Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Sille Habakukk

Geenimarkerite põhjal hinnatud haiguseriski võrdlus perekonnaajalooga

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendajad Krista Fischer, PhD
Kristi Läll, MSc

Tartu 2016

Infoleht

Geenimarkerite põhjal hinnatud haiguseriski võrdlus perekonnaajalooga

Geneetika kiire areng lubab üha enam kasutada perekonnaajaloo kõrval ka geneetilisi teste inimese haiguseriski hindamiseks. Bakalaureusetöö eesmärk on välja selgitada, millistel tingimustel töötab haiguseriski hindamisel paremini perekonnaajalugu ja millistel tingimustel geneetilised andmed. Töös viiakse läbi kaks simulatsioonuuringut erinevatel simuleerimistingimustel ja saadud mudeleid võrreldakse nii andmetega sobivuse kui ka diagnostilise võimekuse poolest. Geneetilisi andmeid kasutavad mudelid edestavad perekonnaajaloopõhiseid mudeleid pea igas olukorras.

Märksõnad: ühenukleotiidsed polümorfismid, geneetiline riskiskoor, perekonnaajalugu, simulatsioon

Teadusala CERCS nimetus ja kood: Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika (P160)

Estimation of disease risk using genetic markers or family history: a simulation study

Fast development in genetics allows to use genetic tests instead of family history for predicting disease risk. The purpose of this thesis is to determine whether the use of genetic markers or family history is more accurate in predicting disease risk. A series of simulation studies are carried out under different scenarios. The resulting models are compared in terms of fit of data and diagnostic ability. The results indicate that genetic data outperforms family history in almost all cases.

Keywords: single nucleotide polymorphism, polygenic risk score, family history, simulation

CERCS classification and code: Statistics, operations research, programming, actuarial mathematics (P160)

Sisukord

Sissejuhatus

Geneetiliste testide areng viimase kümnendi jooksul on olnud kiire. Seetõttu on langenud nii testide tegemise maksumus kui ka ajakulu ning geenitestide kasutamine on üha kättesaadavam personaalse meditsiini seisukohast. Siiski eelistatakse indiviidide haigestumisriskide hindamisel standardpraktikaks kujunenud perekonnaajaloo kasutamist, mille kogumine on kiire ja praktiliselt tasuta. Perekonnaajaloo kasutamisel esineb aga mitmeid puuduseid - info sugulaste haiguste kohta võib olla ebatäpne või üldse puududa. Seetõttu küsitlemisel saadav informatsioon ei pruugi anda piisavalt infot isiku tegeliku haiguseriski kohta.

Bakalaureusetöö eesmärgiks on simulatsioonuuringute põhjal välja selgitada, millistel tingimustel ja kui palju erinevad perekonnaajaloo ja geenimarkerite põhjal hinnatud haiguseriskid. Simulatsioonid ja analüüsid viiakse läbi tarkvaraga R.

Töö on liigendatud kolmeks peatükiks. Esimeses peatükis on toodud kokkuvõte kirjandusest. Täpsemalt on käsitletud geneetikaga seotud mõisteid, päritavust, Hardy-Weinbergi tasakaalu, logistilist regressioonimudelit, Akaike informatsioonikriteeriumit ja ROC kõvera alust pindala. Teises peatükis vaadeldakse olukorda, kus haigestumist mõjutab üks geenimarker. Erinevusi perekonnaajaloo ja geenimarkerite vahel on kirjeldatud nii teoreetilise arutluskäigu kui ka simulatsiooni analüüsiga. Kolmandas peatükis on simuleeritud olukorda, kus haigestumist mõjutab sada geenimarkerit. Töö sisaldab kolme lisa. Lisas 1 on toodud enim kasutatavate mõistete sõnastik, lisas 2 on viide töös kasutatud koodidele ligipääsuks ja lisas 3 on tabel varasemas uuringus saadud tulemuste kohta.

Autor tänab juhendajaid Krista Fischerit ja Kristi Lälli rohkete selgituste ja nõuannete eest.

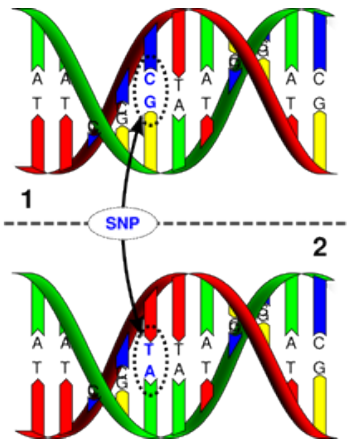
1 Kirjanduse ülevaade

1.1 Alusmõisted geneetikas

Desoksüribonukleiinhape ehk DNA on päriliku informatsiooni edasikandja inimestel. DNA koostises esineb nelja erinevat nukleotiidi: adenosiinfosfaat (tähistatakse A), tümidiinfosfaat (T), tsütidiinfosfaat (C) ja guanosiinfosfaat (G). Nende nukleotiidide liitumisel tekib DNA ahel. Komplementaarsusprintsiibi põhjal ühinevad kaks DNA ahelat DNA molekuliks ehk kui ühes ahelas on A, siis teises ahelas on vastas T, sarnaselt esinevad koos C ja G. (Kasela, 2011)

Täielikku DNA järjestust kutsutakse genoomiks. Inimese genoomis on hinnanguliselt $3.3 \cdot 10^9$ nukleotiidipaari, kusjuures kahe inimese genoom erineb kuni 0.1% ulatuses, see on 3.3 miljonit nukleotiidipaari. Genoomist umbes 3% moodustavad kodeerivad järjestused ehk 30 kuni 40 tuhat geeni (Burton, Tobin ja Hopper, 2005). Geen on DNA segment, mis kodeerib ühte valgu molekuli (Kasela, 2011).

Üks geneetilise varieeruvuse näitaja on SNP (häädatakse „snipp") ehk üksiku nukleotiidi muutusest põhjustatud polümorfism. Polümorfism tähendab kahe või enama geneetilise variandi olemasolu populatsiooni isendite hulgas, sealjuures vähemalt kahe variandi sagedus on suurem kui 1% (Heinaru, 2012, lk 1051).



Joonis 1.1: Näide SNPst. (Dnabaser.com, 2016)

Joonisel 1.1 on kujutatud olukorda, kus ühe inimese DNA ahel (C**C**TAG) erineb teise inimese DNA ahelast (C**T**TAG) ühe nukleotiidi võrra ehk märgitud kohal esineb SNP.

SNPide harvemini esineva variandi ehk minoorse alleeli sagedust kindlas populatsioonis iseloomustab suurus MAF (ingl. *Minor Allele Frequency*). Enim esineb bialleelseid SNPe, vaid 0.1% on trialleelsed. SNPd asuvad genoomis nii kodeerivates kui ka mittekodeerivates alades, esimesel juhul võib SNP kaasa tuua muutuse sünteesitavas valgus. Siiski, SNPd enamasti ei põhjusta haiguseid, vaid neid kasutatakse haigustega seotud geenide identifitseerimiseks. Seeläbi on SNPe võimalik kasutada haiguse tekke ennustamiseks. (Kasela, 2011)

1.2 Päritavus

Geneetika mõju haigestumisele on põhjust uurida vaid siis, kui on alust arvata, et haigus on pärilik. Eeldatakse, et indiviidi fenotüübiline tunnus P, milleks võib olla näiteks vererõhk, avaldub kujul

$$P = G + E, \tag{1.1}$$

kus G on fenotüübi geneetiliselt määratud osa ja E on keskkonna poolt määratud osa. (Kaart, 2012)

Näitena võib tuua ühemunaraku kaksikud, kelle genotüübid on identsed, seega nende vererõhku mõjutav geneetiline komponent G on ühesugune ja vererõhk saab neil erineva vaid keskkonnamõjude E tõttu.

