Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Tuuli Jürgenson

Koopiaarvu variatsioonide mõju ravimi kõrvaltoimete tekkimisele

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendajad: Maarja Lepamets, MSc

Kaido Lepik, MSc

Tartu 2019

Koopiaarvu variatsioonide mõju ravimi kõrvaltoimete tekkimisele

Bakalaureusetöö

Tuuli Jürgenson

Koopiaarvu variatsioonid (CNV-d) on muutused inimese DNA-s, mille käigus on DNA piirkonnad kadunud või mitmekordistunud. Farmakogeenid on geenid, mille produktid osalevad ravimite lagundamisel. Bakalaureusetöö põhieesmärk on hinnata farmakogeenidega ülekattes olevate CNV-de mõju ravimi kõrvaltoimete tekkimisele.

Selleks luuakse esmalt teoreetiline raamistik CNV-de käsitlemiseks. Töös näidatakse, kuidas CNV-sid simuleerida, ning simulatsioonide abil näidatakse Tartu Ülikooli Eesti Geenivaramus välja töötatud CNV kvaliteediskoori paremust võrreldes populaarse CNV-sid määrava algoritmiga. Simulatsioonidega määratakse kindlaks ka CNV ja ravimi kõrvaltoime minimaalsed suhtelised sagedused, mille korral saavutatakse piisavalt suur võimsus seose avastamiseks.

Simulatsioonide tulemusi arvestades viiakse läbi praktiline analüüs CNV-de ja ravimi kõrvaltoimete vaheliste seoste avastamiseks Tartu Ülikooli Eesti Geenivaramu ja UK Biopanga andmetel. Analüüside käigus leiti oluline seos HLA-A geeniga ülekattes olevate duplikatsioonide ning rohtudest ja ravimitest tingitud üldise nahalööbe vahel.

Võtmesõnad: koopiaarvu variatsioonid, ravimite kõrvaltoimed, andmeanalüüs, simulatsioon

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika; B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Associations between copy number variations and adverse drug reactions

Bachelor's thesis

Tuuli Jürgenson

Copy number variations (CNVs) are gains or losses of segments of genomic DNA. Pharmacogenes are genes involved in drug response or metabolism. The aim of this bachelor's thesis is to analyse the effect of CNVs overlapping pharmacogenes on adverse drug reactions (ADRs).

First, a theoretical framework relating to CNVs is created. The thesis shows how to simulate CNV variables which are then used to demonstrate that the CNV quality score developed at Estonian Genome Center at the University of Tartu (EGCUT) is more effective compared to a binary CNV variable found by a popular CNV detection algorithm. The simulations are also used to determine the minimum relative frequencies of CNV and ADR needed to achieve sufficiently high power for detecting associations between the two.

Based on the results of the simulations, a practical analysis is carried out to find associations between CNVs and ADRs using data from EGCUT and UK Biobank. The analysis revealed a significant association between duplications overlapping the HLA-A gene and generalized skin eruption due to drugs and medicaments.

Keywords: copy number variation, adverse drug reactions, data analysis, simulation

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics; B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

Kasutatud lühendid

ADR	ravimi kõrvaltoime (*adverse drug reaction*)
bp	aluspaar (*base pair*)
CGF	kumulante genereeriv funktsioon (*cumulant generating function*)
CNV	koopiaarvu variatsioon (*copy number variation*)
DNA	desoksüribonukleiinhape (*deoxyribonucleic acid*)
EGCUT	Estonian Genome Center at the University of Tartu
ICD	rahvusvaheline haiguste klassifikatsioon (*The International Classification of Diseases*)
NA	puuduv väärtus (*not available*)
OR	šansside suhe (*odds ratio*)
PCA	peakomponentanalüüs (*principal component analysis*)
PharmGKB	The Pharmacogenomics Knowledge Base
SAIGE	Scalable and Accurate Implementation of GEneralized mixed model
SNV	üksiknukleotiidne variant (*single nucleotide variant*)
SPA	sadulpunkti lähendamine (*saddlepoint approximation*)
TÜ EGV	Tartu Ülikooli Eesti Geenivaramu
UKB	UK Biopank (UK Biobank)
VIP	väga oluline farmakogeen (*very important pharmacogene*)

Sissejuhatus

Personaalmeditsiini üks uurimisvaldkondi on ravimivastuse hindamine iga konkreetse patsiendi jaoks tema geneetilisest eripärast lähtudes. See, kui kiiresti inimese organism ravimit lagundab, kas ja kui kiiresti saabub ravimi toime, kui kaua ravimi toime kestab ja kui efektiivne see on ning kas ravimi lagundamise tulemusena tekivad kõrvaltoimed, varieerub indiviiditi. Erinevatele inimestele sobivad erinevad ravimid ja ravimiannused. On näidatud, et 99,8% Eesti inimestest kannab vähemalt ühte sellist geenivarianti, mille tõttu oleks mõnda sagedasti kasutatavat ravimit tarvis võtta tavapärasest erinevas annuses (Reisberg *et al*., 2018).

Tulevikus võib geeniinfo olla osa riiklikust tervishoiusüsteemist. Tänu sellele saaks arst enne ravi alustamist teada, kas patsiendil on mõni selline geenivariant, mille tõttu ravim ei toimi või põhjustab tõsiseid kõrvaltoimeid. Selline individuaalse geneetilise omapära arvestamine muudab meditsiinisüsteemi toimimise efektiivsemaks – kiiremini leitakse mõjuv raviviis, vähendatakse mittetoimiva ravi kasutamist ning välditakse kõrvaltoimete teket.

Farmakogeenideks nimetatakse geene, mille produktid ravimite lagundamises osalevad. Geneetilised variandid nendes geenides on üks põhjustest, miks ühed ja samad ravimid mõjuvad eri patsientidele erinevalt. Palju on uuritud üksiku nukleotiidi variatsioone (SNV, *single nucleotide variant*), kuid muud tüüpi variandid, näiteks koopiaarvu variatsioonid (CNV, *copy number variant*), on jäänud suuresti tähelepanuta.

CNV-d on muutused inimese genoomis, mille käigus on genoomipiirkonnad kadunud või mitmekordistunud. CNV-d hõlmavad rohkem nukleotiide kui üksiknukleotiidsed variatsioonid, võivad olla ülekattes tervete geenidega ja omavad seega tõenäoliselt suurt mõju inimestevahelisele varieeruvusele ning haiguste tekkele.

CNV-de uurimine on keerukas, kuna nad on sageli harvad ja varieeruvad nii pikkuse kui koopiate arvu poolest. CNV-sid määratakse erinevate algoritmide, näiteks PennCNV (Wang *et al*., 2007) abil. Algoritmid pole aga täiuslikud ning suureks probleemiks on valepositiivsete CNV-de rohkus. Selle lahendamiseks on Tartu Ülikooli Eesti Geenivaramus (TÜ EGV) välja tööta-

tud CNV kvaliteediskoor, mis peaks võimaldama valepositiivseid CNV-sid avastada ja seeläbi CNV-de analüüsimise täpsemaks muutma (Lepamets *et al*., 2019).

Bakalaureusetöö põhieesmärk on analüüsida, kas CNV-d mõjutavad ravimi kõrvaltoimete tekkimist. Praktiline analüüs tuvastamaks seoseid CNV-de ja ravimi kõrvaltoimete vahel viiakse läbi TÜ EGV ning UK Biopanga (UKB) andmetel.

Selleks formuleeritakse esmalt matemaatiliselt CNV-de ja kvaliteediskooriga seonduv ning luuakse teoreetiline raamistik, milles saab võrrelda CNV kvaliteediskoori PennCNV algoritmi leitud ja tegelike CNV-dega. Töös kirjeldatakse, kuidas CNV tunnuseid simuleerida, ning simulatsioonide abil näidatakse, et CNV kvaliteediskoor on parem kui PennCNV leitud CNV tunnus. Samuti määratakse tehtud simulatsioonide abil kindlaks, kui suured peavad olema CNV ja ravimi kõrvaltoime sagedused, et nendevaheliste seoste avastamiseks oleks piisavalt võimsust.

Enamasti kasutatakse binaarse tunnuse uurimiseks logistilist regressiooni. Bakalaureusetöös kasutatakse lisaks sellele ka Firth'i meetodit ning näidatakse simulatsioonide abil, et Firth'i meetod toimib väikeste sagedustega kõrvaltoimete ja CNV-de puhul paremini kui tavaline logistiline regressioon. Samuti kasutatakse töös seoste avastamiseks logistilist segamudelit, et arvestada indiviididevahelise sugulusega.

Analüüsi tulemused on sisendiks juhtivteadur Lili Milani juhitavale farmakogeneetika uurimisgrupile TÜ EGV-s. Andmete simuleerimiseks ning analüüsimiseks kasutati statistikatarkvara R ning kõik arvutused viidi läbi Tartu Ülikooli teadusarvutuste keskuse arvutusklastris (Teadusarvutuste keskus, www.hpc.ut.ee).

1 Bioloogiline taust

1.1 Põhimõisted geneetikast

Kõik rakud sisaldavad geneetilist materjali, millest valdav osa asub raku tuumas ja on organiseerunud kromosoomidesse. Kromosoomiks nimetatakse valkudega kokkupakitud DNA-molekuli (Kaart ja Möls, 2010). DNA on polümeer, mis koosneb omavahel ühendatud nukleotiididest, mida on nelja tüüpi: adeniin (A), tsütosiin (C), guaniin (G) ja tümiin (T). Aluspaariks (bp, *base pair*) nimetatakse kahte omavahel vesiniksidemetega seotud nukleotiidi, mis esinevad vastastikustes komplementaarsetes DNA-ahelates. Komplementaarsus tähendab, et DNA-molekulis ühe ahela adeniin seondub teise ahela tümiiniga ja ühe ahela tsütosiin teise ahela guaniiniga.

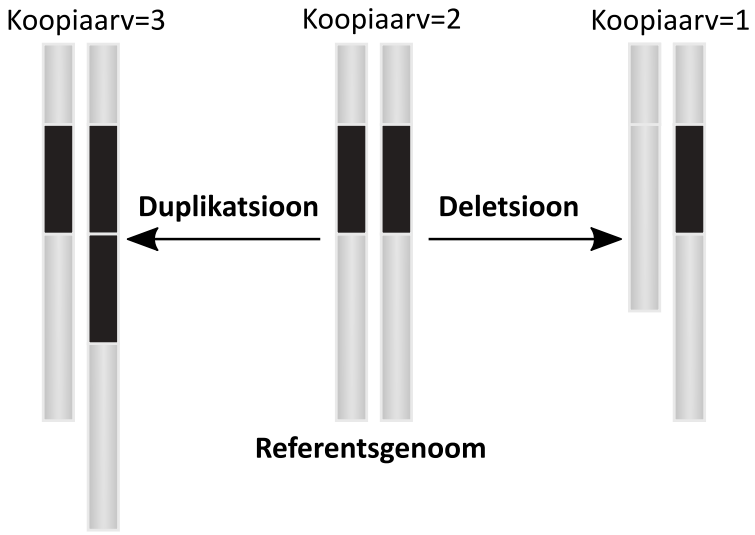
Inimeste genoom on diploidne, mis tähendab, et iga kromosoom esineb kahes koopias. Genoomiks nimetatakse liigiomast ühekordses kromosoomikomplektis sisalduvat kogu organismi pärilikku materjali. Referentsgenoom on mitme indiviidi genoomi põhjal kokku pandud teoreetiline genoom, millega võrreldakse reaalseid genoome.

Geeniks nimetatakse lõiku DNA-st, mis mõjutab mingi tunnuse kujunemist ning mis asub kromosoomil kindlas asukohas. Geeni järjestuse alusel luuakse geeniprodukt, näiteks valk. Geeniprodukti kogus sõltub muuhulgas ka sellest, mitmes koopias vastav geen organismi genoomis paikneb (DeBoever *et al.*, 2017). Genotüüp on organismi geneetiline struktuur. Fenotüübiks nimetatakse organismi avaldunud tunnuseid, mis on määratud tema genotüübi ja keskkonnategurite koostoimes (Heinaru, 2012: 991), näiteks inimese pikkus või tema ravimitaluvus.

Üksiknukleotiidseks variandiks (SNV) nimetatakse kindlas DNA punktis oleva üksiku aluspaari erinevust kahe indiviidi võrdlusel. SNV-d on kõige sagedasemateks variantideks inimese genoomis. Enamik SNV-sid ei mõjuta kuidagi inimese tervist ega arengut, kuid on leitud SNV-sid, mis on seotud näiteks indiviidide ravimivastusega, vastuvõtlikkusega keskkonnafaktoritele ja haigustesse jäämise riskiga (Buniello *et al.*, 2019). SNV-d pole aga ainsad olulise mõjuga variatsioonid.

1.2 Koopiaarvu variatsioonid

Koopiaarvu variatsiooniks (CNV) nimetatakse genoomisegmenti, mis erineb koopiaarvu poolest referentsgenoomist (Zarrei *et al*., 2015). Kui referentsgenoomis on igast genoomipiirkonnast kaks koopiat, siis CNV-de puhul on genoomis toimunud muutused, mille toimel on koopiaarv kas suurenenud või vähenenud ehk on toimunud vastavalt kas genoomipiirkonna duplikatsioon või deletsioon. Joonisel 1 on illustreeritud kahte tüüpi CNV-sid. Kui CNV on ülekattes geeniga, on muutunud ka vastava geeni koopiate arv organismi genoomis.



Joonis 1. Näide genoomipiirkonna duplikatsioonist ja deletsioonist diploidses genoomis.

CNV-del on mitmeid tekkemehhanisme, näiteks DNA-molekuli homoloogsete piirkondade vaheline ristsiire, mittehomoloogne DNA otste liitmine või defektid DNA paljundamises ja parandamises (Hastings *et al*., 2009).

