When can odds ratios mislead? *BMJ*, *316*(7136), 989–991. doi: 10.1136/bmj.316.7136.989

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*(359), 557–565. doi: 10.1080/01621459.1977.10480613

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, *5*(10), 3. doi: 10.2307/1174772

Grambsch, P. M., ja Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515. doi: 10.2307/2337123

Green, M. S., ja Symons, M. J. (1983). A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of Chronic Diseases*, *36*(10), 715–723. doi: 10.1016/0021-9681(83)90165-0

Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (GWAS). *Methods in Molecular Biology*, *1019*, 149–169. doi: 10.1007/978-1-62703-447-0_6

Hong, E. P., ja Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics Informatics*, *10*(2), 117. doi: 10.5808/gi.2012.10.2.117

Ingram, D. D., Makuc, D. M., ja Feldman, J. J. (1997). Re: "Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale". *American Journal of Epidemiology*, *146*(6), 528–529. doi: 10.1093/oxfordjournals.aje.a009309

Joshi, P. K., Fischer, K., Schraut, K. E., Campbell, H., Esko, T., ja Wilson, J. F. (2016). Variants near CHRNA3/5 and APOE have age-and sex-related effects on human lifespan. *Nature Communications*, *7*(1), 1–7. doi: 10.1038/ncomms11174

Kalbfleisch, J. D., ja Prentice, R. L. (1973). Marginal likelihoods based on cox's regression and life model. *Biometrika*, *60*(2), 267. doi: 10.2307/2334538

Kalbfleisch, J. D., ja Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/9781118032985

Klein, J. P., ja Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer-Verlag New York.

Kleinbaum, D. G., ja Klein, M. (2012). *Survival Analysis*. New York, NY: Springer New York. doi: 10.1007/978-1-4419-6646-9

Korn, E. L., Graubard, B. I., ja Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *American Journal of Epidemiology*, *145*(1), 72–80. doi: 10.1093/oxfordjournals.aje.a009034

Leffondré, K., Abrahamowicz, M., ja Siemiatycki, J. (2003). Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: A simulation study. *Statistics in Medicine*, *22*(24), 3781–3794. doi: 10.1002/sim.1674

Leik, I. (2016). *Mis on diabeet?* Kättesaadav: https://www.haigekassa.ee/blogi/mis-diabeet

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., ja Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. doi: 10.1038/ng.3190

Mahajan, A., Spracklen, C. N., Zhang, W., Ng, M. C., Petty, L. E., Kitajima, H., Yu, G. Z., Rüeger, S., Speidel, L., Kim, Y. J., Horikoshi, M., Mercader, J. M., Taliun, D., Moon, S., Kwak, S. H., Robertson, N. R., Rayner, N. W., Loh, M., Kim, B. J., ... Morris, A. P. (2020). Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *medRxiv*. doi: 10.1101/2020.09.22.20198937

Matlab optimization toolbox. (2021). (The MathWorks, Natick, MA, USA) Kättesaadav: https://se.mathworks.com/help/optim/ug/fsolve.html

Maziarz, M., Liu, Y., Qin, J., ja Pfeiffer, R. M. (2019). Inference for case-control studies with incident and prevalent cases. *Biometrics*, *75*(3), 842-852. doi: 10.1111/biom.13023

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., ja Marchini, J. (2020). Computationally efficient whole genome regression for quantitative and binary traits. *bioRxiv*. doi: 10.1101/2020.06.19.162354

Oleckno, W. A. (2008). *Epidemiology: Concepts and Methods*. Long Grove, IL: Waveland Press. Kättesaadav: https://books.google.ee/books?id=KXsbAAAAQBAJ

Peto, R. (1972). Contribution to Discussion on Professor Cox's Paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 205–207. doi: 10.1111/j.2517-6161.1972.tb00900.x

Pilling, L. C., Kuo, C. L., Sicinski, K., Tamosauskaite, J., Kuchel, G. A., Harries, L. W., Herd, P., Wallace, R., Ferrucci, L., ja Melzer, D. (2017). Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging*, *9*(12), 2504–2520. doi: 10.18632/aging.101334

Prasad, R. B., ja Groop, L. (2015). Genetics of type 2 diabetes-pitfalls and possibilities. *Genes*, *6*(1), 87–123. doi: 10.3390/genes6010087