Fenotüübilise tunnuse P kui kahe juhusliku suuruse summa dispersioon avaldub järgnevalt

$$var(P) = var(G) + var(E) + cov(G, E). \tag{1.2}$$

Kui $cov(G, E) = 0$ ehk eeldades genotüübi ja keskkonna sõltumatust, defineeritakse päritavus (ingl. *heritability*) ehk päritavuskoefitsient:

$$h^2 = \frac{var(G)}{var(P)}.$$

Päritavus näitab, kui suure osa tunnuse fenotüübi varieeruvusest moodustab geneetiline varieeruvus.

Valemist 1.2 ning genotüübi ja keskkonna sõltumatusest tuleneb, et

$$0 \leqslant var(G) \leqslant var(P),$$

seega päritavuskoefitsient jääb lõiku $[0, 1]$. Päritavus 0 tähendab, et uuritav fenotüüp ei ole seotud genotüübiga. Vastupidiselt tähendab päritavus 1, et kogu tunnuse varieeruvus on seletatav päriliku muutlikkuse kaudu. (Kaart, 2011)

Haiguseid, mille päritavus on 1 ehk haigestumisel puuduvad keskkonnamõjud ($E = 0$), nimetatakse mendeliaalseteks haigusteks. Selliste haiguste puhul piisab vastava geenitesti tegemisest, et üheselt määrata haigusstaatus ja puudub vajadus statistilise analüüsi järele. Seetõttu käsitletakse järgnevalt töö käigus niinimetatud kompleks-haiguseid, kus $E \neq 0$.

Et eristada haiguseid, mida põhjustab perekonna ühine keskkond, haigustest, mida põhjustavad perekonna jagatud geenid, kasutatakse kaksikute uuringuid. Ühemunarakukaksikud (UK) on genotüübi poolest täiesti identsed, samas kui kahemunarakukaksikute (KK) genotüübid on umbes 50% ulatuses samad. Meetodi idee selgub vaadeldes näiteks jällegi vererõhku: kui varieeruvus vererõhutasemes oleks täielikult geneetilise varieeruvuse põhjal seletatav, siis oleks ühemunarakukaksikutel täpselt sama vererõhk ehk korrelatsioon vererõhutaseme vahel ühemunarakukaksikutel oleks 1. Et vererõhutaset mõjutavad ka keskkonnatingimused, siis avaldatakse päritavus:

$$h^2 = 2(r_{KK} - r_{UK}),$$

kus r_{UK} on tunnuse korrelatsioon identsete kaksikute puhul ja r_{KK} tunnuse korrelatsioon kahemunarakukaksikute puhul. (Wassertheil-Smoller, 2004. lk 173-175)

1.3 Hardy-Weinbergi tasakaal

Vaatleme suurt populatsiooni, kus puuduvad migratsioon, valik ja mutatsioonid, ning ristumine indiviidide vahel on täiesti juhuslik. Hardy-Weinbergi tasakaal on

printsiip, mis väidab, et sellises populatsioonis püsivad alleeli- ja genotüübisagedused põlvkonniti muutumatutena. Öeldakse ka, et taoline populatsioon on geneetilise tasakaalu seisundis. See tähendab, et kui on teada, et geenil on kaks alleeli, A ja a, vastavalt sagedustega $P(A) = p$ ja $P(a) = 1 - p$, siis populatsiooni genotüübisagedused avalduvad alleelisageduste kaudu: $P(AA) = p^2$, $P(Aa) = 2p(1 - p)$ ja $P(aa) = (1 - p)^2$. (Kasela, 2011)

Kuigi sellist ideaalpopulatsiooni reaalselt ei eksisteeri, siis võib lühikeste, mõne põlvkonna pikkuste ajavahemike korral populatsiooni paljude geenide suhtes vaadelda praktiliselt tasakaalulisena. (Kasela, 2011)

1.4 Logistiline regressioonimudel

Järgnev alapeatükk põhineb Ene Kääriku 2015. loengukursuse Andmeanalüüs II peatükil 9.2 „Logistiline regressioon".

Vaatleme tunnust Y, mis iseloomustab teatud omaduse, näiteks mõne haiguse, puudumist ($Y = 0$) või esinemist ($Y = 1$). Selline tunnus on Bernoulli jaotusega $Y \sim B(1, \pi)$, kus π on omaduse esinemise tõenäosus. Et Bernoulli jaotuse korral $EY = \pi$, hindab mudel keskväärtusele ka omaduse esinemise tõenäosust. Lisaks on huvi hinnata seos omaduse esinemise tõenäosuse ja mõõdetud seletavate tunnuste vahel.

Binaarse uuritava tunnuse mudeldamisel kasutatakse *Logit* teisendust ehk naturaallogaritmi omaduse esinemise šansist

$$Logit(\pi) = \ln \frac{\pi}{1 - \pi}$$

Logistiline regressioonimudel eeldab, et $Logit(\pi)$ avaldub seletavate tunnuste $X_1, \ldots, X_k$ väärtuste $x_1, \ldots, x_k$ lineaarkombinatsioonina:

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

Logistilise regressiooni parameetreid hinnatakse suurima tõepära meetodil.

Väärtuse $Y = 1$ esinemise tõenäosus avaldub mudelist:

$$\pi = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}. \tag{1.3}$$

1.5 Akaike informatsioonikriteerium

Lühikokkuvõte Akaike informatsioonikriteeriumist on kirjutatud Kenneth P. Burnhami ja David R. Andersoni 2002. aasta raamatu „Model selection and Multimodel Inference" teise peatükki põhjal.

Hindamaks suhtelist kaugust uuritava mudeli ja tegeliku (varjatud) andmete tekkemehhanismi vahel, defineeris Akaike 1973. aastal Akaike informatsiooni kriteeriumi ehk *AIC* (ingl. *Akaike information criterion*):

$$AIC = -2\ln(L) + 2k,$$

kus L on mudeli tõepärafunktsiooni maksimaalne väärtus ja k on mudeli parameetrite arv.

AIC omab sisukat seletust ainult sama andmestiku pealt hinnatud mudelite hulgas. Seejuures ei oma *AIC* absoluutne väärtus tähendust, olulised on vaid mudelite *AIC* väärtuste vahed:

$$\Delta_i = AIC_i - AIC_{min},$$

kus AIC_{min} on uuritavate mudelite hulgast väikseim *AIC* väärtus ja indeks i tähistab i-nda mudeli vastavat *AIC* väärtust. Rusikareegel Δ_i interpreteerimiseks on järgmine: kui $\Delta_i > 10$ siis mudel i jätab seletamata olulise osa andmete varieeruvusest ja selle võib vaatluse alt välja jätta.