Seoses arengutega DNA analüüsimiseks kasutatavates tehnoloogiates on CNV-de uurimine ja kaardistamine hoogustunud ning tänu sellele on kujunenud selgem arusaam CNV-de ulatusest ja mõjust fenotüübile (Korbel *et al*., 2008). CNV-de osakaal inimese genoomis on erinevate andmete põhjal 4,8-9,7% (Zarrei *et al*., 2015).

Enamik CNV-dest ei oma mingisugust fenotüübilist mõju ning erinevate geenide koopiaarvud võivad erineda ka tervete indiviide vahel (Korbel *et al*., 2008). Osadel CNV-del on aga oluline

roll haiguste tekkimise soodustamisel (Zarrei *et al*., 2015). Näiteks on leitud seoseid CNV-de ja autismi (Sebat *et al*., 2007), skisofreenia (Stone *et al*., 2008), Alzheimeri tõve (Swaminathan *et al*., 2012) ja rinnavähi (Zhang *et al*., 2009) vahel. CNV on ka näiteks Downi sündroomi põhjustav kogu 21. kromosoomi duplitseerumine.

Kui varem on hinnatud inimestevaheliseks geneetiliseks varieeruvuseks 0,1%, millest valdava osa moodustavad SNV-d, siis viimastel aastatel on leitud, et väga oluliseks teguriks indiviididevahelise geneetilise erinevuse taga on just CNV-d. Võttes arvesse CNV-sid, on hinnang inimestevahelisele geneetilisele sarnasusele vaid 99,5%. See tähendab, et geneetiline varieeruvus indiviidide vahel on viis korda suurem, kui varem arvatud (Levy *et al*., 2007).

Koopiaarvu variatsioonide tuvastamine

Koopiaarvu variatsioonide tuvastamiseks genotüpiseerimisandmetelt on loodud erinevaid algoritme. Üheks populaarseks meetodiks on varjatud Markovi mudelil põhinev PennCNV (Wang *et al*., 2007). CNV-de tuvastamise algoritmid annavad aga ka nii valepositiivseid (leitakse CNV, kui seda tegelikult ei eksisteeri) kui valenegatiivseid (tegelik CNV jääb tuvastamata) tulemusi. PennCNV algoritmi leitud valepositiivsete CNV-de avastamiseks kasutatakse CNV kvaliteediskoori. Selle skoori väljatöötamisel on kasutatud eeldust, et tõene CNV mõjutab ülekattes olevatelt geenidelt loodud geeniprodukti kogust (ja metülatsioonikiibilt mõõdetud markeri koguintensiivsust), samal ajal kui valepositiivne CNV seda ei mõjuta. Lisaks arvestab kvaliteediskoor seda, kas CNV on leitud ka sama indiviidi täisgenoomi sekveneerimisandmetelt. Saadud kvaliteediskoor on arv lõigust [0,1], mis määratakse igale PennCNV poolt leitud CNV-le, kusjuures valepositiivsed leiud peaksid saama nullilähedase skoori ja õigepositiivsete CNV-de skoor peaks jääma ühe lähedale. Valenegatiivsete leidude puhul kvaliteediskoor ei aita.

1.3 Farmakogeneetika

Erinevused indiviidide ravimivastuses on sagedased ja seda mõjutavad mitmed eri tegurid – näiteks inimese geneetiline iseärasus, sugu, vanus, keskkond, elustiil ning teised tarbitavad ravi-

mid. Ravimite mõju erinevatele inimestele võib varieeruda ravimi mittemõjumisest kuni eluohtlike kõrvalmõjude tekkeni. Enamik sageli kasutatavatest ravimitest on efektiivsed vaid 25-60% patsientide jaoks (Wilkinson, 2005). Peaaegu kõik Eesti inimesed kannavad vähemalt ühte sellist geenivarianti, mille tõttu vajaksid nad mõne sagedasti kasutatava ravimi puhul tavapärasest erinevat annust (Reisberg *et al*., 2018).

Farmakogeneetika on teadusharu, mis analüüsib inimese geneetika mõju tema individuaalsele ravimivastusele. Farmakogeneetikal on oluline osa personaalmeditsiinis – selle lõppeesmärk on anda arstidele informatsiooni inimese farmakoloogilise fenotüübi kohta, mis võimaldaks optimeerida ravimite kasutamist vastavalt konkreetse inimese genotüübile nii, et ta saaks ravist maksimaalset kasu ning kõrvalmõjude tekke risk oleks minimaalne.

Geene, mis on seotud erinevate farmakoloogiliste protsessidega, mille hulka kuuluvad näiteks ravimite imendumine, jaotumine, metabolism ja eritumine, nimetatakse farmakogeenideks. The Pharmacogenomics Knowledge Base (PharmGKB, www.pharmgkb.org) on kliiniliste farmakogeneetiliste markerite andmebaas, mis koondab ja jagab informatsiooni selle kohta, kuidas geneetilised variatsioonid mõjutavad ravimite toimet. PharmGKB lehel on olemas nimekiri väga olulistest farmakogeenidest (VIP, *very important pharmacogene*). Sellesse nimekirja kuuluvad geenid, mis kas osalevad paljude erinevate ravimite metabolismis või on seotud variatsioonidega, mis võivad soodustada tõsiste kõrvalmõjude teket.

2 Teoreetiline raamistik CNV-de hindamiseks

CNV olemasolu konkreetses kromosoomi piirkonnas on binaarne tunnus. Tegelik CNV olemasolu ei ole aga teada, see hinnatakse geeniandmete pealt näiteks PennCNV algoritmi abil, mis aga ei ole kunagi täiesti täpne. Algoritmi vigade mõju vähendamiseks on välja töötatud CNV kvaliteediskoor (Lepamets *et al*., 2019).

Bakalaureusetöös pakub huvi see, kas kvaliteediskoor toimib paremini kui PennCNV algoritmi poolt tuvastatud binaarne CNV tunnus. Eesmärk on välja selgitada, kas CNV osakaalude võrdlemisel uuritava fenotüübi juhtude ja kontrollide seas annab CNV kvaliteediskoor suurema võimsuse.

Järgnevalt antakse matemaatiline formuleering nii binaarse PennCNV tunnuse kui kvaliteediskoori kohta ja kirjeldatakse, kuidas neid suuruseid simuleerida.

2.1 PennCNV tunnus

Olgu $X \sim \mathrm{B}(1, p)$ CNV esinemist kirjeldav juhuslik suurus, kus p on CNV suhteline sagedus uuritavas populatsioonis. See tähendab, et X on juhuslik suurus, mille võimalikud väärtused on 1 ja 0, mis tähistavad vastavalt CNV olemasolu või selle puudumist.

Olgu $Y = f(X)$ juhuslik suurus, mis tähistab PennCNV poolt määratud CNV-d ning olgu PennCNV meetodi valepositiivsuse määr

$$p_{01} = P(Y = 1|X = 0)$$

ning valenegatiivsuse määr

$$p_{10} = P(Y = 0|X = 1).$$

Olgu õigepositiivsete ja õigenegatiivsete tulemuste määrad tähistatud vastavalt $p_{11} = 1 - p_{10}$ ning $p_{00} = 1 - p_{01}$. Siis saab kirjutada

$$Y = \begin{cases} 1, & \text{tõenäosusega } p_{11}, \text{ kui } X = 1 \\ 1, & \text{tõenäosusega } p_{01}, \text{ kui } X = 0 \\ 0, & \text{tõenäosusega } p_{10}, \text{ kui } X = 1 \\ 0, & \text{tõenäosusega } p_{00}, \text{ kui } X = 0 \end{cases} \sim \mathrm{B}(1, pp_{11} + (1-p)p_{01}). \qquad (*)$$

Valenegatiivsuse ja valepositiivsuse määrad pole teada, kuid neid on võimalik hinnata, kasutades geenidoonorite kohta kogutud sõltumatuid bioloogilisi andmeid. TÜ EGV andmete põhjal on hinnatud, et $p_{01} \approx 0{,}008$ ning $p_{10} \approx 0{,}4$ nii deletsioonide kui duplikatsioonide korral. Kuna CNV-d on harvad, siis on loogiline, et ka valepositiivsuse määr on väike. Samal ajal tähendab CNV harv esinemine seda, et ka väike valepositiivsete leidude määr toob kaasa suhteliselt suure valede CNV leidude hulga.

Valepositiivsete CNV leidude olemasolu muudab CNV-de sageduse võrdlemise juhtude ja kontrollide seas raskemaks. Näiteks, olgu CNV-de reaalne suhteline sagedus kontrollide seas $p_0 = 0{,}001$ ning juhtude seas $p_1 = 0{,}002$. Siis on tegelik šansside suhe

$$\mathrm{OR}_X = \frac{p_1(1-p_0)}{p_0(1-p_1)} = \frac{0{,}002 \cdot 0{,}999}{0{,}001 \cdot 0{,}998} \approx 2.$$

Kasutades aga arvutamiseks PennCNV abil leitud CNV-de osakaalusid, on šansside suhe palju väiksem:

$$\mathrm{OR}_Y = \frac{(p_1p_{11} + (1-p_1)p_{01})(1 - (p_0p_{11} + (1-p_0)p_{01}))}{(p_0p_{11} + (1-p_0)p_{01})(1 - (p_1p_{11} + (1-p_1)p_{01}))} \approx \frac{0{,}0092 \cdot 0{,}9914}{0{,}0086 \cdot 0{,}9908} \approx 1{,}07$$

ja seose avastamine diagnoosi ning CNV kandmise vahel on keeruline ja nõuab suuremat valimimahtu.

Simuleerimine

Nii tegeliku CNV kui PennCNV määratud CNV olemasolu on Bernoulli jaotusega ning seega on CNV-de andmeid lihtne simuleerida. PennCNV tunnuse puhul peab arvestama valepositiivsete

ja valenegatiivsete leidude määraga – kui reaalne CNV-de suhteline sagedus on p, siis PennCNV poolt leitud CNV-de suhteline sagedus on $pp_{11} + (1-p)p_{01}$ (vt valem $(*)$). Seega tehniliselt tähendab teadmine tegelikust CNV olemasolust, et valepositiivseid ega valenegatiivseid CNV leide ei ole ehk $p_{11} = 1$ ja $p_{01} = 0$.

Võimsuse juhtude ning kontrollide CNV-de osakaalude erinevuse leidmisel binaarse tunnuse korral saab leida analüütiliselt (Ma *et al.*, 2013). Olgu n_0 ja n_1 vastavalt kontrollide ja juhtude arv valimis ning p_0 ja p_1 CNV-de osakaal vastavalt kontrollide ja juhtude seas (kas tegelik osakaal või PennCNV tunnuse osakaal). Olgu $X_{0,i} \sim \mathrm{B}(1, p_0),\ i = 1, \ldots, n_0$ ja $X_{1,j} \sim \mathrm{B}(1, p_1),\ j = 1, \ldots, n_1$ sõltumatud juhuslikud suurused, mis tähistavad CNV esinemist vastavalt kontrollide ja juhtude seas.

Siis CNV-de arvud on vastavalt kontrollidel

$$T_0 = \sum_{i=1}^{n_0} X_{0,i} \sim \mathrm{B}(n_0, p_0)$$

ja juhtudel

$$T_1 = \sum_{j=1}^{n_1} X_{1,j} \sim \mathrm{B}(n_1, p_1).$$

Olgu α maksimaalne lubatav I liiki vea tõenäosus. Siis võimsuse saab arvutada valemiga

$$\text{võimsus} = P(P_t < \alpha | p_0 \neq p_1) = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} P(T_0 = i, T_1 = j) \cdot I(P_{t,i,j} < \alpha),$$

kus P_t tähistab statistilise testi t (näiteks logistilise regressiooni puhul Waldi testi) p-väärtuste jaotusega juhuslikku suurust konkreetse simulatsiooni konfiguratsiooni (n_0,n_1,p_0,p_1) korral, $P_{t,i,j}$ on testi t p-väärtus $T_0 = i$ ja $T_1 = j$ korral ning I on indikaatorfunktsioon. Juhul kui $p_0 = p_1$, saab eelneva valemi abil arvutada I liiki vea tõenäosuse.

Eeldades, et CNV kandjad juhtude ja kontrollide seas on sõltumatud, saab iga paari (i, j) korral

arvutada

$$P(T_0 = i, T_1 = j) = P(T_0 = i)P(T_1 = j) = \mathrm{C}_{n_0}^{i} p_0^i (1-p_0)^{n_0-i} \cdot \mathrm{C}_{n_1}^{j} p_1^j (1-p_1)^{n_1-j}.$$

Kood PennCNV ja tegeliku CNV tunnuse simuleerimiseks on toodud lisas 1.

2.2 Kvaliteediskoor

Kvaliteediskoor on arv lõigust [0,1], mis määratakse igale PennCNV poolt määratud CNV-le. Ideaalis peaksid kõik valepositiivsed CNV-d saama skoori, mis on 0 lähedal, ja õigepositiivsed CNV-d skoori, mis on 1 lähedal.

Olgu X ja Y defineeritud nii nagu eelnevalt. Olgu $Z = g(X,Y)$ kvaliteediskoori kirjeldav juhuslik suurus ja $\mathcal{S}_0$ ning $\mathcal{S}_1$ tundmatud jaotused, millest on genereeritud vastavalt valepositiivsete ja õigepositiivsete CNV leidude kvaliteediskoorid.