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., ja Sham, P. C. (2007). PLINK: a tool set

for whole-genome association and population-based linkage analyses. *American journal of human genetics*, *81*(3), 559–575. doi: 10.1086/519795

R Core Team. (2019). R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria) Kättesaadav: https://www.R-project.org/

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239. doi: 10.2307/2335876

Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., Pervjakova, N., Pers, T. H., Johnson, A. D., Eicher, J. D., Jackson, A. U., Ferreira, T., Lee, Y., Ma, C., Steinthorsdottir, V., Thorleifsson, G., Qi, L., Van Zuydam, N. R., Mahajan, A., ... Prokopenko, I. (2017). An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, *66*(11), 2888–2902. doi: 10.2337/db16-1253

Staley, J. R., Jones, E., Kaptoge, S., Butterworth, A.Š., Sweeting, M. J., Wood, A. M., ja Howson, J. M. M. (2017). A comparison of cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *European Journal of Human Genetics*, *25*(7), 854–862. doi: 10.1038/ejhg.2017.78

Sutradhar, R., ja Austin, P. C. (2018). Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology*, *28*(1), 54–57. doi: 10.1016/j.annepidem.2017.10.014

Symons, M. J., ja Moore, D. T. (2002). Hazard rate ratio and prospective epidemiological studies. *Journal of Clinical Epidemiology*, *55*(9), 893-899. doi: 10.1016/S0895-4356(02)00443-2

Therneau, T. M. (2020). A Package for Survival Analysis in R. Kättesaadav: https://CRAN.R-project.org/package=survival

Therneau, T. M., Crowson, C., ja Atkinson, E. (2021). Multi-state models and competing

risks. *Vignette included in R package survival, version 3.2-11*. Kättesaadav: https://CRAN.R-project.org/web/packages/survival/vignettes/compete.pdf

Therneau, T. M., ja Grambsch, P. M. (2000). *Modeling survival data: Extending the cox model*. New York, NY: Springer New York. doi: 10.1007/978-1-4757-3294-8

Therneau, T. M., Grambsch, P. M., ja Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, *77*(1), 147–160. doi: 10.1093/biomet/77.1.147

Thiébaut, A. C., ja Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: A simulation study. *Statistics in Medicine*, *23*(24), 3803–3820. doi: 10.1002/sim.2098

Timmers, P. R., Wilson, J. F., Joshi, P. K., ja Deelen, J. (2020). Multivariate genomic scan implicates novel loci and haem metabolism in human ageing. *Nature Communications*, *11*(1), 1–10. doi: 10.1038/s41467-020-17312-3

van der Net, J. B., Janssens, A. C. J., Eijkemans, M. J., Kastelein, J. J., Sijbrands, E. J., ja Steyerberg, E. W. (2008). Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. *European Journal of Human Genetics*, *16*(9), 1111–1116. doi: 10.1038/ejhg.2008.59

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. doi: 10.18637/jss.v036.i03

Zemunik, T., ja Borask, V. (2011). Genetics of Type 1 Diabetes. In *Type 1 diabetes - pathogenesis, genetics and immunotherapy* (pp. 529–548). InTech. doi: 10.5772/21880

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W. Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., ja Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), 1335–1341. doi: 10.1038/s41588-018-0184-y

Lisad

Lisa 1. Schoenfeldi jäägid

Nagu nimigi ütleb, on Coxi võrdeliste riskide mudeli tähtsaks eelduseks see, et riskimäärade suhe püsib ajas muutumatuna. Selle eelduse kontrollimiseks on võimalik kasutada Schoenfeldi jääke (Schoenfeld, 1982). Need erinevad klassikalistest jääkidest selle poolest, et igale vaatluse-le arvutatakse mitu jääki – üks jääk iga argumenttunnuse kohta. Olgu meil n indiviidi põhjal hinnatud p argumenttunnusega Coxi mudel