Samas tuvastab AIC parima mudeli ainult uuritud mudelite seast, mistõttu võib väljavalitud mudel ikkagi osutuda reaalsusest kaugeks. Seetõttu on oluline, et mudelid oleks uuritavate mudelite hulka valitud läbimõeldult.

1.6 ROC kõvera alune pindala ehk AUC

Ülevaade ROC kõverast ja AUCst on kokku pandud Sylvia Wassertheil-Smolleri 2004. aastal välja antud õpiku „Biostatistics and Epidemiology" viiendal peatükil ja Tom Fawcetti 2006. aasta artikli „An introduction to ROC analysis" põhjal.

Haiguse *levimus*eks nimetataks suurust:

$$levimus = \frac{\text{Haigete inimeste arv populatsioonis}}{\text{Inimeste koguarv populatsioonis}}.$$

Tabel 1.1: Tähistused erinevatele võimalikele testitulemustele

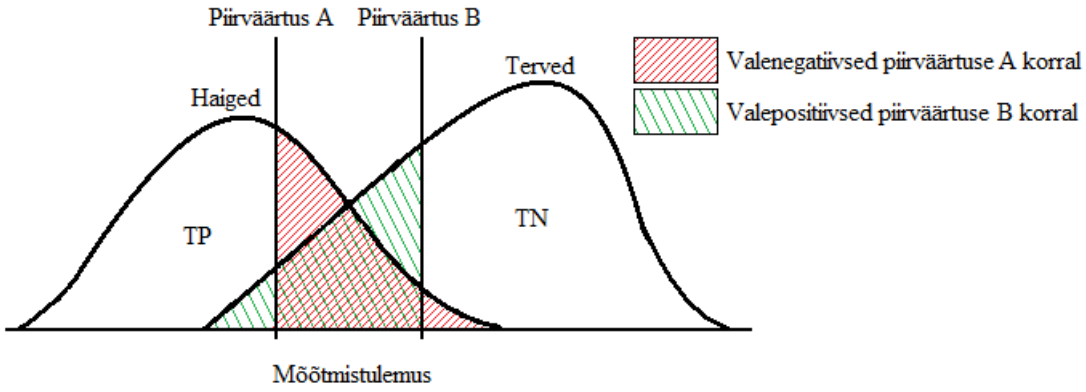
		Haige: Jah	Haige: Ei
Testi tulemus	+	Õige positiivne (**TP**)	Valepositiivne (**VP**)
	−	Valenegatiivne (**VN**)	Õige negatiivne (**TN**)

Diagnostiliste testide tegemisel on oluline tuvastada tegelikult haiged inimesed ja samal ajal vältida haiguse diagnoosimist tegelikult tervetel inimestel. Tabelis 1.1 on toodud tähistused võimalikele testi tulemustele vastavalt tegelikule haigusstaatusele. Testide iseloomustamiseks kasutatakse suuruseid *tundlikkus* ja *spetsiifilisus*.

$$tundlikkus = \frac{TP}{TP + VN}, \qquad spetsiifilisus = \frac{TN}{TN + VP}$$

Diagnostilise testi tundlikkus ja spetsiifilisus sõltuvad testi piirväärtusest, mis määrab ära, kes testitutest diagnoositakse haigeks ja kes terveks. Joonisel 1.2 on toodud kaks piirväärtust A ja B. Piirväärtuse A puhul on testi tundlikkus madalam kui piirväärtusel B, sest valenegatiivsete osakaal on piirväärtuse A korral kõrgem. Samas piirväärtusele A vastab kõrgem spetsiifilisus, kuna valepositiivsete osakaal on madalam kui piirväärtusele B vastaval testil.

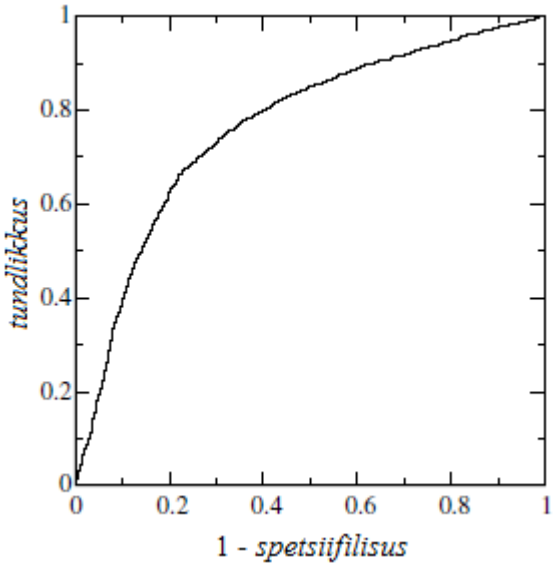
Tundlikkus ja spetsiifilisus on omavahel monotoonselt seotud. Kui muuta testi piirväärtust tundlikkuse tõstmiseks, siis võib suureneda valepositiivsete arv ja spet-



Joonis 1.2: Piirväärtusest A vasakule poole jääv roheliselt viirutatud ala tähistab valepositiivseid tulemusi piirväärtuse A korral. Piirväärtusest B paremale jääv punaselt viirutatud ala tähistab valenegatiivsete tulemuste arvu piirväärtuse B korral.

siifilisus võib langeda, ja vastupidiselt, kui muuta testi piirväärtust spetsiifilisuse tõstmiseks, siis tundlikkus võib langeda.

Illustreerimaks seost tundlikkuse ja spetsiifilisuse vahel erinevate võimalike piirväärtuste korral esitatakse tihti ROC (ingl. *Reciever Operating Characteristic*) kõver. Sellel kahemõõtmelisel graafikul on x-teljel kujutatud valepositiivse tulemuse tõenäosus ehk $1 - spetsiifilisus$ ja y-teljel *tundlikkus*, nagu on ka näidatud joonisel 1.3.



Joonis 1.3: Näide ROC kõverast (Fawcett, 2006, lk 870)

ROC kõverat saab kasutada erinevate testide võrdlemiseks. Selle jaoks leitakse igale

testile ROC kõvera alune pindala ehk AUC (ingl. *Area Under the ROC Curve*). Kuna AUC on osa ühikruudu pindalast, siis tema väärtused jäävad vahemikku $[0, 1]$. Kui test töötaks sama hästi kui aus mündivise, siis oleks AUC väärtus 0.5. AUC väärtust saab tõlgendada kui tõenäosust, et test määrab juhuslikult valitud, tegelikult haige, inimese haigeks suurema tõenäosusega kui juhuslikult valitud, tegelikult terve, inimese haigeks.

2 Ühe geenimarkeriga mudel

Vaadeldakse kõige lihtsamat juhtu, kus haiguse geneetiline komponent G on määratud vaid ühe SNP poolt ja võrreldakse tõenäosusi haigestuda lihtsaima perekonnaajaloo ehk ainult vanemate haigusstaatuse põhjal leitud haigestumise tõenäosustega.

2.1 Perekonnaajaloo kasutamise teoreetiline tuletuskäik

Järgnevalt avaldatakse vanemate haigusstaatuse põhjal lapse haigestumise tõenäosuse. Tuletuskäigu selguse mõttes kasutatakse järgmisi tähistusi:

- H_V, kus $V \in \{E, I\}$ ja tähistab ühe vanema (ema või isa) haigusstaatust.
- g_V, kus $V \in \{E, I\}$ ja tähistab ühe vanema(ema või isa) genotüübi väärtust. $g_V \in \{0, 1, 2\}$ vastavalt sellele, kas isikul on null, üks või kaks riskialleeli.
- f_0, f_1, f_2 tähistavad vastavalt nulli, ühe ja kahe riskialleeliga genotüübi suhtelisi sagedusi.