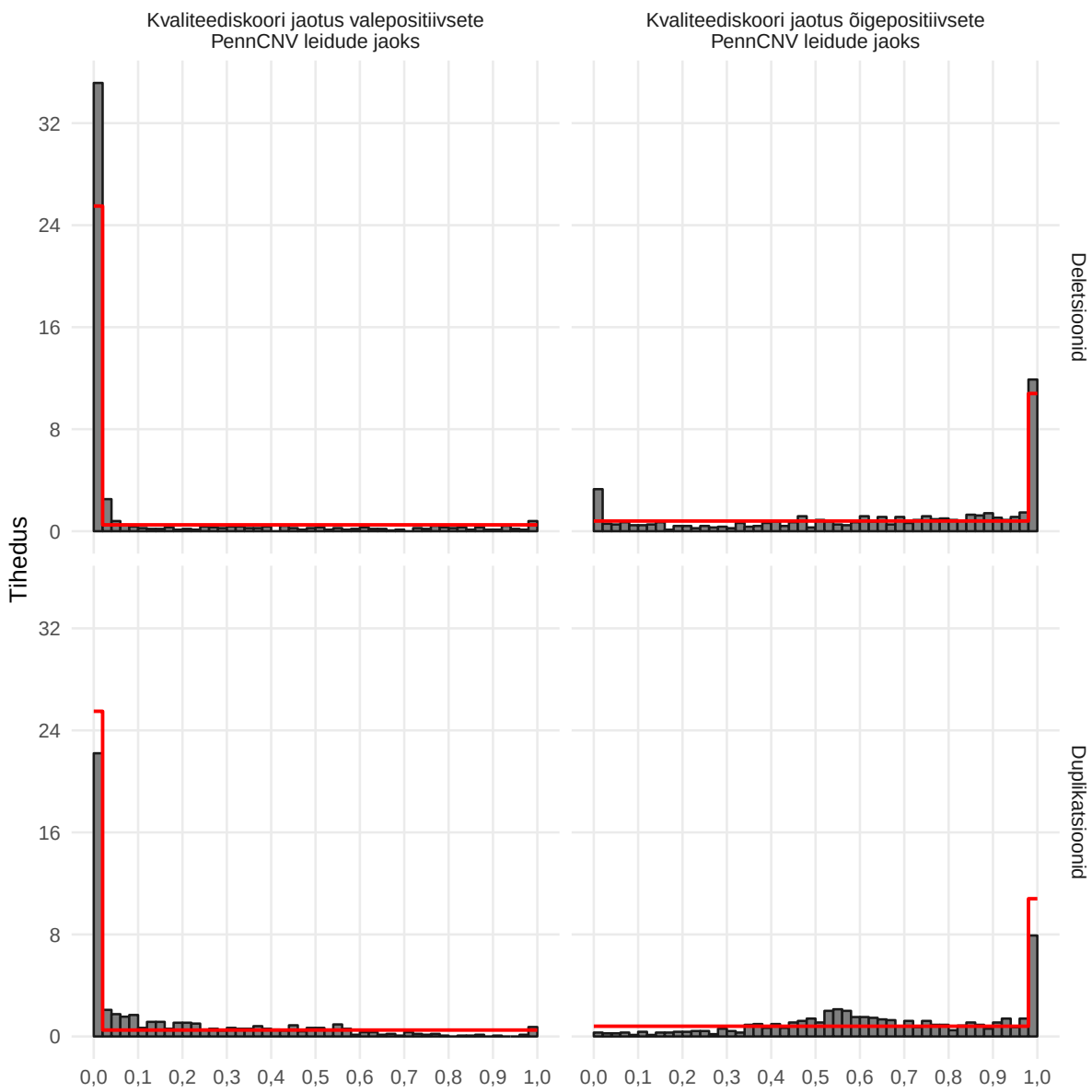
Siis

$$Z = g(X,Y) = \begin{cases} S_1 \leftarrow \mathcal{S}_1, & X = 1 \wedge Y = 1 \\ S_0 \leftarrow \mathcal{S}_0, & X = 0 \wedge Y = 1 \\ 0, & \text{muidu} \end{cases} = \begin{cases} S_1 \leftarrow \mathcal{S}_1, & \text{tõenäosusega } pp_{11} \\ S_0 \leftarrow \mathcal{S}_0, & \text{tõenäosusega } (1-p)p_{01} \\ 0, & \text{muidu,} \end{cases}$$

kus tähistus $S_i \leftarrow \mathcal{S}_i \ \ (i = 0, 1)$ märgib juhusliku suuruse S_i valimist jaotusest $\mathcal{S}_i$.

Teisisõnu, Z saab juhusliku väärtuse jaotusest $\mathcal{S}_0$, kui $X = 0 \wedge Y = 1$, juhusliku väärtuse jaotusest $\mathcal{S}_1$, kui $X = 1 \wedge Y = 1$ ning muudel juhtudel ehk vale- ja õigenegatiivsete CNV leidude puhul saab Z väärtuseks 0.

Jaotuseid $\mathcal{S}_0$ ja $\mathcal{S}_1$ on võimalik hinnata TÜ EGV andmete põhjal. Joonisel 2 on näha kvaliteediskoori empiirilised jaotused ja lähendina leitud treppfunktsioon. Kasutades jaotuste $\mathcal{S}_0$ ja $\mathcal{S}_1$ lähendamiseks treppfunktsiooni, on nende hinnangulisteks tihedusfunktsioonideks vastavalt



Joonis 2. Kvaliteediskoori empiiriline jaotus valepositiivsete (vasakul) ja õigepositiivsete (paremal) PennCNV leidude jaoks TÜ EGV andmete põhjal (histogrammi tulba laius on 0,02). Punase joonega on märgitud lähendina leitud kaheastmeline treppfunktsioon jaotustele $\mathcal{S}_0$ (vasakul) ja $\mathcal{S}_1$ (paremal).

$$f_{S_0}(x) = \begin{cases} 25{,}5, & 0 < x \leqslant 0{,}02 \\ 0{,}5, & 0{,}02 < x \leqslant 1 \\ 0, & \text{muidu} \end{cases}$$

ja

$$f_{S_1}(x) = \begin{cases} 0{,}8, & 0 < x \leqslant 0{,}98 \\ 10{,}8, & 0{,}98 < x \leqslant 1 \\ 0, & \text{muidu.} \end{cases}$$

Jaotusega $\mathcal{S}_0$ juhuslike suuruste genereerimiseks tuleb seega genereerida juhuslikke suurusi ühtlasest jaotusest $\mathcal{U}(0; 0{,}2)$ tõenäosusega $25{,}5 \cdot 0{,}02$ ja ühtlasest jaotusest $\mathcal{U}(0{,}02; 1)$ tõenäosusega $0{,}5 \cdot (1 - 0{,}02)$. Jaotusega $\mathcal{S}_1$ juhuslike suuruste genereerimine on analoogiline – tõenäosusega $0{,}8 \cdot 0{,}98$ tuleb genereerida juhuslikke suuruseid ühtlasest jaotusest $\mathcal{U}(0; 0{,}98)$ ja tõenäosusega $10{,}8 \cdot (1 - 0{,}98)$ ühtlasest jaotusest $\mathcal{U}(0{,}98; 1)$.

Kaheastmelise treppfunktsiooni kasutamine kvaliteediskoori lähendamisel on hea, kuna see on lihtsasti interpreteeritav – tuvastatakse umbes 50% valepositiivsetest CNV leidudest, umbes 20% õigepositiivsetest CNV leidudest ning muudel juhtudel genereeritakse skooriks juhuslik suurus ühtlasest jaotusest 0 ja 1 vahel. Samuti on kaheastmelise treppfunktsiooni puhul ülesobitamise oht väike ning see on oma lihtsuse tõttu lihtsasti üldistatav ka teistele kohortidele peale TÜ EGV.

Simuleerimine

Selleks, et hinnata kvaliteediskoori kasutamisel statistilise testi võimsust, tuleb genereerida suur arv m korda juhuslikke suurusi $Z_{0,i},\ \ i = 1, \ldots, n_0$ ning $Z_{1,j},\ j = 1, \ldots, n_1$, kus

$$Z_{0,i} = \begin{cases} S_1 \leftarrow \mathcal{S}_1, & \text{tõenäosusega } p_0 p_{11} \\ S_0 \leftarrow \mathcal{S}_0, & \text{tõenäosusega } (1 - p_0) p_{01} \\ 0, & \text{muidu} \end{cases}$$

ning

$$Z_{1,j} = \begin{cases} S_1 \leftarrow \mathcal{S}_1, & \text{tõenäosusega } p_1 p_{11} \\ S_0 \leftarrow \mathcal{S}_0, & \text{tõenäosusega } (1 - p_1) p_{01} \\ 0, & \text{muidu,} \end{cases}$$

kus n_0 ja n_1 on nagu ennegi vastavalt kontrollide ja juhtude arvud ning p_0 ja p_1 on CNV osakaalud vastavalt kontrollide ja juhtude seas. Siis tuleb iga genereeritud valimi jaoks läbi viia statistiline analüüs (näiteks logistiline regressioonanalüüs) ning seejärel saab empiirilise võimsuse arvutada kui nende testide osakaalu, mille puhul p-väärtus on väiksem valitud olulisuse nivoost α:

$$\text{võimsus} = P(P_t < \alpha | p_0 \neq p_1) = \frac{1}{m} \sum_{k=1}^{m} I(P_{k,t} < \alpha),$$

kus $P_{k,t}$ on statistilise testi t p-väärtus k-nda valimi korral.

Juhul kui $p_0 = p_1$, saab eelneva valemi abil arvutada empiirilise I liiki vea tõenäosuse.

Kood kvaliteediskoori simuleerimiseks on toodud lisas 2.

3 Andmete kirjeldus

Analüüs CNV-de ja ravimi kõrvaltoimete tekke vahelise seose uurimiseks viiakse läbi nii TÜ EGV kui ka UKB geenidoonorite andmetega. Geenidoonorite kohta on olemas CNV-de ja ravimi kõrvaltoimete diagnooside andmed.

TÜ EGV-s on info üle 52 000 geenidoonori kohta. Nendest geenidoonoritest 66,9% on naised ja 33,1% mehed. Nende vanus geenivaramuga liitumise ajal on olnud vahemikus 18-103 aastat, kusjuures keskmine vanus oli 44,2 aastat.

UKB-s on üle 500 000 geenidoonori andmed. Doonorite enamiku (88%) moodustavad Euroopa päritolu etnilisest grupist inimesed ning töös kasutatakse andmeid vaid nende kohta. Euroopa päritolu doonoritest 54,3% on naised ja 45,7% on mehed. Doonorite vanus liitumise ajal oli vahemikus 39-73 aastat ning keskmine vanus 56,8 aastat.

Ülesehituselt on töös kasutatavad TÜ EGV ja UKB andmestikud ühesugused.

Koopiaarvu variatsioonide andmestik

CNV-d on TÜ EGV ja UKB geenidoonoritele leitud PennCNV programmi abil. PennCNV väljundis on igal real info ühe CNV kohta. Analüüsi läbiviimise jaoks kasutatakse CNV-de kohta järgmisi tunnuseid:

- CNV-d kandva indiviidi kood;
- kromosoom, millel CNV asub;
- CNV alguskoordinaat kromosoomil;
- CNV lõppkoordinaat kromosoomil;
- koopiaarv (deletsiooni korral 0 või 1 ning duplikatsiooni korral 3 või 4);
- CNV kvaliteediskoor.

CNV-de andmestikule on juba eelnevalt tehtud kvaliteedikontroll ning kõik nõuetele mittevastavad read on andmestikust eemaldatud. TÜ EGV-s läbivad kvaliteedikontrolli indiviidid, kelle

puhul on täidetud järgmised tingimused:

- geneetiline sugu vastab geenidoonori ankeedis olevale soole;
- genotüüp on määratud vähemalt 98% genotüpiseerimise kiibi peal olevatest positsioonidest;
- heterosügootsete genotüüpide osakaal vastab ligikaudu kogu andmestiku keskmisele ehk jääb vahemikku keskmine $\pm$ 3 standardhälvet;
- ei ole leitud üle 200 CNV piirkonna (väga suur CNV-de arv viitab kehva kvaliteediga proovile, mille puhul enamik CNV-dest on valepositiivsed).

UKB andmete puhul on läbi viidud analoogiline kvaliteedikontroll. Lisaks eelnevalt välja toodud tingimustele kontrollitakse UKB puhul ka seda, kas doonori ankeedile märgitud rass vastab tema geneetilisele rassile.

Ravimi kõrvaltoimete andmestik

Ravimite kõrvaltoimed on tähistatud ICD-10-koodide abil. ICD-10 on rahvusvahelise haiguste klassifikatsiooni kümnes versioon. ICD (*The International Classification of Diseases*) eesmärk on teha võimalikuks erinevates riikides kogutud andmete süstemaatiline analüüsimine ja võrdlemine. Kõikidel haigusseisunditel, sealhulgas ravimite kõrvaltoimetel, on oma kood, mis koosneb ühest tähest ja kahest numbrist, millele võib järgneda ka täpsustav arv (Küng ja Bogovski, 1996). Näiteks ICD-10-kood L23 tähistab allergilist kontaktdermatiiti ja kood L23.3 tähistab nahale toimivate ravimite põhjustatud allergilist kontaktdermatiiti.

TÜ EGV geenidoonorite ravimite kõrvaltoimete andmestik on kokku pandud info põhjal, mis on pärit doonorite enda täidetud küsimustikest, Eesti Haigekassa retseptidest, Tartu Ülikooli Kliinikumist, Põhja-Eesti Regionaalhaiglast, E-tervise andmetest ja surma- ning vähiregistrist. Kui TÜ EGV doonoril on vähemalt ühes allikas vähemalt ühe korra mingi ravimi kõrvaltoime diagnoositud, on ta selle kõrvaltoime suhtes juht. Juhul kui doonoril ei ole konkreetset kõrvaltoimet diagnoositud ning tal ei ole E-tervise ega Haigekassa andmetes ka ühegi teise haiguse

diagnoosi, siis määrati ta selle haiguse osas NA-ks (puuduv väärtus, *not available*), kuna ei saa olla kindel, kas doonori andmed olid ikka geenivaramu andmebaasis olemas. NA-ks märgiti doonor ka siis, kui haigus oli küll diagnoositud, kuid seda oli tehtud ebamääraselt, näiteks oli antud mingi ICD-koodide vahemik. Kui doonor oli mingi kõrvaltoime osas juht, siis ülejäänud kõrvaltoimete osas muudeti ta vastava kõrvaltoime analüüsi ajaks NA-ks. Kui doonor ei ole juht ega NA, siis on ta vastava kõrvaltoime osas kontroll.