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i).$$

Siis i-ndale $(i = 1, \ldots, n)$ indiviidile vastab p-mõõtmeline Schoenfeldi jääkide vektor kujul

$$\hat{\boldsymbol{s}}_i = \delta_i \left\{ \boldsymbol{x}_i - \frac{\sum_{l \in R_i} \boldsymbol{x}_l \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l)}{\sum_{l \in R_i} \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l)} \right\},$$

kus δ_i on sündmuse toimumise indikaator, $\boldsymbol{x}_i$ on i-nda indiviidi argumenttunnuste vektor ja R_i on hetkele t_i vastav riskigrupp. Definitsioonist näeme, et Schoenfeldi jääk i-nda indiviidi j-nda $(j = 1, \ldots, p)$ argumenttunnuse x_{ji} jaoks on erinevus selle argumenttunnuse tegeliku väärtuse ja riskigrupi R_i kaalutud keskmise vahel, kus l-nda $(l \in R_i)$ indiviidi kaaluks on

$$\frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l)}{\sum_{k \in R_i} \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_k)}.$$

Samuti paneme tähele, et tsenseeritud vaatluste jäägid on alati nullid, seega sageli loetakse tsenseeritud vaatluste Schoenfeldi jäägid lihtsalt puuduvateks väärtusteks.

On näidatud, et praktikas on parem kasutada kaalutud Schoenfeldi jääke. Nende vektor on defi-neeritud tavaliste Schoenfeldi jääkide kaudu järgmiselt:

$$\boldsymbol{s}_i^* = d\text{var}(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{s}}_i,$$

kus d on toimunud sündmuste arv ja var($\hat{\boldsymbol{\beta}}$) on hinnatud parameetrite $\hat{\boldsymbol{\beta}}$ kovariatsioonimaatriks. On näidatud, et j-ndale argumenttunnusele vastava kaalutud Schoenfeldi jäägi korral kehtib

$$\mathrm{E}(s_{ji}^*) + \hat{\beta}_j \approx \beta_j(t_i),$$

kus $\hat{\beta}_j$ on j-nda argumenttunnuse parameetri hinnang ja $\beta_j(t_i)$ on j-nda argumenttunnuse ajas muutuva kordaja väärtus i-nda indiviidi sündmuse toimumise ajal (Grambsch ja Therneau, 1994).

Tänu sellele on võimalik j-nda argumenttunnuse võrdeliste riskide eeldust kontrollida graafikult, kuhu on kantud väärtused $s_{ji}^* + \hat{\beta}_j$ ja vaadeldud sündmuste toimumise ajad. Kui graafiku punktid paiknevad horisontaalselt, siis on j-nda argumenttunnuse kordaja ajas konstantne ja võrdeliste riskide eeldus selle tunnuse korral kehtib. Statistiline test, mis võimaldab kontrollida võrdeliste riskide eelduse täidetust, testibki sellele graafikule sobitatud sirge tõusu erinevust nullist, R-is saab selle testi teha *survival* paketi funktsiooniga *cox.zph*.

Lisa 2. Riski- ja üleelamisfunktsiooni hindamine

See peatükk põhineb allikal (Collett, 2015: 107-116).

Olgu meil vaatluse all n subjekti, kellest r-il toimus sündmus, ja ülejäänud $n - r$ on paremalt tsenseeritud. Kui meil on hinnatud Coxi võrdeliste riskide mudel

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$$

ehk on olemas hinnang $\hat{\boldsymbol{\beta}}$, siis saab hinnata ka vastavad riski- ja üleelamisfunktsioonid. Riskifunktsiooni hinnang i-ndale indiviidile avaldub kujul

$$\hat{h}_i(t) = \hat{h}_0(t) \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i), \tag{6}$$

kus $\hat{h}_0(t)$ on hinnang baasriskifunktsioonile. Vaatame esmalt, kuidas leitakse hinnang $\hat{h}_0(t)$.

Olgu sündmuste toimumiste ajad järjestatud: $t_{(1)} \leqslant \ldots \leqslant t_{(r)}$. Tähistagu iga $j = 1, \ldots, r$ korral d_j hetkel $t_{(j)}$ toimuvate sündmuste arvu, n_j hetkel $t_{(j)}$ riskigrupis $R_{(j)}$ olevate indiviidide arvu ning olgu $D_{(j)}$ nende indiviidide hulk, kellel toimus sündmus hetkel $t_{(j)}$.