Tuletuskäigus lähtutakse eeldustest:

- Genotüüpide sagedused populatsioonis f_0, f_1, f_2 on teada.
- Tõenäosused haigestuda on vastavalt genotüübile p_0, p_1, p_2.
- Vanemate haiguse staatused $H_E = 0/1$, $H_I = 0/1$ on teada.

Tõenäosus, et kindel genotüüp ja haigus esinevad vanemal sama-aegselt avaldub kui

$$P(g_V = k, H_V = 1) = P(H_V = 1 \mid g_V = k)P(g_V = k) = p_k f_k \tag{2.1}$$

Kuna vanemal on üks ja ainus genotüüp, siis moodustab vanema genotüüp täissüsteemi ja vanema haigestumise tõenäosuse saab avaldada

$$P(H_V = 1) = \sum_{k=0}^{2} P(H_V = 1 \mid g_V = k)P(g_V = k) = \sum_{k=0}^{2} p_k f_k \tag{2.2}$$

Tabel 2.1: Lapse genotüübi võimalikud väärtused vanemate genotüüpide põhjal

g_E \ g_I	0	1	2
0	0	0/1	1
1	0/1	0/1/1/2	1/2
2	1	1/2	2

Vanema haigusseisundi põhjal ja kasutades valemeid 2.1 ja 2.2 saab ennustada vanema genotüüpi

$$P(g_V = k|H_V = 1) = \frac{P(g_V = k, H_V = 1)}{P(H_V = 1)} = \frac{p_k f_k}{\sum_{k=0}^{2} p_k f_k} := \alpha_k^1 \tag{2.3}$$

Sarnaselt saab leida ka

$$P(g_V = k|H_V = 0) = \frac{P(g_V = k, H_V = 0)}{P(H_V = 0)} = \frac{(1 - p_k) f_k}{1 - \sum_{k=0}^{2} p_k f_k} := \alpha_k^0 \tag{2.4}$$

Olgu teise vanema samad tõenäosused tähistatud $\beta_k^{0/1}$.

Lapse genotüüp g_L sõltub vanemate genotüüpidest. Tabelis 2.1 toodud lapse genotüüpide väärtuste esinemise tõenäosused on määratud Mendeli 1. ja 2. seadustega (vt Lisa 1) ja on võimalik kirja panna tõenäosuste 2.3 ja 2.4 kaudu:

$$\begin{aligned} P(g_L = 0 \mid H_E, H_I) &= P(g_E = 0 \mid H_E)P(g_I = 0 \mid H_I) + \\ &+ \frac{1}{2}P(g_E = 0 \mid H_E)P(g_I = 1 \mid H_I) + \\ &+ \frac{1}{2}P(g_E = 1 \mid H_E)P(g_I = 0 \mid H_I) + \\ &+ \frac{1}{4}P(g_E = 1 \mid H_E)P(g_I = 1 \mid H_I) = \\ &= \alpha_0^{H_E}\beta_0^{H_I} + \frac{1}{2}\alpha_1^{H_E}\beta_0^{H_I} + \frac{1}{2}\alpha_0^{H_E}\beta_1^{H_I} + \frac{1}{4}\alpha_1^{H_E}\beta_1^{H_I} \end{aligned}$$

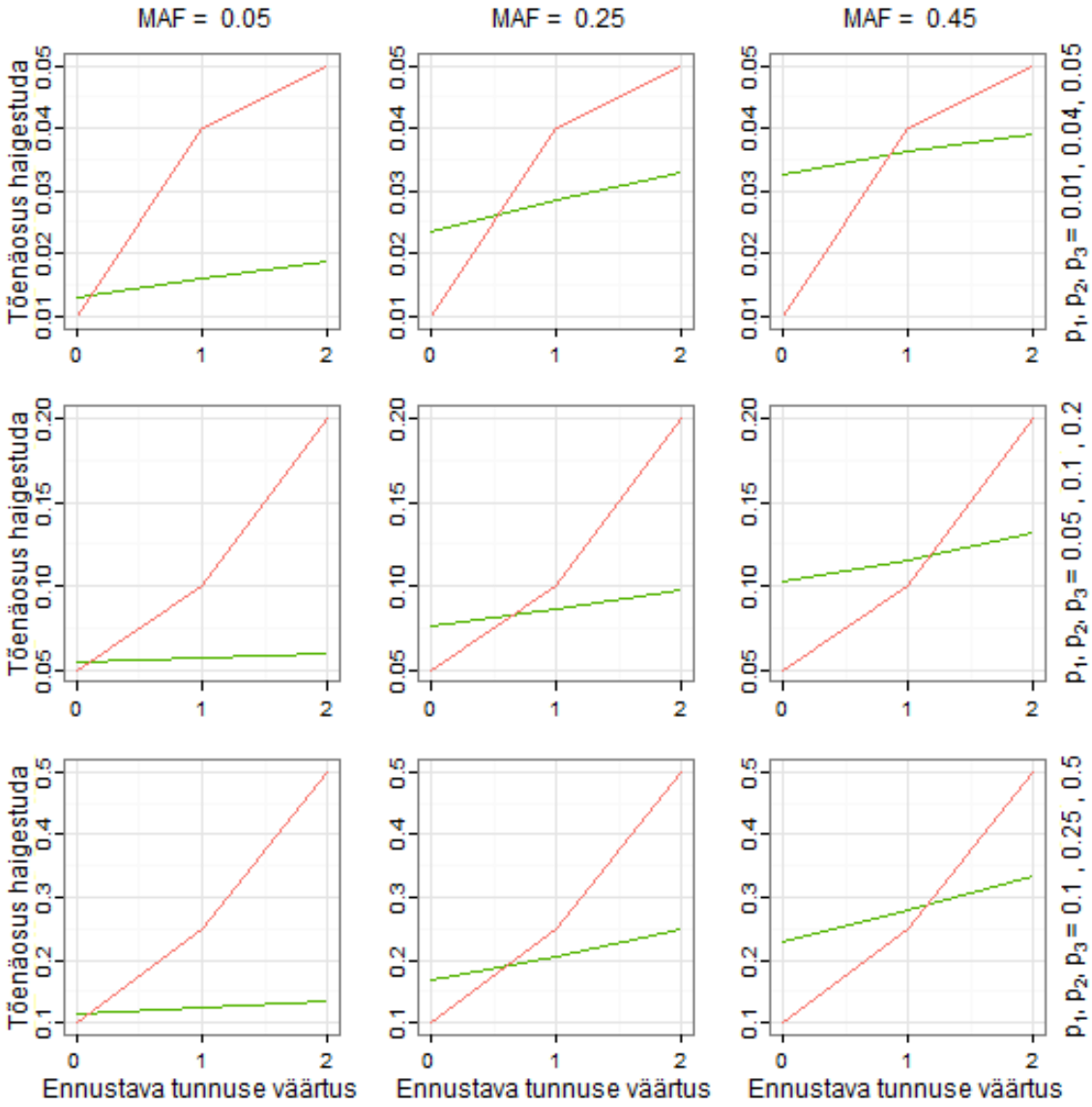
Teised kaks tõenäosust avalduvad sarnaselt:

$$\begin{aligned} P(g_L = 1 \mid H_E, H_I) &= \frac{1}{2}\alpha_0^{H_E}\beta_1^{H_I} + \alpha_0^{H_E}\beta_2^{H_I} + \frac{1}{2}\alpha_1^{H_E}\beta_0^{H_I} + 2 \cdot \frac{1}{4}\alpha_1^{H_E}\beta_1^{H_I} + \\ &+ \frac{1}{2}\alpha_1^{H_E}\beta_2^{H_I} + \alpha_2^{H_E}\beta_0^{H_I} + \frac{1}{2}\alpha_2^{H_E}\beta_1^{H_I} \\ P(g_L = 2 \mid H_E, H_I) &= \frac{1}{4}\alpha_1^{H_E}\beta_1^{H_I} + \frac{1}{2}\alpha_1^{H_E}\beta_2^{H_I} + \frac{1}{2}\alpha_2^{H_E}\beta_1^{H_I} + \alpha_2^{H_E}\beta_2^{H_I} \end{aligned}$$

Arutluskäigu tulemusena saab kirja panna tõenäosuse, et laps on haige, kui on teada tema vanemate haigusseisund:

$$P(H_L = 1 \mid H_E, H_I) = \sum_{k=0}^{2} P(g_L = k \mid H_E, H_I) p_k \tag{2.5}$$

Eelneva tuletuskäigu rakendamiseks on võimalik tuua näide perekonnaajaloo ja genotüübi põhjal hinnatud haigestumistõenäosuse muutumisest erinevate haigestumist mõjutavate parameetrite korral.



Joonis 2.1: Rohelisega on kujutatud haigete vanemate arvu põhjal hinnatud haigestumise tõenäosused ja punasega on ära toodud lapse riskialleelide arvule vastava haigestumise tõenäosused. X-teljel on kujutatud vastavalt kas riskialleelide või haigete vanemate arv.

Joonisel 2.1 on toodud haigete vanemate arvu põhjal hinnatud lapse haigestumise tõenäosused, mis on leitud valemi 2.5 abil. Lapse ehk indeksisiku riskialleelide arvule vastava haigestumise tõenäosused on p_0, p_1, p_2, mis joonise ridades on valitud kui $(p_0, p_1, p_2) = (0.01, 0.04, 0.05), (0.05, 0.1, 0.2)$ ja $(0.1, 0.25, 0.5)$. Tulpades on kasutatud erinevaid MAF väärtuseid - $0.05, 0.25$ ja 0.45.

Kuigi rangelt võttes tähendab MAF lihtsalt harvemini esineva alleeli sagedust, siis käsitletakse järgnevas arutelus väikseima sagedusega alleeli kui riskialleeli.

Näeme, et perekonnaajaloo põhjal ennustatav haigestumise tõenäosus sõltub tugevalt MAF väärtusest: mida suurem on riskialleelide sagedus populatsioonis, seda kõrgem on hinnatud risk haigestuda. Lisaks muutub haigete vanemate arvu kasvades haigestumise tõenäosus palju lineaarsemalt kui genotüübipõhine tõenäosus. Seetõttu perekonnaajaloo järgi haigestumise tõenäosust hinnates väikese MAF väärtusega riskialleeli puhul jäädakse väga konservatiivseks ja riskihinnangud nulli, ühe ja kahe haige vanema puhul erinevad vähe. Vastupidiselt, suurte MAF väärtuste puhul, ülehindab vanematemudel märgatavalt tõenäosust haigestuda riskialleelideta isikul, $MAF = 0.45$ korral isegi üle kahe korra. Seevastu genotüübipõhine haigestumise tõenäosus ei sõltu vaadeldud juhul MAFst.

2.2 Simulatsioon

Riskialleelide ja haigete vanemate arv jagavad kumbki populatsiooni kolmeks rühmaks, millest esimeste rühmade suurus sõltub riskialleeli sagedusest ja teistel haiguse levimusest. Vastavad rühmad ei ole suuruselt võrreldavad ja joonis 2.1 ei anna veel selget vastust kumb prognoos on praktikas parem. Lisaks tuleb arvestada, et praktikas ei ole genotüübile vastava haigestumise tõenäosused teada ja kasutada tuleb hinnanguid varasematest uuringutest.

Seetõttu, saamaks aimdust haigestumise prognoosimise võimalikkusest reaalses olukorras, viiakse läbi simulatsioonuuring. Simuleerides eeldame, et kõige väiksema sagedusega alleel tõstab haigestumise riski, ehk tõlgendame MAF-i kui riskialleeli sagedust populatsioonis. Andmestiku genereerimine toimub järgnevate reeglite põhjal:

1. Määratakse kindlaks soovitavate indeksisikute arv n ja riskialleeli esinemise sageduse populatsioonis MAF.

2. Genereeritakse n vanemate paari genotüübid G_E ja G_I, mis on mõlemad jaotusega $B(2, MAF)$, kuna igal vanemal võib olla kuni kaks riskialleeli.

3. Genereeritakse n lapse ehk indeksisikute genotüübid vastavalt vanematelt saadavatele alleelidele:

   - Kui vanema genotüüp on 0, siis pärandub lapsele 0 riskialleeli.
   - Kui vanema genotüüp on 2, siis pärandub lapsele 1 riskialleel.
   - Kui vanema genotüüp on 1, siis pärandub lapsele riskialleel tõenäosusega 0.5 ehk jaotusega $B(1, 0.5)$.

4. Määratakse kindlaks genotüüpidele vastavad tõenäosused haigestuda p_0, p_1, p_2.

5. Genereeritakse igale indiviidile haigusstaatuse $H \sim B(1, p)$, kus p on indiviidi genotüübile vastav tõenäosus haigestuda. Arvutatakse ka haigete vanemate arv $H_{VV} = H_E + H_I$.

Simulatsioon viidi läbi valimisuurusega $n = 10^6$ indeksisikut, kasutades $MAF = 0.2$ ja kolme erineva seti genotüüpidele vastavate haigestumistõenäosustega: $(p_0, p_1, p_2) = (0.01, 0.04, 0.05)$, $(0.05, 0.1, 0.2)$ ja $(0.1, 0.25, 0.5)$.

Saadud andmestikust 10^5 indeksindiviidi põhjal sobitati logistilised regressioonimudelid $H \sim H_{VV}$ (nn vanemamudel) ja $H \sim G$ (nn geenimudel) iga genotüüpidele vastavate haigestumistõenäosuste seti jaoks. Seejärel ennustati saadud mudelite põhjal ülejäänud andmestiku $9 \cdot 10^5$ indeksindiviidile haigestumise riskid ja määrati binoomjaotusega $B \sim (1, p)$, kus p on mudelile vastav haigestumise risk, indeksisikule haigusstaatus.

2.3 Tulemused

Hinnatud mudelitest osutus ebaoluliseks kõige madalamate haigestumistõenäosuste seti $(0.01, 0.04, 0.05)$ puhul vanemamudel. Seda võib põhjendada asjaoluga, et ha-

ruldaste haiguste puhul on väga tõenäoline, et haige lapse vanemad ei ole haiged ja seega pole haigestumist võimalik ennustada perekonnaajalugu kasutades.