Analoogilised erinevate terviseregistrite andmed on olemas ka UKB geenidoonorite jaoks.

Kõrvaltoimete andmestikus on igas reas info ühe geenidoonori kohta. Veergudes on kirjas 85 erinevat kõrvaltoimete ICD-10-koodi, mille väärtusteks on kas 0 (kõrvaltoimet ei esinenud), 1 (kõrvaltoime esines) või NA. UKB puhul on ravimi kõrvaltoime tüübina kirjas ka doonori enda poolt teada antud ravimi kõrvaltoime (*self reported ADR*). Peale selle on andmestikus kuus tunnust, millesse on koondatud omavahel seotud kõrvaltoimed ning mille väärtusteks on samuti 0, 1 või NA (tabel 1). Lisaks diagnoosidele on igas reas kirjas indiviidi kood, sünniaasta, sugu, genotüpiseerimisel kasutatud kiip ning genotüübiandmetelt leitud peakomponendid.

Tabel 1. Grupeeritud ravimi kõrvaltoimed

Grupp	Diagnooside ICD-10-koodid
Kesknärvisüsteemi toksilisus	G25.4, G25.6, G44.4
Kilpnäärmega seotud kõrvaltoimed	E03.2, E06.4, E23.1
Kontaktdermatiit	L23.3, L24.4, L25.1
Nahaga seotud kõrvaltoimed (L-koodid)	L23.2, L24.4, L25.1, L27.0, L27.1, L56.0, L64.0
Kõrvaltoimed toimeaine järgi (Y-koodid)	Y4*, Y5*
Ülitundlikkus	L10.5, L27.0, L27.1, M32.0, M34.2, T88.6, Z88

TÜ EGV geenidoonorite ravimi kõrvaltoimete andmed on seisuga 31.12.2015 ja UKB andmed seisuga 19.10.2017.

Farmakogeenide andmestik

Analüüsi jaoks on kasutatud ka farmakogeenide andmestikku. Selles andmestikus on 65 PharmGKB VIP-ide nimekirjast pärit farmakogeeni, mille kohta on olemas järgmised tunnused:

– geeni nimi;

– kromosoom, millel see geen asub;

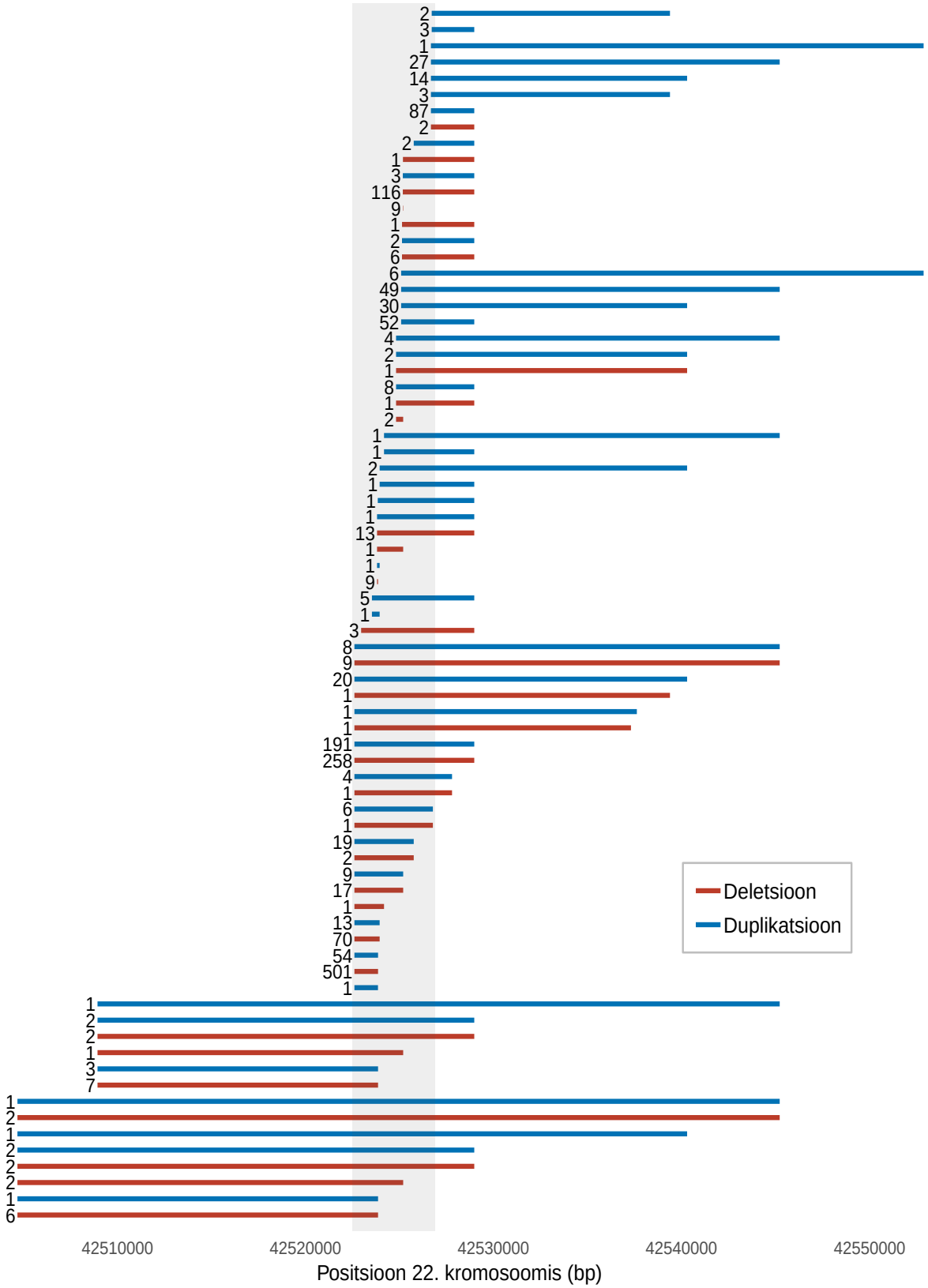
– geeni alguskoordinaat kromosoomil;

– geeni lõppkoordinaat kromosoomil.

Geenide ja CNV-de ülekatte andmestik

CNV-de ning farmakogeenide andmete põhjal koostati uus andmestik, kus on info farmakogeenide ja nendega ülekattes olevate CNV-de kohta. Iga geeni kohta leiti CNV-de andmestikust need read, milles olev CNV asus antud geeniga samal kromosoomil ning mille kromosoomipiirkond kattus mingis osas selle geeni asukohaga kromosoomil. Seletav CNV tunnus hõlmab seega kõiki konkreetse farmakogeeniga ülekattes olevaid CNV-sid. Edaspidi nimetame seda lihtsalt CNV tunnuseks.

Joonisel 3 on näha TÜ EGV geenidoonorite CNV-d, mis on ülekattes CYP2D6 geeniga.



Joonis 3. Geeniga CYP2D6 (22: 42522501-42526883) ülekattes olevad CNV-d TÜ EGV doonoritel. Punased ja sinised lõigud joonisel tähistavad vastavalt deletsioone ja duplikatsioone ning arvud nende kõrval tähistavad vastavat CNV-d kandvate doonorite arvu. Hall osa tähistab CYP2D6 asukohta 22. kromosoomil.

4 Kasutatav analüüsimetoodika

4.1 Logistiline regressioonimudel

Ravimi kõrvaltoime on binaarne tunnus – sellel on kaks võimalikku väärtust: esineb/ei esine. Binaarse sõltuva tunnuse ja mingi sõltumatu tunnuse vahelise seose leidmiseks kasutatakse tavaliselt logistilist regressiooni. Logistilise regressioonimudeliga prognoositakse uuritava sündmuse toimumise tõenäosust sõltuvalt argumenttunnuste väärtuste muutumisest.

Olgu Y meile huvipakkuv sündmus ehk ravimi kõrvaltoime tekkimine. Tähistame sündmuse esinemise tõenäosust $P(Y = 1) = \pi$. Binaarse uuritava tunnuse puhul on kasutusel logit-seosefunktsioon

$$\text{logit}(\pi) = \ln \frac{\pi}{1-\pi}, \quad \text{kus } \frac{\pi}{1-\pi} \text{ on sündmuse toimumise šanss.}$$

Sündmuse toimumise šanss näitab, mitu korda on uuritava sündmuse toimumine tõenäolisem kui sündmuse mittetoimumine. Logistiline mudel hindab sündmuse esinemise logaritmitud šansi

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

kus $\beta_0, \beta_1, \ldots, \beta_k$ (k on argumenttunnuste arv) on mudeli tundmatud parameetrid ning $x_1, \ldots, x_k$ on argumenttunnused. Argumenttunnusteks on selles töös CNV-d ja kovariaatidena lisatakse mudelisse ka näiteks sugu ja sünniaasta.

Tundmatute parameetrite β_j $(j = 0, \ldots, k)$ hindamiseks valimi põhjal kasutatakse suurima tõepära meetodit. Kui hinnatud parameeter β_j on positiivne, siis on vastava argumendi ja uuritava tunnuse vahel samapidine seos, kui negatiivne, siis vastupidine seos.

4.2 Suurima tõepära meetod

Suurima tõepära meetodi korral leitakse parameetritele sellised väärtused, mis maksimeerivad olemasoleva valimi saamise tõepära. Järgnevalt kirjeldatakse suurima tõepära hinnangute leidmist logistilise regressioonimudeli korral.

Olgu $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ hinnatavate parameetrite vektor ning olgu i-nda vaatluse argumenttunnuste vektor koos vabaliikmega $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$, kus $i = 1, \ldots, n$ ning n on vaatluste arv. Iga vaatluse korral on olemas binaarse uuritava tunnuse väärtus y_i, mille puhul $P(y_i = 1) = \pi_i$.

Valimi i-nda vaatluse $(y_i, \boldsymbol{x}_i)$ tõepära on kujul

$$\pi_i^{y_i}(1 - \pi_i)^{1-y_i},$$

kus

$$\pi_i := \pi(\boldsymbol{x}_i\boldsymbol{\beta}) = \frac{\mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + \mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}}.$$

Eeldusel, et valimi vaatlused on sõltumatud, avaldub kogu valimi tõepärafunktsioon vaatluste tõepärade korrutisena:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}.$$

Vastav maksimeeritav logaritmitud tõepärafunktsioon on kujul

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \ln \frac{\pi_i}{1 - \pi_i} + \ln(1 - \pi_i) \right] = \sum_{i=1}^{n} \left[y_i \boldsymbol{x}_i\boldsymbol{\beta} - \ln(1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})) \right].$$

Selleks, et leida tõepära maksimeerivad parameetrid, võetakse logaritmitud tõepärafunktsioonist tuletis $\boldsymbol{\beta}$ järgi, võrdustatakse see nulliga ning avaldatakse $\boldsymbol{\beta}$.

Olgu $U(\boldsymbol{\beta})$ logaritmitud tõepärafunktsiooni esimene tuletis $\boldsymbol{\beta}$ järgi. Siis

$$U(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \pi_i).$$

$U(\boldsymbol{\beta})$ nimetatakse ka skoorifunktsiooniks. Seega hinnangud parameetervektorile $\boldsymbol{\beta}$ leitakse võrrandist $U(\boldsymbol{\beta}) = 0$.

Hesse maatriksiks ehk hessiaaniks $H(\boldsymbol{\beta})$ nimetatakse logaritmitud tõepärafunktsiooni teist järku tuletiste maatriksit. Pannes tähele, et

$$\frac{\partial \pi_i}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \frac{\mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + \mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}} = \boldsymbol{x}_i \frac{\mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + \mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}} \left(1 - \frac{\mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + \mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}}}\right) = \boldsymbol{x}_i\pi_i(1 - \pi_i),$$

on vastav Hesse maatriks kujul

$$\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i \pi_i(1 - \pi_i).$$

Arvestades, et $\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i = \mathbf{X}^T\mathbf{X}$, kus $\mathbf{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$, ja tähistades $\mathbf{W}$-ga diagonaalse maatriksi, mille peadiagonaali elemendid on $\pi_i(1 - \pi_i)$, kus $i = 1, \ldots, n$, saame hessiaani kirjutada maatrikskujul

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X}.$$

Hinnangute arvutamiseks kasutatakse iteratiivseid meetodeid. Logistilise regressioonimudeli saab R-is hinnata funktsiooniga *glm*, mille puhul arvutatakse hinnangud Fisheri skoorimeetodil.

Fisheri skoorimeetodi algoritm on

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j + \mathbf{I}^{-1}(\boldsymbol{\beta}_j)\mathbf{U}(\boldsymbol{\beta}_j),$$

kus $\boldsymbol{\beta}_j$ on parameetervektori $\boldsymbol{\beta}$ väärtus j-ndal iteratsioonisammul ja

$$\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

nimetatakse Fisheri informatsioonimaatriksiks.

4.3 Eraldavuse probleem ja Firth'i meetod

Nii CNV-d kui ka ravimite kõrvaltoimete diagnoosid on harvad. Seega võib valimis olla vähe neid, kes on kõrvaltoime osas juhud, ning veelgi vähem võib olla CNV-d kandvaid juhtusid. Kui juhtusid on vähe, siis tavalise logistilise regressiooni puhul kasutatav suurima tõepära meetod annab nihkega hinnanguid (Firth, 1993).

Samuti põhjustab juhtude väike arv kas osalist või täielikku eraldavust (*quasi-complete separation, complete separation*). Eraldavus on olukord, kus seletav tunnus (või seletavate tunnuste komplekt) prognoosib (peaaegu) täpselt binaarse uuritava tunnuse väärtuse. Kui uuritav tunnus on diagnoosi olemasolu ning seletav tunnus sugu, siis täieliku eraldavusega on tegemist näiteks juhul, kui valimis igal mehel on diagnoos ning ühelgi naisel pole diagnoosi. Osalise eraldavusega on tegemist aga näiteks siis, kui igal mehel on diagnoos, kuid naisi on nii diganoosiga kui diagnoosita.