Kalbfleisch ja Prentice (1973) tuletasid suurima tõepära meetodil baasriskifunktsiooni hinnanguks astmefunktsiooni:

$$\hat{h}_0(t) = \frac{1 - \hat{\alpha}_j}{t_{(j+1)} - t_{(j)}}, \quad t_{(j)} \leqslant t < t_{(j+1)}, \, j = 1, \ldots, r-1,$$

ja $\hat{h}_0(t) = 0$, kui $t < t_{(1)}$, ning kus iga $j = 1, \ldots, r$ korral on $\hat{\alpha}_j$ lahend võrrandile

$$\sum_{l \in D_{(j)}} \frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l)}{1 - \hat{\alpha}_j^{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l)}} = \sum_{l \in R_{(j)}} \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l). \tag{7}$$

Suurust $\hat{\alpha}_j$ saab tõlgendada kui hinnangut tõenäosusele, et indiviid elab üle ajavahemiku hetkest $t_{(j)}$ kuni hetkeni $t_{(j+1)}$. Selle baasriskifunktsiooni hinnangu tuletamisel on tehtud eeldus, et järjestikuste sündmuste toimumiste aegade vahel on riskimäär konstantne.

Kui kõik sündmuste toimumiste ajad on erinevad, siis koosneb iga $j = 1, \ldots, r$ korral $D_{(j)}$ ainult ühest vaatlusest ja võrrand (7) lahendub analüütiliselt:

$$\hat{\alpha}_j = \left\{ 1 - \frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{(j)})}{\sum_{l \in R_{(j)}} \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_l)} \right\}^{\exp(-\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{(j)})},$$

kus $\boldsymbol{x}_{(j)}$ on selle indiviidi seletavate tunnuse vektor, kellel toimus sündmus hetkel $t_{(j)}$. Muudel juhtudel ehk võrdsete mittetsenseeritud aegade esinemisel tuleb võrrandi (7) lahendamiseks kasutada iteratiivseid meetodeid.

Baas-üleelamisfunktsiooni hinnang avaldub $\hat{\alpha}_j$ kaudu järgmise astmefunktsioonina:

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\alpha}_j, \quad t_{(k)} \leqslant t < t_{(k+1)}, \, k = 1, \ldots, r-1,$$

ja $\hat{S}_0(t) = 1$, kui $t < t_{(1)}$.

Märgime ka, et juhul kui seletavad tunnused puuduvad ja meil on andmed ainult vaatlusaegade kohta, siis võrrand (7) lihtsustub kujule $d_j/(1-\hat{\alpha}_j) = n_j$, millest $\hat{\alpha}_j = 1 - d_j/n_j$ ja vastav baasriskifunktsiooni hinnang hetkel $t_{(j)}$ on siis d_j/n_j ning üleelamisfunktsiooni hinnang on

$$\prod_{j=1}^{k} \left(1 - \frac{d_j}{n_j}\right),$$

mis on tuntud Kaplan-Meieri hinnang üleelamisfunktsioonile.

Kumulatiivse riskifunktsiooni ja üleelamisfunktsiooni omavahelise seose (2) kaudu saame kumulatiivse baasriskifunktsiooni hinnanguks

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = -\sum_{j=1}^{k} \log \hat{\alpha}_j, \quad t_{(k)} \leqslant t < t_{(k+1)},\, k = 1, \ldots, r-1,$$

ja $\hat{H}_0(t) = 0$, kui $t < t_{(1)}$.

Nüüd, kus baasfunktsioonide hinnangud on olemas, saame leida vastavad hinnangud ka igale indiviidile. Integreerides võrduse (6) mõlemaid pooli, saame i-nda indiviidi kumulatiivse riskifunktsiooni hinnanguks

$$\hat{H}_i(t) = \hat{H}_0(t) \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i),$$

ja kasutades kumulatiivse riskifunktsiooni ning üleelamisfunktsiooni vahelist seost (1), saame, et hinnang i-nda indiviidi üleelamisfunktsioonile on

$$\hat{S}_i(t) = \left\{\hat{S}_0(t)\right\}^{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)},$$

kus $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, \ldots, r-1$.