Tabelis 2.2 on toodud simulatsiooni põhjal leitud mudelite AICd ja nende mudelite abil haigestumise prognoosimise headuse näitajad AUCd.

Tabel 2.2: Ühe geenimarkeri põhiste mudelite täpsushinnangud

Haigestumistõenäosused	Vanemamudel		Geenimudel	
p_0, p_1, p_2	AIC	AUC	AIC	AUC
0.01, 0.04, 0.05	*20662*	*0.505*	19807	0.674
0.05, 0.10, 0.20	51361	0.508	49872	0.614
0.10, 0.25, 0.50	89195	0.529	83325	0.660

Esmalt on märgata, et geenimudelid sobivad iga haigestumistõenäosuste seti korral andmetega kõige paremini. Vanemamudelitele vastavad $\Delta = AIC - AIC_{min}$ väärtused on 855, 1489 ja 5870. Kuna need erinevused on vähemalt suurusjärgu võrra suuremad, kui Δ-de puhul suurim sisuliselt oluline erinevus 10, siis on selge, et geeniandmete põhjal hinnatud haigestumismudelid sobivad andmetega paremini kui perekonnaajaloo põhjal hinnatud mudelid.

Sarnasele järeldusele võib jõuda ka AUCde võrdlemisel. Vanemamudeli korral jäävad AUC väärtused väga lähedale väärtusele 0.5 ehk nad on diagnostilises mõttes sama edukad kui mündivise. Kuigi on märgata trendi, et vanemamudeli ennustusvõimsus kasvab, kui SNPi mõju haigestumisele kasvab. Geenimudelite AUCd saavutavad iga haigestumistõenäosuste seti korral mõõduka tulemuse (> 0.6).

3 Saja markeriga mudel

Tegelikkuses sõltuvad paljud haigused mitmetest, isegi tuhandetest SNPdest. Selliste haiguste puhul leitakse iga üksiku SNP mõju haigestumisele ülegenoomsete assotsiatsiooniuuringute (GWAS - ingl. *Genome Wide Association Study*) käigus. Summeerides inimese genotüübis leiduva iga mõju omava SNP efektid saadakse nn. riskiskoor.

Riskiskoori kasutatakse diagnostiliste testide tegemisel. Inimesel, kelle riskiskoor on üle mingi kokkulepitud piirväärtuse, öeldakse, et on kõrgendatud risk haigestuda. Sarnase diagnostilise testi saab üles ehitada ka haigete pereliikmete arvu põhjal.

3.1 Simulatsioon

Järgnevalt vaadeldakse juhtu, kus haigestumist mõjutab, lisaks keskkonnamõjudele, täpselt sada teadaolevat SNPi. Iga SNP teoreetilise efekti suurused (tähistatakse $\beta_1, \beta_2, \ldots, \beta_{100}$) on valitud võrdseks hinnangutega varem läbiviidud GWASst (Morris jt, 2012), valides uuringus leitud saja väikseima p-väärtusega SNPd ehk haigusega kõige kindlamalt seostatud geenimarkerid. Valitud SNPd järjestati seejärel efektide absoluutväärtuste järgi, sest eeldusel, et ainult need SNPd omavad haigestumisele mõju, sõltub haigusrisk kõige rohkem absoluutväärtuselt suurima efektiga SNPst, seejärel absoluutväärtuselt järgmise suurima efektiga SNPst jne. Lisaks kasutatakse simulatsioonis GWAS uuringus toodud SNPde sagedusi vaadeldud populatsioonis.

GWAS uuringus on SNP efektid hinnatud logistilise regressioonmudeli abil. Teades haiguse levimust p populatsioonis, saab avaldada mudeli vabaliikme β_0 väärtuse:

$$\hat{\beta}_0 = \ln \frac{p}{1-p}$$

Teades kõiki logistilise regressioonmudeli kordajaid ja inimese genotüüpi, on võimalik leida valemi 1.3 põhjal hinnata inimese haigestumise tõenäosus.

Arvestades kirjeldatud taustsüsteemi haigestumise mõjudest, kusjuures keskkonna-

mõjud haigestumisele jäetakse vaatluse alt välja, genereeritakse järgneva juhise järgi andmestik:

1. Määratakse kindlaks soovitavate indeksisikute arv n.

2. Genereeritakse n vanemate paari, kusjuures vanemate genotüübid on üksteisest sõltumatud. Ühe vanema genotüüp koosneb 100st SNPst x_i ehk 200st alleelipaarist x_{i1} ja x_{i2}. Alleelid x_{i1} ja x_{i2} on mõlemad sõltumatud sama jaotusega $B(1, eaf_i)$. Vanema riskialleelide arvu saame kui: $x_i = x_{i1} + x_{i2}$.

3. Igale vanematepaarile genereeritakse 3 last. Lapse i-nda riskialleelide arv moodustub ühe vanema i-nda SNPi ühe alleeli x^1_{iy} ja teise vanema i-nda SNPi ühe alleeli x^2_{iz} summa. y ja z tähistavad, kumb vanema alleelist lapsele pärandub, seega suurused $y - 1$ ja $z - 1$ on mõlemad sõltumatud suurused jaotusest $Be(0.5)$.

4. Määratakse kindlaks uuritava haiguse levimus p populatsioonis ja leitakse levimusele vastav logistilise regressioonimudeli vabaliige β_0.

5. Leitakse igale genereeritud inimesele riskiskoori põhjal haigestumise tõenäosused π_k, kus k on inimese identifikaator. Et haigestumise mediaantõenäosus jääks samaks teoreetilise levimusega, tuleb riskiskoor tsentreerida. Selleks lahutatakse enne haigestumistõenäosuse avaldamist inimese riskiskoorist kõigi $5n$ genereeritud inimeste riskiskooride keskmine. Valem 1.3 saab kokkuvõttes kuju:

$$\pi_k = \frac{1}{1 + \exp^{\beta_0 + \sum_{i=1}^{100} \beta_i x_i^k - \frac{1}{5n} \sum_{k=1}^{5n} \sum_{i=1}^{100} \beta_i x_i^k}},$$

   kus x_i^k on k-nda inimese i-nda SNPi riskialleelide arv.

6. Vastavalt inimese tõenäosusele haigestuda genereeritakse inimese haigusstaatus jaotusega $B(1, \pi_k)$. Seejärel loetakse iga lapse jaoks kokku haigete vanemate arv H_{VV} ja haigete pereliikmete (indeksisik va) arv H_P ning tähistagu H_{IND} indeksisiku haigusstaatust.

Nüüd on kirjeldatud andmete simuleerimise eeskiri. Praktikas ei ole tegelikud SNPde efektid teada ja need tuleb hinnata. Selleks viiakse indeksisikute genotüüpide ja haigusstaatuste andmeid kasutades läbi GWAS uuring, mille eeskiri on samm 7.

7. Hinnatakse SNPde efektid. Selle jaoks sobitatakse logistilised regressioonmudelid $H_{IND} \sim x_i$, kus x_i on indeksisiku i-nda SNPi riskialleelide arv. Saadud mudeli regressioonikordaja loetakse hinnanguks $\hat{\beta}_i$.