Eraldavus on logistilise regressiooni puhul sage probleem. Eriti tihti tuleb seda ette väikeste valimite korral. Eraldavus esineb aga ka suurtes valimites; näiteks siis, kui mõni seletav tunnus on uuritava tunnusega väga tugevalt seotud või kui seletavad tunnused on kvalitatiivsed ja uuritava tunnuse puhul esineb ühte väärtustest väga vähe.

Kui andmetes esineb kas täielik või osaline eraldavus, siis huvipakkuva parameetri hindamiseks kasutatav iteratsiooniprotsess ei koondu. Sel juhul suurima tõepära hinnangut parameetrile ei leidu ning tagastatakse n-ö lõpmatu parameeter.

Eespool mainitud probleemide korral võib tavalise logistilise regressiooni asemel kasutada Firth'i meetodit (Firth, 1993). Firth'i meetodi puhul maksimeeritakse tavalise tõepära $L(\boldsymbol{\beta})$ asemel parandusliikmega tõepära (*penalized likelihood*)

$$L(\boldsymbol{\beta})^* = L(\boldsymbol{\beta})|\mathbf{I}(\boldsymbol{\beta})|^{1/2}$$

ning vastavat log-tõepära

$$l(\boldsymbol{\beta})^* = l(\boldsymbol{\beta}) + \frac{1}{2}\ln|\mathbf{I}(\boldsymbol{\beta})|.$$

Vastav skoorifunktsioon on sel juhul

$$U(\boldsymbol{\beta})^* = U(\boldsymbol{\beta}) + \frac{1}{2}tr\left(\mathbf{I}^{-1}(\boldsymbol{\beta})\frac{\partial \mathbf{I}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right).$$

Firth'i meetodi puhul koondub iteratsioonimeetod kindlasti ning parameetritele leiduvad alati lõplikud hinnangud (Heinze ja Schemper, 2002).

4.4 Populatsiooni struktuur ja sugulus

Tulenevalt sellest, et partneri valik järglaste saamiseks ei ole juhuslik (näiteks geograafilise eraldatuse tõttu), esineb erinevate populatsioonide geno- ja fenotüübis tihti süstemaatilisi erinevusi. Struktureeritust esineb isegi väikeste populatsioonide ja riikide siseselt, näiteks on tuvastatud Eesti-sisene struktureeritus geenidoonorite sünnimaakonna järgi (Nelis *et al*., 2009). Populatsiooni struktuur on segav tunnus, mis võib mõjutada nii uuritavat genotüüpi (CNV-d) kui fenotüüpi (ravimi kõrvaltoimet) ning tekitada nende vahele seeläbi fiktiivse seose. Kui assotsiatsiooniuuringutes populatsiooni struktuuri arvesse ei võeta, siis võidakse leida valepositiivseid seoseid ja lisanduva varieeruvuse tõttu võivad avastamata jääda ka reaalsed seosed (Marchini *et al*., 2004).

Assotsiatsiooniuuringutes populatsiooni struktuuriga arvestamiseks lisatakse regressioonimudelisse kovariaatidena peakomponendid, mis on leitud genotüübiandmete (enamasti SNV-de) pealt tehtud peakomponentanalüüsi (PCA, *principal component analysis*) abil. PCA on mitmemõõtmelise statistika meetod, mille eesmärk on leida esialgsete tunnuste lineaarkombinatsioonid, mis kirjeldaksid võimalikult suure osa esialgsete näitajate varieeruvusest. Neid uusi tunnuseid nimetatakse peakomponentideks ning need on valitud nii, et esimene peakomponent kirjeldab maksimaalse võimaliku osa algsete tunnuste varieeruvusest, teine peakomponent on

esimesega mittekorreleeritud ning kirjeldab võimalikult suure osa allesjäänud varieeruvusest ja nii edasi. PCA abil saab suuremõõtmelisi andmeid paremini visualiseerida ning näha nende andmete klasterdumist. Genotüübiandmete pealt arvutatud esimeste peakomponentide järgi klasterduvad indiviidid populatsiooni struktuuri (näiteks elukohast tingitud) alusel (Nelis *et al*., 2009).

Geenidoonorite seas on palju ka inimesi, kes on omavahel lähisugulased ja seega geneetiliselt väga sarnased. Sõltumatuse eeldus logistilise regressiooni ja Firth'i meetodi korral ei ole nende inimeste puhul täidetud ning peakomponentide lisamine mudelitesse ei aita. Probleemi lahendamiseks jäetakse analüüsi jaoks alles vaid need doonorid, kes pole omavahel sugulased.

Sugulaste määramiseks kasutatakse programmi Plink 1.9 (Purcell *et al*., 2007), mille abil 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 andmestikust välja jätta. Kui kahe lähisugulase korral ühe kohta on olemas info ravimi kõrvaltoime kohta ja teise kohta mitte, siis jäetakse andmestikku alles just esimene.

4.5 Logistiline segamudel

Indiviididevahelise suguluse arvestamiseks kasutatakse ka segamudeleid, mis lubavad vaatlustevahelist sõltuvust. Kuna segamudelite jaoks ei tule indiviide eemaldada, siis suurema valimimahu tõttu on ka statistiline võimsus seoste avastamiseks suurem. Segamudelid sisaldavad fikseeritud ja juhuslikke mõjusid; geneetilise sõltuvuse ehk suguluse efekt modelleeritakse juhusliku vektorina, uuritavad geneetilised variandid ja kovariaadid aga fikseeritud efektidena.

Olgu endiselt uuritav tunnus Y ravimi kõrvaltoime, mille väärtus iga indiviidi puhul on kas 1 või 0 vastavalt sellele, kas tegemist on juhu või kontrolliga. Olgu n indiviidide arv valimis ja p kovariaatide arv.

Vaatame logistilist segamudelit

$$\text{logit}(\pi_i) = \boldsymbol{x}_i\boldsymbol{\alpha} + g_i\beta + b_i,$$

kus $\pi_i = P(y_i = 1|\boldsymbol{x}_i, g_i, b_i)$ on i-nda indiviidi juhuks olemise tõenäosus, $\boldsymbol{x}_i$ on $1 \times (p+1)$ kovariaatide vektor koos vabaliikmega, $\boldsymbol{\alpha}$ on $(p+1) \times 1$ kovariaatide ja vabaliikme efektide vektor, g_i on i-nda indiviidi genotüüp uuritava geneetilise variandi (CNV) osas ning β on selle CNV efekt.

Eeldame, et $\boldsymbol{b} = (b_1, \ldots, b_n)^T$ on juhuslike efektide vektor, kus $b_i = \sum_{j=1}^{m} V_{ij}u_j$, ja seega $\boldsymbol{b} = \mathbf{Vu}$. Siin $\mathbf{V}$ on $m \times n$ standardiseeritud SNV-de maatriks, kus m on SNV-de arv, $\boldsymbol{u} = (u_1, \ldots, u_m)^T$ on SNV-de efektide vektor ja $\boldsymbol{u} \sim \mathcal{N}(0, \sigma_b^2\mathbf{I})$, kus σ_b^2 tähistab aditiivset geneetilist variatsiooni. Seega $\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \sigma_b^2\mathbf{V}\mathbf{V}^T)$. Maatriksit $\boldsymbol{\Psi} = \mathbf{V}\mathbf{V}^T$ nimetatakse sugulusmaatriksiks. Sugulusmaatriksi element $(\boldsymbol{\Psi})_{ij}$ on arv, mis kirjedab i-nda ja j-nda doonori vahelist geneetilist sarnasust (Lippert *et al*., 2011).

SAIGE

Logistiliste segamudelite hindamine on väga arvutusintensiivne ja seda just suurte geeniuuringute korral, mille puhul on vaatlusi palju ning iga uuritava variandi jaoks on vaja hinnata uus mudel.

R-is on olemas pakett SAIGE (Scalable and Accurate Implementation of GEneralized mixed model; Zhou *et al*., 2018), mis hindab logistilise segamudeli, kasutades arvutuste lihtsustamiseks erinevaid optimeerimismeetodeid. SAIGE on loodud hindamaks SNV-de seoseid, kuid selles töös kasutatakse seda CNV-de seoste hindamiseks. Logistiline segamudel hinnatakse kahe sammuga. Kõigepealt kasutatakse kvaasi-tõepära meetodit, et sobitada nullmudel kujul

$$\text{logit}(\pi_{i0}) = \boldsymbol{x}_i\boldsymbol{\alpha} + b_i$$

ning leida hinnangud $\hat{\pi}_{i0}$ tõenäosustele $\pi_{i0} = P(y_i = 1|\boldsymbol{x}_i, b_i)$.

Teise sammuna hinnatakse seos iga uuritava geneetilise variandi (CNV) ja fenotüübi (ravimi kõrvaltoime) vahel. Nullhüpoteesi $H_0 : \beta = 0$ kontrollimiseks leitakse skoor ehk log-tõepära

tuletis β järgi kujul

$$T = \sum_{i=1}^{n} g_i(y_i - \hat{\pi}_{i0}).$$

Teises sammus kasutatakse skooritesti läbiviimiseks sadulpunkti lähendamist (SPA, *saddlepoint approximation*). Traditsiooniliselt lähendatakse skoorifunktsiooni jaotust normaaljaotusega, kasutades keskväärtust ja dispersiooni. SPA aga kasutab jaotuse hindamiseks kumulante genereerivat funktsiooni. Juhusliku suuruse kumulante genereeriv funktsioon (CGF, *cumulant generating function*) on naturaallogaritm vastavast momente genereerivast funktsioonist.

Arvestades, et $y_i \sim \mathrm{B}(1,\pi_i)$, avaldub skoori T kumulante genereeriv funktsioon kujul

$$K(t) = \ln \mathrm{E}(\mathrm{e}^{tT}) = \sum_{i=1}^{n} \ln\left(1 - \hat{\pi}_i + \hat{\pi}_i \, \mathrm{e}^{t g_i}\right) - t \sum_{i=1}^{n} g_i \hat{\pi}_i.$$

Testi p-väärtuse arvutamiseks kasutatakse valemit

$$P(T < q) \approx \Phi\left(w + \frac{1}{w} \ln \frac{v}{w}\right),$$

kus q on arvutatud teststatistiku väärtus, $w = \mathrm{sgn}(\hat{t})\sqrt{2(\hat{t}q - K(\hat{t}))}$, $v = \hat{t}\sqrt{K''(\hat{t})}$, $K'(t)$ ja $K''(t)$ on vastavalt K esimene ja teine tuletis, $\hat{t}$ nimetatakse sadulpunktiks ja on võrrandi $K'(\hat{t}) = q$ lahend ning Φ on standardse normaaljaotuse jaotusfunktsioon (Kuonen, 1999).

SAIGE metoodikat kasutataksegi just hüpoteeside testimiseks, efektide suuruste leidmiseks kasutatakse selles töös logistilist regressiooni ja Firth'i meetodit.

4.6 Mitmese testimise probleem

Töös kontrollitakse mitmeid hüpoteesipaare ja see toob kaasa mitmese testimise probleemi. Mitmese testimise korral suureneb I liiki vea tegemise tõenäosus ja seega ka valepositiivsete oluliste seoste leidmine.

Näiteks, olgu ühe testi korral olulisuse nivooks $\alpha = 0{,}05$ ehk I liiki viga lubatakse teha 5%

tõenäosusega. Sel juhul tõenäosus, et saja sõltumatu testi korral ei tehta ühtegi I liiki viga, on $(1 - 0{,}05)^{100}$ ja seega tõenäosus, et saja testi korral tehakse vähemalt üks I liiki viga, on $1 - (1 - 0{,}05)^{100} \approx 99{,}4\%$.

I liiki vea kontrolli all hoidmiseks kasutatatakse Bonferroni parandust. Bonferroni parandus põhineb teadmisel, et I liiki vea tõenäosus n testi korral on väiksem või võrdne kui n üksiku testi I liiki vea tõenäosuste summa. Seega, piiramaks kõigi n testi puhul ühe või enama vea tegemise tõenäosust olulisuse nivooga α, peab iga üksiku testi olulisuse nivoo olema α/n.