Kui andmetes esineb võrdseid sündmuste toimumiste aegu, siis võrrandi (7) iteratiivse lahendamise asemel saab kasutada ka järgmist lähendamist: võrrandi (7) vasaku poole saab ümber

kirjutada kujul

$$\sum_{l\in D_{(j)}} \frac{\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)}{1-\hat{\alpha}_j^{\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)}} = \sum_{l\in D_{(j)}} \frac{\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)}{1-\exp(\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)\log\hat{\alpha}_j)} \approx$$
$$\approx \sum_{l\in D_{(j)}} \frac{\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)}{1-(1+\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)\log\hat{\alpha}_j)} =$$
$$= \sum_{l\in D_{(j)}} \frac{1}{-\log\hat{\alpha}_j} =$$
$$= -\frac{d_j}{\log\hat{\alpha}_j},$$

ja seega võrrandi (7) asemel lahendame

$$-\frac{d_j}{\log\tilde{\alpha}_j} = \sum_{l\in R_{(j)}} \exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l),$$

millest

$$\tilde{\alpha}_j = \exp\left\{\frac{-d_j}{\sum_{l\in R_{(j)}}\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)}\right\}.$$

Seega baas-üleelamisfunktsiooni ja kumulatiivse baasriskifunktsiooni hinnangud saab iga $t_{(k)} \leqslant t < t_{(k+1)}, k = 1, \ldots, r-1$ jaoks kirja panna kujul vastavalt

$$\tilde{S}_0(t) = \prod_{j=1}^{k}\tilde{\alpha}_j = \prod_{j=1}^{k}\exp\left\{\frac{-d_j}{\sum_{l\in R_{(j)}}\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)}\right\}$$

ja

$$\tilde{H}_0(t) = -\log\tilde{S}_0(t) = \sum_{j=1}^{k}\frac{d_j}{\sum_{l\in R_{(j)}}\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_l)}. \tag{8}$$

Viimast hinnangut nimetatakse ka Breslow' või Nelson-Aaleni hinnanguks kumulatiivsele baasriskifunktsioonile.

Lisa 3. Haigestumisandmete simuleerimine

```
library(dplyr)
# N - valimi suurus
# MAF - harvema alleeli sagedus
# HR - riskimäärade suhe
# a, b - haigestumise vanuse Weibulli jaotuse baasparameetrid

sim_andmed = data.frame(g = rbinom(n = N, size = 2, prob = MAF)) %>% # genotüüp 0:AA,1:AB,2:BB
  mutate(sünniaeg = runif(n = n(), min = 1920, max = 1980),
         elukestus = rweibull(n = n(), shape = 9, scale = 90),  # elukestus T ~ W(9, 90)
         b_g = b * ((exp(log(HR) * g)) ** (-1/a)), # genotüübist sõltuv skaalaparameeter b_g
             ↪ haigestumise vanusele
         haigestumise_vanus = rweibull(n = n(), shape = a, scale = b_g), # haigestumise vanus
             ↪ H_g ~ W(a, b_g)
         vaatlusvanus = pmin(haigestumise_vanus, elukestus),  # vaatlusvanus on minimaalne
             ↪ haigestumise vanusest ja elukestusest
         on_juht = haigestumise_vanus == vaatlusvanus)  # haigus on neil, kes haigestusid enne
             ↪  surma

# 2. ja 3. stsenaariumi jaoks genereerime liitumisaja ja võtame arvesse katkestusaja
sim_andmed2 = sim_andmed %>%
  mutate(liitumisvanus = runif(n(), 2000, 2020) - sünniaeg,
         vaatlusvanus = pmin(haigestumise_vanus, elukestus, 2050 - sünniaeg),  # vaatlusvanus
             ↪ on minimaalne haigestumise vanusest, elukestusest, ja vanusest katkestusajal
         on_juht = haigestumise_vanus == vaatlusvanus) %>% # haigus on neil, kes haigestusid
             ↪ enne surma ja katkestusaega
  filter(elukestus >= liitumisvanus) # jätame alles vaid need, kes elasid vähemalt liitumiseni
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Tuuli Jürgenson,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Retrospektiivsete ja prospektiivsete andmete kombineerimine ülegenoomsetes seoseuuringutes“, mille juhendajad on Anastassia Kolde, Krista Fischer ja Reedik Mägi, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Tuuli Jürgenson

25.05.2021