8. Hinnatakse indeksisikute riskiskoorid $RS_{k,m}$, kasutades $m = 100, 50, 25, 10, 5$ ja 1 kõige suurema efektiga SNPe:

$$RS_{k,m} = \sum_{j=1}^{m} \beta_j \cdot x_{k,j}, \text{ kus } x_{k,j} \text{ on } k\text{-nda indeksisiku } j\text{-nda markeri väärtus.}$$

Andmestik genereeritakse $n = 10^5$ indeksisikuga. Sammud 4 - 8 viiakse läbi valides haiguse levimuseks $0.005, 0.05$ ja 0.339.

Iga valitud levimuse korral sobitatakse logistilised regressioonimudelid $H_{IND} \sim H_{VV}$ (nn vanemamudel), $H_{IND} \sim H_P$ (nn peremudel) ja riskiskooridele vastavad mudelid $H_{IND} \sim RS_m$, kus $m = 100, 50, 25, 10, 5, 1$ (nn SNP100 mudel, SNP50 mudel jne).

Mitmetel põhjustel võib juhtuda, et reaalses olukorras ei ole teada kõige suuremat mõju omavad geenimarkerid - tihti on uuritava tunnuse päritavus märksa suurem kui geenimarkerite poolt kirjeldatud osa tunnuse varieeruvusest. Et jäljendada seda olukorda, hinnati iga valitud levimuse korral lisaks mudelid riskiskooridega, mis moodustati 5 ja 10 juhusliku markeri põhjal. Selle jaoks genereeriti ühtlasest jaotusest vastavalt 5 ja 10 juhuslikku arvu vahemikus $[1, 100]$. Viie markeri mudelisse osutusid valituks suurima efekti põhjal järjestatult markerid järjekorranumbriga $10, 21, 45, 65, 78$ ja kümne markeriga mudelisse markerid järjekorranumbriga $9, 36, 42, 47, 49, 58, 60, 62, 71, 80$. Sarnaselt sammule 8 saadakse riskiskoorid RS_{5J} ja RS_{10J}. Seejärel sobitatakse logistilised regressioonimudelid $H_{IND} \sim RS_{5J}$ (nn SNP5 J mudel) ja $H_{IND} \sim RS_{10J}$ (nn SNP10 J mudel).

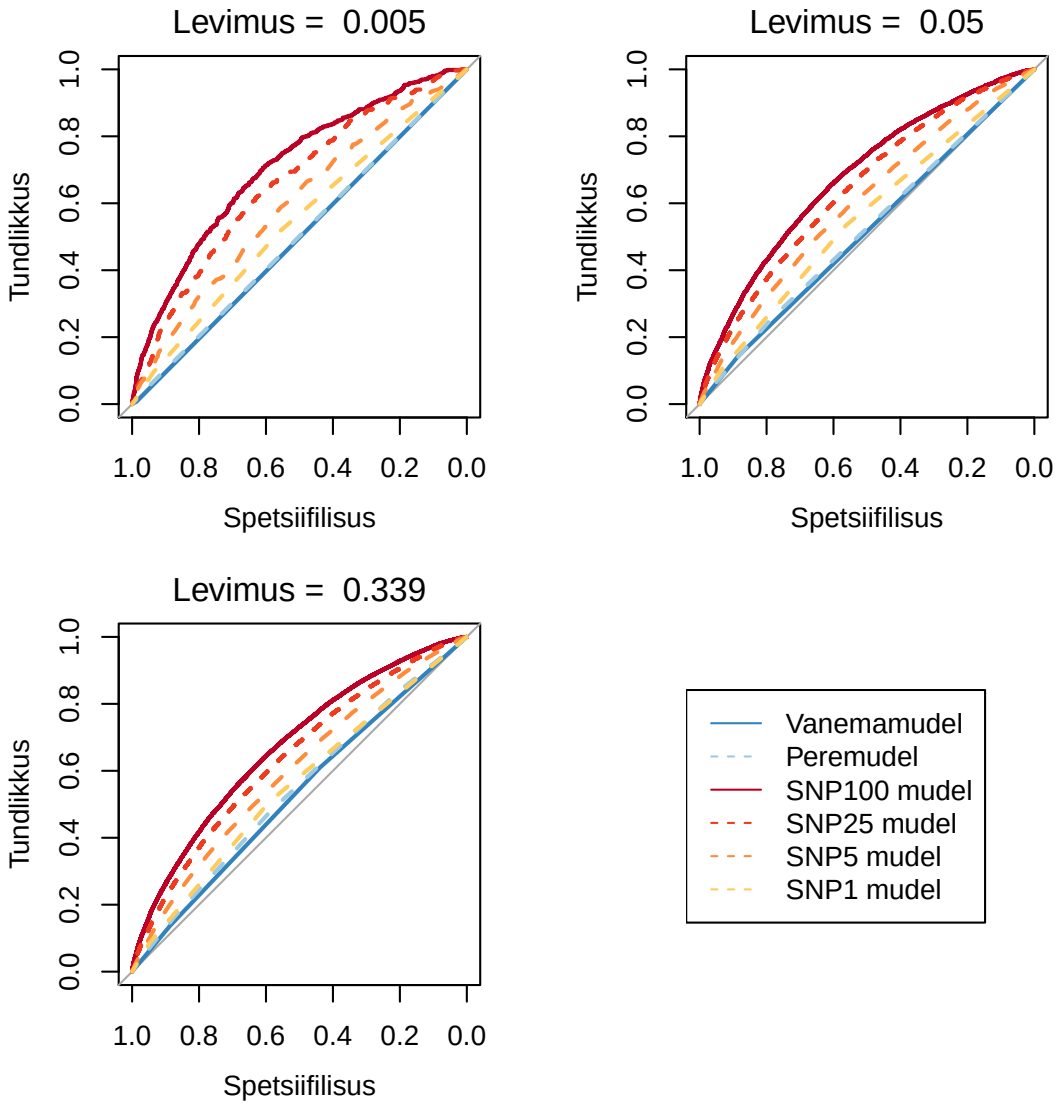
3.2 Tulemused

Üle kõigi hinnatud mudelite osutusid ebaolulisteks vanema- ja peremudelid levimuse 0.005 korral. Ilmselt on see jällegi tingitud asjaolust, et kui populatsioonis esineb haigust väga vähe, siis on ebatõenäoline, et indeksisikul on üks või rohkem haiget pereliiget, isegi kui indeksisik on haigestunud, mistõttu ei ole võimalik luua head mudelit.