5 Tulemused

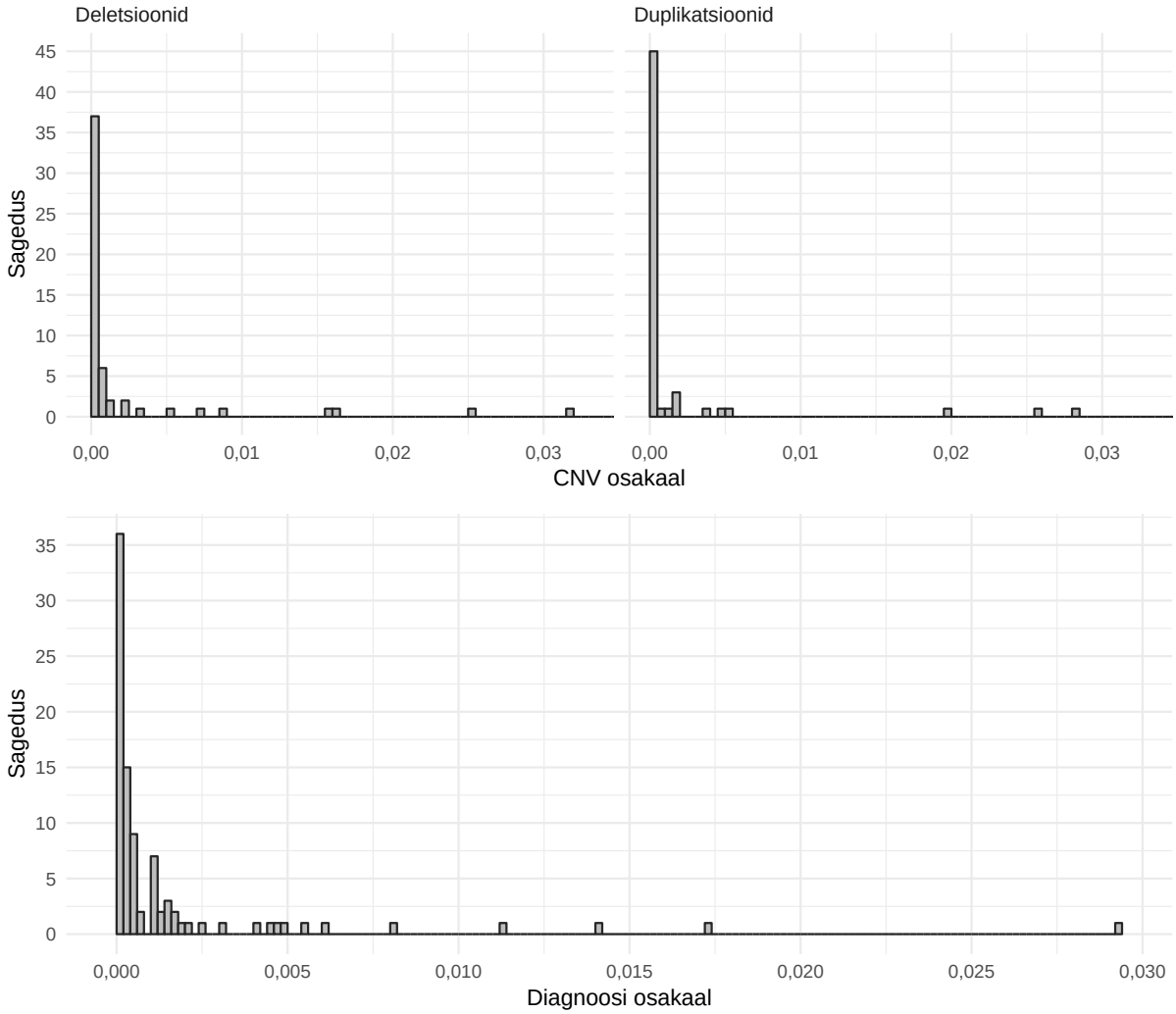
5.1 Kirjeldav analüüs

TÜ EGV andmestikus, millest sugulased on välja jäetud, on kokku 32 999 geenidoonori andmed. TÜ EGV doonoritel on CNV-sid ülekattes 56 erineva farmakogeeniga. Enamike farmakogeenidega on ülekattes väga vähe CNV-sid. Keskmine deletsioonide suhteline sagedus geenide kaupa on 0,525% ja duplikatsioonide suhteline sagedus geenide kaupa on 0,174%. Seejuures on 32 farmakogeeni puhul deletsioonide osakaal alla 0,1% ning 36 farmakogeeni puhul on duplikatsioonide osakaal alla 0,1% (joonis 4). Kõige rohkem oli deletsioone ülekattes geeniga KCNH2 (deletsioonide suhteline sagedus 16,7%, duplikatsioonide suhteline sagedus 2,56%) ja duplikatsioone geeniga CYP2E1 (duplikatsioonide suhteline sagedus 2,84%, deletsioonide suheline sagedus 0,0727%).

Harvad on ka ravimi kõrvaltoimete diagnoosid. Keskmine diagnooside suhteline sagedus on 0,167% ning 63 diagnoosi esinevad vähem kui 0,1% doonoritest (joonis 4). Kõige sagedasem diagnoos TÜ EGV andmestikus oli grupeeritud kõrvaltoime, mille alla kuuluvad kõik naha- ja nahaaluskoe haigustega seotud ravimi kõrvaltoimete ICD-koodid (suhteline sagedus 2,93%).

UKB ilma sugulasteta andmestikus on kokku 372 133 doonori andmed. UKB andmestikus on CNV-sid ülekattes 64 erineva farmakogeeniga. Ka nende andmete puhul on näha, et CNV-d esinevad harva. Keskmiselt on ühe farmakogeeniga ülekattes olevate deletsioonide osakaal 0,0581% ning duplikatsioonide osakaal 0,156%. 55 farmakogeeni puhul on nii deletsioonide kui duplikatsioonide osakaal alla 0,1%. Kõige rohkem on deletsioone ülekattes farmakogeeniga CYP2D6 (deletsioonide osakaal 1,48%, duplikatsioonide osakaal 0,508%) ja duplikatsioone farmakogeeniga KCNH2 (duplikatsioonide osakaal 2,92%, deletsioonide osakaal 0,114%).

Ravimi kõrvaltoimete diagnooside keskmine suhteline sagedus UKB geenidoonorite puhul on 0,237%, kusjuures 75 diagnoosi osakaal on alla 0,1%. Sagedaseim diagnoos on grupeeritud kõrvaltoime, mille alla kuuluvad kõik ülitundlikkusega seotud ravimi kõrvaltoimed (suhteline sagedus 6,96%).

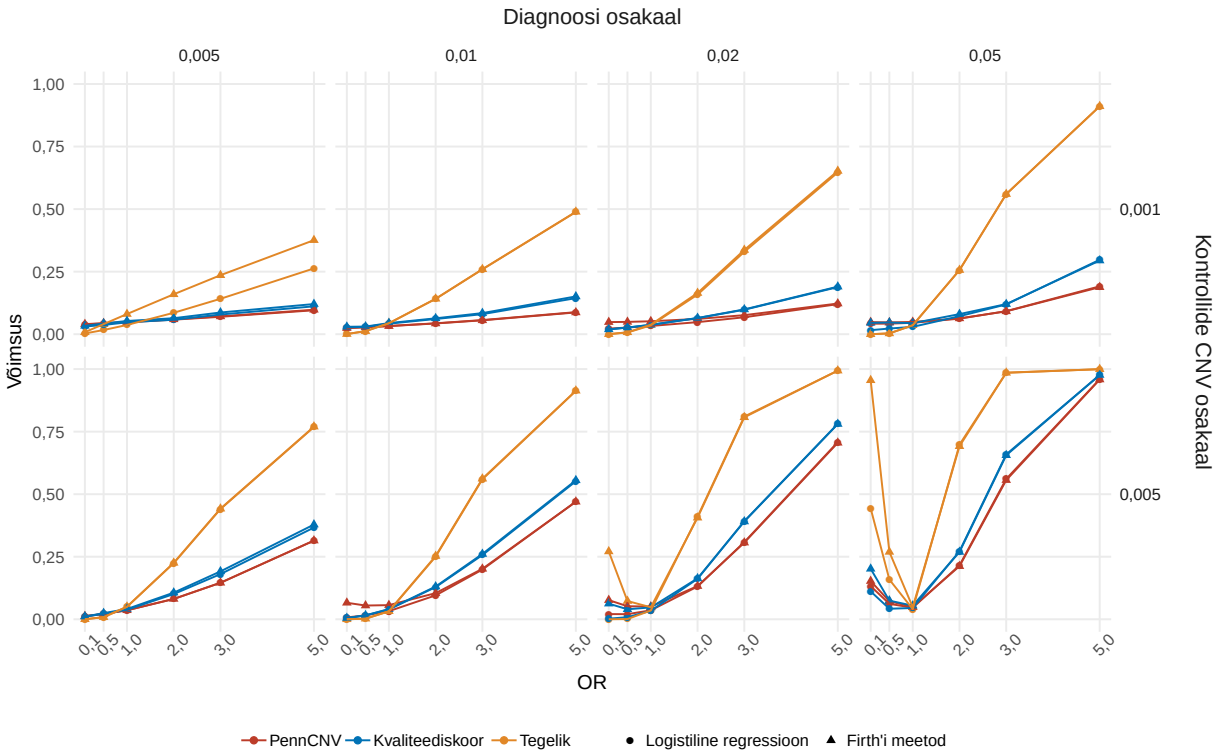


Joonis 4. CNV-de (üleval) ja ravimi kõrvaltoimete diagnooside (all) osakaalude jaotused TÜ EGV andmete põhjal. Deletsioonide histogrammi puhul pole arvestatud ühe farmakogeeniga, millega ülekattes olevate deletsioonide osakaal on 0,167.

5.2 Simulatsioonid

Simulatsioonid viidi läbi erinevate kontrollide CNV osakaalude $p_0 \in (0{,}001; 0{,}005; 0{,}01; 0{,}02)$ ja ravimi kõrvaltoimete osakaalude $p \in (0{,}001; 0{,}005; 0{,}01; 0{,}02; 0{,}05)$ korral. Need osakaalud määrati vastavalt TÜ EGV ja UKB diagnooside ning CNV-de osakaalude jaotuste järgi (joonis 4). Simulatsioonide jaoks eeldati, et šansside suhted $OR \in (0{,}1; 0{,}5; 1; 2; 3; 5)$ ning valimimahuks valiti 33 000.

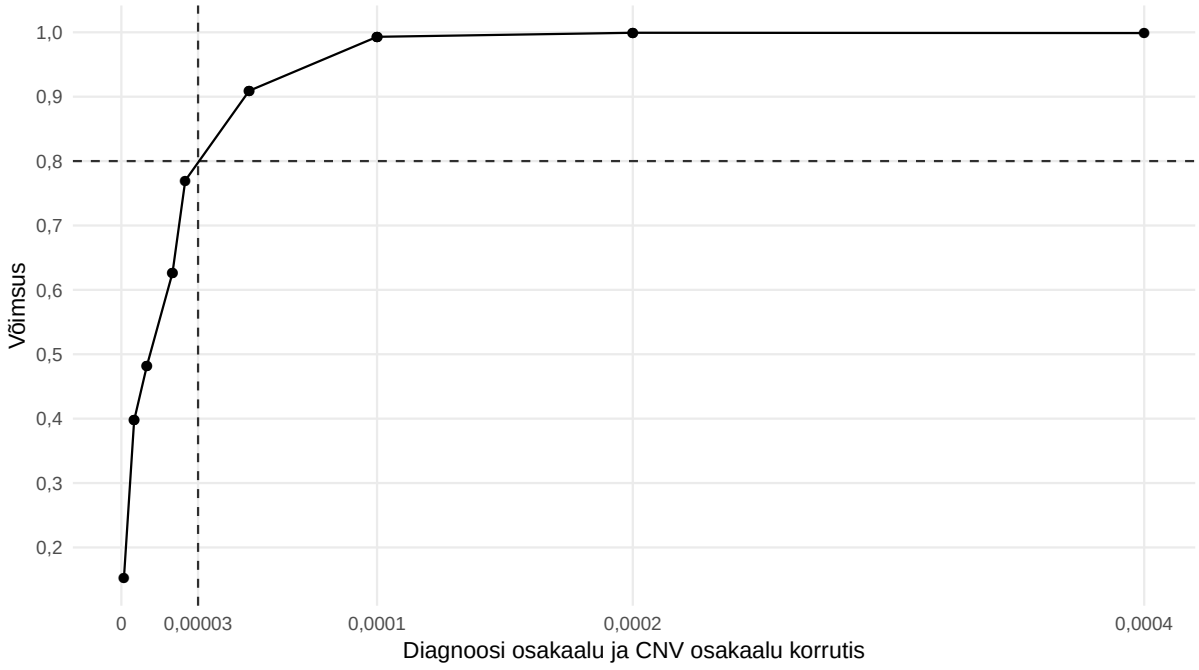
Simulatsioonide tulemused näitasid oodatavalt, et tegeliku CNV mitteteadmine vähendab oluliselt testimise võimsust. Samuti on näha, et kvaliteediskoori kasutamine PennCNV tunnuse asemel parandab testimise võimsust madala CNV sageduse korral (joonis 5). Näiteks kui diagnoosi osakaal on 0,05, CNV osakaal on 0,001 ning tegelik šansside suhe on 5, siis võimsuse erinevus PennCNV ja tegeliku CNV tunnuse vahel on 72,3%. Kvaliteediskoori kasutamine nende osakaalude korral aga parandab võrreldes PennCNV tunnusega võimsust 10,7% võrra.



Joonis 5. Simulatsioonide põhjal arvutatud (empiirilised) võimsused erinevate diagnooside osakaalude ja kontrollide CNV osakaalude korral.

Diagnoosi ja CNV piirsagedused määratakse nii, et testimise võimsus olulisuse nivool 0,05 oleks vähemalt 80%. Reaalsed efektide suurused ega CNV-de määramise valepositiivsuse ja -negatiivsuse määrad pole teada, seega peavad määratavad piirid olema pigem madalad. See tähendab, et kui reaalselt peaks kehtima, et mingi CNV korral on valenegatiivsete ja -positiivsete leidude määr 0 ja seos ravimi kõrvaltoimega on suur, siis peab olema võimalik selline efekt avastada ehk vastavat CNV-diagnoos paari tuleb testida.

Eelneva tõttu määratakse CNV ja diagnoosi piirsagedused eeldusel, et reaalne CNV on teada ja tegelik šansside suhe on 5. Jooniselt 5 on näha, et võimsus sõltub nii diagnoosi kui CNV osakaalust, st kui CNV osakaal on suur, siis võib diagnoosi osakaal olla väiksem ja vastupidi. Mõlema osakaalu korraga arvesse võtmiseks kasutatakse piiride määramiseks nende osakaalude korrutist. Joonisel 6 on näha, et 80% võimsuse saavutamiseks peab diagnoosi ja CNV osakaalude korrutis olema $3 \cdot 10^{-5}$. Kuna UKB valimimaht on umbes kümme korda suurem kui TÜ EGV oma, siis UKB puhul võeti osakaalude korrutise piiriks kümme korda väiksem arv ehk $3 \cdot 10^{-6}$. Umbkaudu sama piiri annavad ka simulatsioonid UKB valimimahuga (joonis L1 lisas 3).

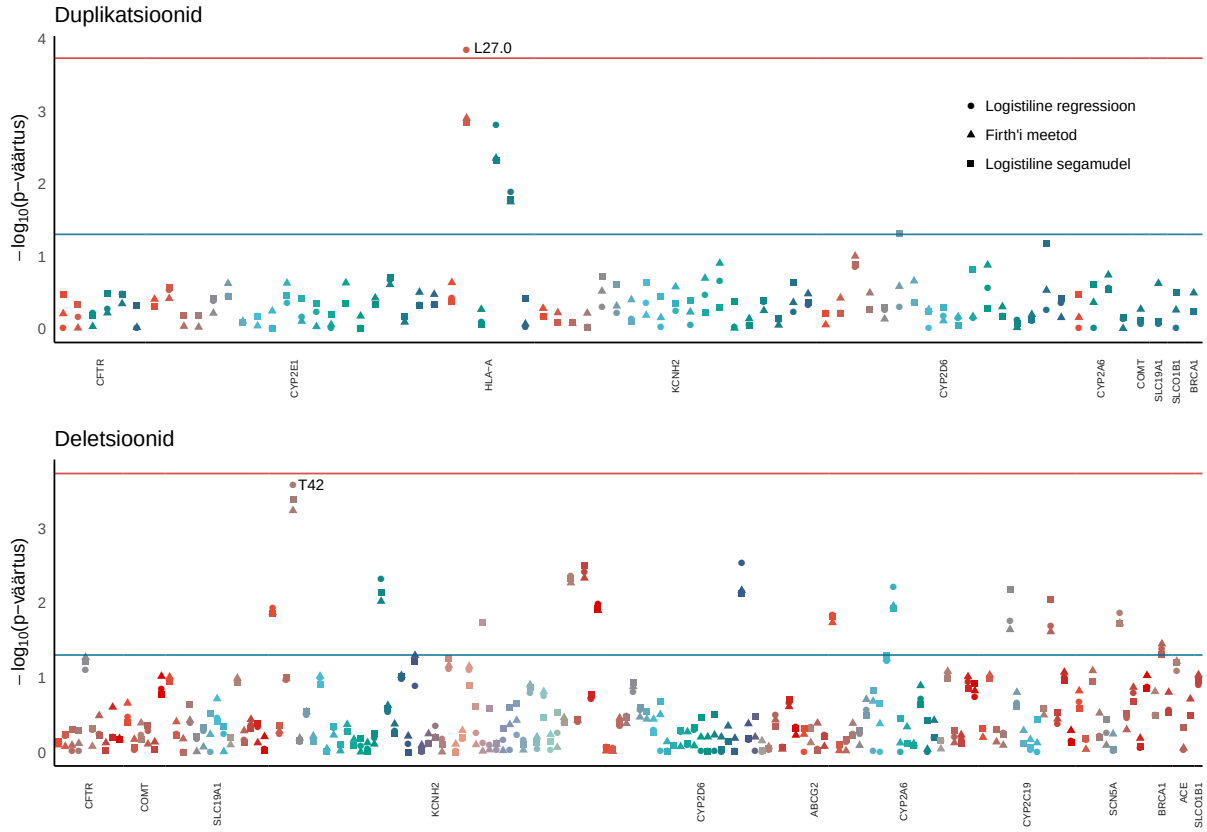


Joonis 6. Firth'i meetodi võimsus valimimahu N=33 000, OR=5 ja tegeliku CNV tunnuse korral. Horisontaalne ja vertikaalne kriipsjoon tähistavad vastavalt 80% võimsuse piiri ja osakaalude korrutise piiri 80% võimsuse saavutamiseks.

5.3 Analüüsi tulemused

Simulatsioonide tulemustest lähtuvalt kasutati CNV-de ja ravimi kõrvaltoimete vaheliste seoste analüüsimisel argumenttunnusena CNV kvaliteediskoori.

TÜ EGV andmete puhul kasutati analüüsideks tavalist logistilist regressioonimudelit, Firth'i meetodit ning logistilist segamudelit. Simulatsioonide tulemusi arvestades analüüsiti TÜ EGV andmete puhul neid CNV-diagnoos paare, mille korral CNV ja diagnoosi osakaalu korrutis oli vähemalt $3{\cdot}10^{-5}$. Sellisele tingimusele vastas 182 deletsioon-diagnoos paari ja 88 duplikatsioon-diagnoos paari. Korrigeeritud olulisuse nivooks on $0{,}05/(182+88) \approx 0{,}000185$. Enamike testitud CNV-diagnoos paari korral olid p-väärtused kõigi kolme meetodiga sarnased (joonis 7).



Joonis 7. Testide p-väärtused TÜ EGV andmete korral. Punktid on värvitud diagnooside järgi ning iga punktikolmik tähistab ühte testitud CNV-diagnoos paari. Punane joon vastab korrigeeritud olulisuse nivoole 0,000185 ning sinine joon olulisuse nivoole 0,05. Duplikatsioonide korral on korrigeeritud olulisuse nivoo suhtes oluline seos HLA-A ja L27.0 vahel. Deletsioonide korral on korrigeeritud olulisuse nivoole kõige lähemal KCNH2 – T42 testi p-väärtus.

TÜ EGV andmeid analüüsides leiti tavalise logistilise regressiooniga korrigeeritud olulisuse nivoo suhtes oluline seos geeniga HLA-A ülekattes olevate duplikatsioonide ja diagnoosi L27.0 vahel ($\beta = 3{,}88$, $SE = 1{,}02$, $p = 0{,}000143$, Firth'i $\beta = 4{,}09$, $SE = 0{,}968$,$p = 0{,}001242$, SAIGE $p = 0{,}001442$). ICD-10-kood L27.0 tähistab üldist nahalöövet rohtudest ja ravimitest. Firth'i meetodi ja logistilise segamudeli korral ei olnud see seos oluline. Kuna näitasime eelnevalt Firth'i meetodi paremust tavalise logistilise regressiooniga võrreldes, on võimalik, et logistilise regressiooniga saadud tulemus ei ole täpne. Farmakogeeniga HLA-A ülekattes olevate CNV-de ja diagnoosi L27.0 vahelist seost UKB valimiga ei testitud, kuna seal oli nii L27.0 kui HLA-A-ga ülekattes olevate CNV-de sagedus liiga väike.

TÜ EGV andmete analüüsimisel ei leitud olulisi seoseid deletsioonide ja ravimi kõrvaltoimete vahel. Vähimate p-väärtustega tulemusi tavalise logistilise regressiooni korral andsid järgmised deletsioon-diagnoos paarid: KCNH2 – T42 ($\beta = 0{,}937$, $SE = 0{,}257$, $p = 0{,}00026$), CYP2D6 – M80.4 ($\beta = 2{,}05$, $SE = 0{,}690$, $p = 0{,}0029$).

Logistilise segamudeli rakendamine ei andnud täiendavaid tulemusi. Põhjus võib olla selles, et sugulaste väljajätmine andmestikest ei ole juhuslik – eelistatult jäetakse välja kontrollid, mitte juhud. Seega logistilise segamudeli puhul on valimimaht küll suurem, kuid seda eelkõige kontrollide arvelt, mistõttu võimsus ei suurene.

UKB andmete puhul kasutati CNV ja ravimi kõrvaltoime vahelise seose uurimiseks tavalist logistilist regressioonanalüüsi ja Firth'i meetodit. Logistilist segamudelit UKB andmete puhul ei kasutatud, kuna seal moodustasid väljajäetud sugulased väga väikese osa kõikidest doonoritest ning seega poleks segamudeli kasutamine täiendavat võimsust lisanud. Simulatsioonide tulemusi arvestades analüüsiti UKB andmete puhul neid CNV-diagnoos paare, mille korral CNV ja diagnoosi osakaalu korrutis oli vähemalt $3 \cdot 10^{-6}$. Sellele tingimusele vastas 177 deletsioon-diagnoos paari ja 303 duplikatsioon-diagnoos paari. Korrigeeritud olulisuse nivooks on $0{,}05/(177 + 303) \approx 0{,}000104$.

Korrigeeritud olulisuse nivoo suhtes olulisi tulemusi ei olnud. Deletsioonide ja ravimi kõrvaltoimete testimisel olid vähimad p-väärtused tavalise logistilise regressiooni puhul järgmiste

deletsioon-diagnoos paaride korral: BCR – Z88 ($\beta = 1{,}30,\ SE = 0{,}421,\ p = 0{,}0021$), BCR – ülitundlikkus ($\beta = 1{,}28,\ SE = 0{,}421,\ p = 0{,}0024$). Duplikatsioon-diagnoos paarid vähimate p-väärtustega on CYP2A6 – Y-koodid ($\beta = 3{,}15,\ SE = 0{,}875,\ p = 0{,}00032$) ja SCN5A – M81.4 ($\beta = 2{,}55,\ SE = 0{,}761,\ p = 0{,}00082$).

Mõne CNV-diagnoos paari korral esines TÜ EGV ja UKB andmete analüüsimisel eraldavus. Sel juhul ei andnud tavaline logistiline regressioon parameetritele korrektseid hinnanguid. Firth'i meetodiga olulisi tulemusi küll ei leitud, kuid Firth'i meetod andis lõplikke parameetrite hinnanguid ka eraldavuse korral.

6 Arutelu

Töö käigus leiti oluline seos farmakogeeniga HLA-A ülekattes olevate duplikatsioonide ja diagnoosi L27.0 vahel. Varem on väga mitmeid erinevaid HLA-piirkondi (kuhu kuulub ka HLA-A geen) seostatud ravimist tingitud ülitundlikkusreaktsioonide tekkimisega (Sousa-Pinto *et al*., 2016). Täpne mehhanism, kuidas HLA-A mõjutab ravimist ülitundlikkuse tekkimist, ei ole kindlalt teada, kuid ilmselt võivad muutused HLA-A geenis põhjustada ravimiaine esitlemist T-rakkudele, mis osalevad organismile immunoloogilise kaitse andmises. Nende muutuste tulemusena tekib ülitundlikkusreaktsioon (Sousa-Pinto *et al*., 2016). HLA-A duplikatsioone on uuritud vaid väga vähesel määral ning käesolevas töös tuvastatud seos on huvitavaks aluseks edasisteks uuringuteks.

Farmakogeenidega ülekattes olevate CNV-de ja ravimi kõrvaltoimete diagnooside vaheliste seoste tuvastamine on keeruline. Tuleb arvestada, et CNV-de ja diagnooside andmed ei ole täiuslikud.

Kui doonoril on mingi ravimi kõrvaltoime, siis on teada, et ta on võtnud mingit ravimit, mis seda kõrvaltoimet põhjustab. Kui aga doonoril ravimi kõrvaltoimet ei ole, siis võib see olla lihtsalt seetõttu, et ta pole võtnud ühtegi ravimit, mis vastavat kõrvaltoimet tekitada võiks. See tähendab, et kontrollide seas on tegelikult palju n-ö varjatud juhte, ning see asjaolu vähendab oluliselt CNV-de ja kõrvaltoimete vaheliste seoste leidmise võimsust. Seda probleemi saab leevendada nende ravimi kõrvaltoimete korral, mille kohta on täpsemalt teada, millised ravimid neid põhjustada võivad. Sel juhul saaks analüüsi kaasata vaid need doonorid, kellele on vastavat kõrvaltoimet põhjustavat ravimit välja kirjutatud.

Geenidega ülekattes olevad CNV-d asuvad geenide suhtes erinevates piirkondades – geeni alguses, keskel või lõpus. Samuti on CNV-d erineva pikkusega ja ka geeniga ülekattes oleva osa suurus on CNV-del väga erinev. Selles töös aga eeldati, et kõigi konkreetse farmakogeeniga ülekattes olevate CNV-de mõju ravimi kõrvaltoimele on samasugune ja CNV-de pikkust ega ülekatte suurust pole arvesse võetud. Edaspidi tuleks täpsemate tulemuste saamiseks arvestada ka CNV ulatuse ja ülekatte suurusega.

Tulevikus võiks koos deletsioonidega uurida ka funktsioonikaoga SNV-de mõju ravimi kõrvaltoimete tekkimisele. Funktsioonikaoga mutatsioonide puhul nõrgeneb või kaob geeni avaldumine või geeni produkti funktsioon ja samasugust mõju võivad avaldada ka deletsioonid, seega võib nende mõju koos uurimine anda suuremat võimsust.

Kokkuvõte

Geneetilised variandid farmakogeenides on üks põhjustest, miks erinevatele inimestele mõjuvad ühed ja samad ravimid erinevalt. Bakalaureusetöö põhieesmärk oli TÜ EGV ja UKB andmete põhjal kindlaks teha, kas ja kuidas farmakogeenidega ülekattes olevad koopiaarvu variatsioonid mõjutavad ravimi kõrvaltoime tekkimist.

Selleks pandi esmalt kirja teoreetiline raamistik CNV-de hindamiseks, et selle abil kontrollida, kas kvaliteediskoor on parem kui PennCNV tunnus, kas ja millal toimib Firth'i meetod paremini kui tavaline logistiline regressioon, ning määrata kindlaks testimisse kaasatavate CNV-de ja diagnooside piirsagedused.

Töös näidati, kuidas erinevaid CNV tunnuseid – PennCNV tunnust ja kvaliteediskoori – simuleerida ja selle abil testide võimsust hinnata. Simulatsioonide abil selgitati välja, et kvaliteediskoor on parem kui PennCNV tunnus. Samuti leiti, et Firth'i meetod on väikeste CNV ja diagnoosi osakaalude korral võimsam kui tavaline logistiline regressioon.

CNV-diagnoos paaride vahel esinevate seoste olemasolu testiti logistilise regressiooni, Firth'i meetodi ja logistilise segamudeli abil, kasutades CNV tunnusena kvaliteediskoori ja arvestades simulatsioonide põhjal määratud sageduspiiridega.

Analüüside tulemusena leiti seos farmakogeeniga HLA-A ülekattes olevate duplikatsioonide ja kõrvaltoime L27.0 vahel. Kuna CNV-de seoseid ravimi kõrvaltoimetega on vähe uuritud, siis võiks leitud seos olla aluseks edasiseks teadustööks.

Edaspidi võiks CNV-de ja ravimi kõrvaltoimete vaheliste seoste uurimisel kasutada lisainfot ravimite tarbimise kohta, võtta arvesse ka CNV-de pikkusi ning farmakogeenidega ülekatete ulatusi. Lisaks võiks koos deletsioonidega uurida ka funktsioonikaoga SNV-sid. Reaalsete seoste avastamine CNV-de ja kõrvaltoimete vahel ning leitud seoste ja patsiendi genotüübi kohta käiva info arstidele kättesaadavaks tegemine aitaks arendada personaalset meditsiini.

Viidatud kirjandus

Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H.Š., Trevanion, S. J., Hall, P., Junkins, H., ... Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. doi: 10.1093/nar/gky1120

DeBoever, C., Li, H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K. M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M., Jepsen, K., Matsui, H., Arias, A., Ren, B., Nariai, N., Smith, E. N., D'Antonio-Chronowska, A., Farley, E. K., ja Frazer, K. A. (2017). Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell*, *20*(4), 533–546.e7. doi: 10.1016/J.STEM.2017.03.009

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. doi: 10.1093/biomet/80.1.27

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., ja Ira, G. (2009). Mechanisms of change in gene copy number. *Nature reviews. Genetics*, *10*(8), 551–564. doi: 10.1038/nrg2593

Heinaru, A. (2012). *Geneetika. Õpik kõrgkoolile*. Tartu: Tartu Ülikooli Kirjastus.

Heinze, G., ja Schemper, M. (2002). A solution to the problem of separation in logistic regression. *STATISTICS IN MEDICINE Statist. Med*, *21*, 2409–2419. doi: 10.1002/sim.1047

Kaart, T., ja Möls, T. (2010). *Populatsioonigeneetika genotüüpide tasemel. Loengukonspekt.* Kättesaadav: http://www.eau.ee/~ktanel/MTMS_02_007/loeng_01_2010web.pdf

Korbel, J.Õ., Kim, P. M., Chen, X., Urban, A. E., Weissman, S., Snyder, M., ja Gerstein, M. B. (2008). The current excitement about copy-number variation: how it relates to ge-

ne duplication and protein families. *Current opinion in structural biology*, *18*(3), 366. doi: 10.1016/J.SBI.2008.02.005

Küng, T. A., ja Bogovski, T. P. (1996). *Rahvusvaheline haiguste ja nendega seotud tervi-seprobleemide statistiline klassifikatsioon.* Kättesaadav: http://pub.e-tervis.ee/classifications/RHK-10/6

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, *86*(4), 929–935. doi: 10.1093/biomet/86.4.929

Lepamets, M., Lepik, K., Jürgenson, T., Kals, M., Carmeli, C., Claringbould, A., Bochud, M., Stringhini, S., Wijmenga, C., Franke, L., Mägi, R., ja Kutalik, Z. (2019). New CNV quality score enables discovering novel phenotype associations from genome-wide CNV analysis. *Ilmumas*.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., ... Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS biology*, *5*(10), e254–e254. doi: 10.1371/journal.pbio.0050254

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., ja Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835. doi: 10.1038/nmeth.1681

Ma, C., Blackwell, T., Boehnke, M., ja Scott, L. J. (2013). Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. *Genetic Epidemiology*, *37*(6), 539–550. doi: 10.1002/gepi.21742

Marchini, J., Cardon, L. R., Phillips, M.Š., ja Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, *36*(5), 512–517. doi: 10.1038/ng1337

Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskáčková, T., Balaščák, I., Peltonen, L., Jakkula, E., Rehnström, K., Lathrop, M., Heath, S., Galan, P., Schreiber, S., Meitinger, T., Pfeufer, A., Wichmann, H.-E., ... Metspalu, A. (2009). Genetic Structure of Europeans: A View from the North–East. *PLoS ONE*, *4*(5), e5472. doi: 10.1371/journal.pone.0005472

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., ja Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. doi: 10.1086/519795

Reisberg, S., Krebs, K., Lepamets, M., Kals, M., Mägi, R., Metsalu, K., Lauschke, V. M., Vilo, J., ja Milani, L. (2018). Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genetics in Medicine*, 1–10. doi: 10.1038/s41436-018-0337-5

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., ... Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, *316*(5823), 445–9. doi: 10.1126/science.1138659

Sousa-Pinto, B., Correia, C., Gomes, L., Gil-Mata, S., Araújo, L., Correia, O., ja Delgado, L. (2016). HLA and Delayed Drug-Induced Hypersensitivity. *International archives of allergy and immunology*, *170*(3), 163–79. doi: 10.1159/000448217

Stone, J. L., O'Donovan, M. C., Gurling, H., Kirov, G. K., Blackwood, D. H. R., Corvin, A., Craddock, N. J., Gill, M., Hultman, C. M., Lichtenstein, P., McQuillin, A., Pato, C. N., Ruderfer, D. M., Owen, M. J., St Clair, D., Sullivan, P. F., Sklar, P., Purcell (Leader), S. M., Stone, J. L., ... Sklar, P. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, *455*(7210), 237–241. doi: 10.1038/nature07239

Swaminathan, S., Shen, L., Kim, S., Inlow, M., West, J. D., Faber, K. M., Foroud, T., Mayeux, R., Saykin, A. J., Alzheimer's Disease Neuroimaging Initiative, ja NIA-LOAD/NCRAD Family Study Group. (2012). Analysis of copy number variation in Alzheimer's disease: the NIALOAD/ NCRAD Family Study. *Current Alzheimer research*, *9*(7), 801–14. Kättesaadav: http://www.ncbi.nlm.nih.gov/pubmed/22486522

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., ja Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17*(11), 1665–1674. doi: 10.1101/gr.6861907

Wilkinson, G. R. (2005). Drug Metabolism and Variability among Patients in Drug Response. *New England Journal of Medicine*, *352*(21), 2211–2221. doi: 10.1056/NEJMra032424

Zarrei, M., MacDonald, J. R., Merico, D., ja Scherer, S. W. (2015, mar). A copy number variation map of the human genome. *Nature Reviews Genetics*, *16*(3), 172–183. doi: 10.1038/nrg3871

Zhang, F., Gu, W., Hurles, M. E., ja Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, *10*(1), 451–481. doi: 10.1146/annurev.genom.9.081307.164217

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

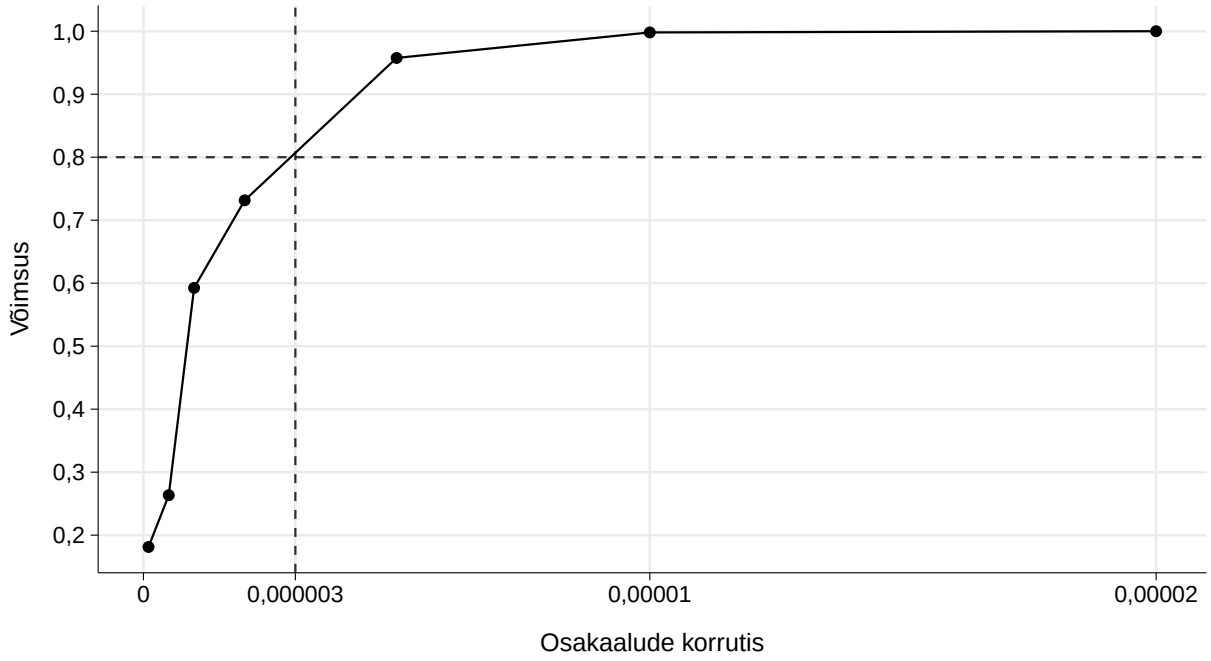
Lisa 1. Binaarse CNV tunnuse simuleerimine

```
library(logistf)
library(dplyr)

# N - valimimaht
# p - diagnoosi osakaal, p0 - kontrollide CNV osakaal,
# OR - tegelik šansside suhe
# p01 - valepositiivsuse määr, p10 - valenegatiivsuse määr

simulatsioon1 <- function(N, p, p0, OR, p01, p10){
    n1 = N * p # juhtude arv
    n0 = N - n1 # kontrollide arv
    p1 = p0 * OR / (p0 * (OR - 1) + 1) # juhtude CNV osakaal

    p11 = 1 - p10 #õigepositiivsuse määr
    p1_penn = (p1*p11+(1-p1)*p01) # juhtude CNV osakaal PennCNV tunnuse kaudu
    p0_penn = (p0*p11+(1-p0)*p01) # kontrollide CNV osakaal PennCNV tunnuse kaudu

    T0 = seq(0,n0) # CNV kandjate võimalikud arvud kontrollide hulgas
    T1 = seq(0,n1) # CNV kandjate võimalikud arvud juhtude hulgas

    pT0 = dbinom(T0, n0, p0_penn) # tõenäosused P(T_0=i)
    names(pT0) = T0
    pT1 = dbinom(T1, n1, p1_penn) # tõenäosused P(T_1=j)
    names(pT1) = T1

    m = outer(pT0, pT1) # paaride (i,j) tõenäosused P(T_0=i,T_1=j)

    # jätame alles need paarid  (i,j), mille tõenäosus on piirist suurem
    piir = 1e-10
    v = which(m > piir, arr.ind = T)
    paarid = data.frame(i = as.numeric(rownames(m)[v[, 1]]),
                        j = as.numeric(colnames(m)[v[, 2]]))

    tulemused = NULL
    for (k in 1:nrow(paarid)){
        i = paarid[k,1] # kontrollid
        j = paarid[k,2] # juhud

        x = c(c(rep(1, i), rep(0, n0-i), c(rep(1, j), rep(0, n1-j)))) # CNV tunnus
        y = c(rep(0, n0), rep(1, n1)) #diagnoosi tunnus

        m1 = glm(y~x, family = binomial(link = "logit"))
        m2 = logistf(y~x, pl=F, firth=F)

        tulemused = rbind(tulemused,data.frame(glm.pvalue=last(coef(summary(m1))),
                                               glm.OR = last(exp(coef(m1))),
                                               firth.pvalue = last(m2$prob),
                                               firth.OR = last(exp(coef(m2))),
                                               p_ij = m[as.character(i),as.character(j)],
                                               p=p, p0=p0, OR=OR, p01=p01, p10=p10))
    }
    return(tulemused)
}
```

Lisa 2. CNV kvaliteediskoori simuleerimine

```
library(logistf)
library(dplyr)

# valepositiivsete jaotus
S0 = function(n=1){
    u = runif(n,0,1)
    s1 = c(runif(sum((u<=0.02*25.5)),0,0.02),runif(sum(u>0.02*25.5),0.02,1))}

# õigepositiivsete jaotus
S1 = function(n=1){
    u = runif(n,0,1)
    s0 = c(runif(sum(u<=0.98*0.8),0,0.98),runif(sum(u>0.98*0.8),0.98,1))}

# N - valimimaht
# p - diagnoosi osakaal, p0 - kontrollide CNV osakaal,
# OR - tegelik šansside suhe
# p01 - valepositiivsuse määr, p10 - valenegatiivsuse määr

simulatsioon2 <- function(N, p, p0, OR, p01, p10, arv){
    n1 = N * p # juhtude arv
    n0 = N - n1 # kontrollide arv
    p1 = p0 * OR / (p0 * (OR - 1) + 1) # juhtude CNV osakaal
    p11 = 1 - p10 # õigepositiivsuse määr

    tulemused = NULL
    for (i in 1:arv) {
        # kvaliteediskoori simuleerimine
        QS = function(n,cnv_freq){
            u = runif(n,0,1)
            qs = c(S1(sum(u <= cnv_freq*p11)),
                S0(sum((u > cnv_freq*p11) & (u <= (cnv_freq*p11 + (1-cnv_freq)*p01)))),
                rep(0, sum(u>cnv_freq*p11 + (1-cnv_freq)*p01)))}

        x = c(QS(n0, p0), QS(n1, p1)) # CNV tunnus
        y = c(rep(0, n0), rep(1, n1)) # diagnoosi tunnus

        m1 = glm(y~x, family = binomial(link = "logit"))
        m2 = logistf(y~x)

        tulemused = rbind(tulemused, data.frame(glm.pvalue=last(coef(summary(m1))),
                                                glm.OR = last(exp(coef(m1))),
                                                firth.pvalue = last(m2$prob),
                                                firth.OR = last(exp(coef(m2))),
                                                p=p, p0=p0, OR=OR, p01=p01, p10=p10))
  }
  return(tulemused)
}
```

Lisa 3. Võimsuse sõltuvus osakaalude korrutisest



Joonis L1. Logistilise regressiooni võimsus valimimahu N = 400 000, OR=5 ja tegeliku CNV tunnuse korral. Horisontaalne ja vertikaalne kriipsjoon tähistavad vastavalt 80% võimsuse piiri ja osakaalude korrutise piiri 80% võimsuse saavutamiseks.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Tuuli Jürgenson,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Koopiaarvu variatsioonide mõju ravimi kõrvaltoimete tekkimisele“, mille juhendajad on Maarja Lepamets ja Kaido Lepik, 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

08.05.2019