Tabel 3.1: Saja geenimarkeri simulatsiooni mudelite $\Delta = AIC - AIC_{min}$ väärtused

	Mudel									
Levimus	Vanema	Pere	SNP1	SNP5	SNP10	SNP25	SNP50	SNP100	SNP5 J	SNP10 J
0.005	*325*	*327*	311	262	203	146	79	0	313	293
0.05	2310	2287	2103	1636	1300	860	403	0	2278	2162
0.339	8469	8180	7712	5991	4914	3227	1474	0	8449	8141

Tabelis 3.1 on toodud leitud mudelite AIC väärtuste erinevused vastava levimuse minimaalsest AIC väärtusest. Ootuspäraselt sobivad andmetega kõige paremini SNP100 mudelid, kuna nendes on kasutatud kõik tekkemehhanismis esinevad SNPd. Tähelepanuväärne on, et kõige väiksema levimuse korral on SNP10 J mudel parem SNP1 mudelist ja SNP5 J mudel on võrreldav SNP1 mudeli headusega, aga kõrgemate levimuste korral jätavad juhuslike SNPdega mudelid seletamata olulise osa andmete varieeruvusest võrreldes kõigi teiste geenimudelitega. Levimuse 0.339 korral sobib peremudel andmetega paremini kui SNP5 J mudel, aga see on ainus olukord, kus perekonnaajaloopõhine mudel edestab mõnda riskiskooripõhist mudelit. Perekonnaajaloopõhiseid mudeleid võrreldes on näha, et kui mõlemad mudelid on olulised, siis õvedega kaasamisega mudelisse on seletatud oluliselt suurem osa andmete varieeruvusest, kui ainult vanemamudeli korral.



Joonis 3.1: Simulatsiooni mõnede mudelite ROC kõverad.

Tabel 3.2: Saja geenimarkeri simulatsiooni mudelite AUC väärtused

	Mudel									
Levimus	Vanema	Pere	SNP1	SNP5	SNP10	SNP25	SNP50	SNP100	SNP5 J	SNP10 J
0.005	*0.497*	*0.502*	0.542	0.594	0.630	0.659	0.682	0.706	0.523	0.570
0.05	0.514	0.521	0.553	0.600	0.621	0.644	0.664	0.679	0.532	0.553
0.339	0.528	0.542	0.553	0.597	0.614	0.637	0.657	0.672	0.532	0.545

Joonisel 3.1 on välja toodud kuue mudeli ROC kõverad ja tabelis 3.2 on näha kõigile mudelitele vastavaid AUC väärtuseid. On näha, et mida suurem on haiguse levimus, seda sarnasemad on mudelite ennustused haigestumisele. Perekonnaajalugu kasutavate mudelite AUCd erinevad juhuslikust arvamisest nähtavalt vaid levimuse 0.339 korral, ja ka siis väga vähesel määral. Seega võib järeldada, et perekonnaajaloo seos haigestumisega on meie simulatsiooni tingimustes liiga nõrk, et indiviidi haigusstaatust ennustada. Ootuspäraselt, mida rohkem kõige suurema mõjuga SNPe on mudelis, seda paremini klassifitseerivad mudelid indiviidide haigestumise põhjal. Juhuslikult valitud markeritega mudelid edestavad perekonnaajaloopõhiseid mudeleid, kuid on selgelt nõrgema eristamisvõimega, kui teised geenimudelid. Ainus olukord, kus perekonnaajaloopõhine mudel edestab geneetilistelt andmetelt hinnatud mudelit on levimuse 0.339 korral, kui peremudel on parema eristamisvõimega kui SNP5 J mudel. Vaid levimuse 0.005 korral oli SNP10 J mudel parema AUC väärtusega, kui SNP1 mudel.

Kuna reaalsuses leidub haiguseid, mille haiguseriski saab ennustada perekonnaajaloo põhjal täpsemalt kui geneetilise riskiskoori põhjal (Do jt, 2012), siis võib arvata, et juhuslikult valitud markeritega mudelid sarnanevad enim päriselt kasutuses olevatele mudelitele. Seetõttu võib arvata, et kui oskused haiguse geneetilist komponenti mõõta veelgi paranevad, tuleks praktikas eelistada geenimarkerite kasutamist perekonnaajaloole.

Lisas 3 on toodud varasema sarnase uuringu (Do jt, 2012) AUC väärtused erinevate haiguste korral, mille levimused on võrreldavad siin töös kasutatutega. Kuigi erinevate haiguste puhul omavad erinevad SNPd erinevate suurustega efekte, siis võrreldes Do uuringus SNP mudeleid ligilähedaselt samade SNPde arvuga mudelitega siit tööst (vt tabel 3.2), on näha, et tulemused on küllalt sarnastes suurusjärkudes. Suurimad erinevused esinevad perekonnaajaloopõhiste mudelite hinnangute headuses. Võib oletada, et suur osa sellest erinevusest on tingitud kasutatud perekonnastruktuurist, kuna Do uuringus on modelleeritud palju suuremat perekonda.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli simulatsioonuuringu käigus välja selgitada, kui hästi prognoosivad haiguseriske erinevatel tingimustel geenimarkerite ja perekonnaajaloo põhjal hinnatud mudelid. Töös vaadeldi kahte olukorda: haigestumist mõjutab täpselt üks SNP ja haigestumist mõjutavad täpselt sada SNPi.

Läbi viidi simulatsioonid erinevate riskialleelide haigestumistõenäosuste, riskialleelide sageduste ja haiguse levimuste korral. Saadud andmetelt hinnati mitmed mudelid perekonnaajaloo ja geeniandmete põhjal ning seejärel võrreldi mudelite sobivust andmetega ja nende diskrimineerimisvõimet haigete eristamisel tervetest.

Selgus, et geenimarkerite põhjal haigusriski hindavad mudelid edestasid perekonnaajaloo põhjal hinnatud mudeleid pea kõigil kontrollitud tingimustel. Eriti suur erinevus uuritavate meetodite vahel esineb siis, kui haiguse levimus on madal. Selliste haruldaste haiguste puhul osutusid perekonnaajaloopõhised mudelid ebaolulisteks ja geenimarkerite põhjal hinnatud mudelid eristasid haigeid või kõrge riskiga inimesi tervetest kõige edukamalt.

Töös uuritud olukorrad erinevad reaalsusest, kuna eeldasime, et haiguse geneetiline komponent on täielikult määratud teadaolevate geenimarkerite poolt. Tegelikkuses on GWASi tulemuste põhjal enamasti võimalik kirjeldada vaid väga väikest osa pärilikkusest (Do jt, 2012). See võib olla seotud nii sellega, et GWAS ei võimalda leida harva alleelisagedusega ($< 1\%$), aga tugeva mõjuga markereid kui ka sellega, et haiguse pärilik komponent on mõjutatud ka otseselt DNAs mitte sisalduvate geneetiliste tegurite (nt DNA metülatsiooni) poolt. Ka selles töös saadud tulemustest on näha, et kui riskiskoorist välja jätta kõige suurema efektiga markerid, on geenimarkerite põhised mudelid vahel kehvemad perekonnaajaloo põhjal saadud mudelitest.

Varasemate uuringute põhjal on alust arvata, et kui on teada suurema sugupuu haiguslugu, siis ka perekonnaajaloopõhiste mudelite headus kasvab, kuid lisaks tuleks uurida, kui laialdased on praktikas patsientide teadmised oma sugulaste haigustest.

Kasutatud kirjandus

Burnham, K. ja Anderson, D. (2002). Model selection and multimodel inference. New York: Springer.

Burton, P., Tobin, M. ja Hopper, J. (2005). Key concepts in genetic epidemiology. The Lancet, 366(9489), lk 941-951. doi: 10.1016/s0140-6736(05)67322-9.

Dnabaser.com. (2016). Single nucleotide polymorphism (SNP) detection and analysis software. Kättesaadav: http://www.dnabaser.com/articles/SNP/SNP-single-nucleotide-polymorphism.html [Vaadatud 27.04.2016].

Do, C., Hinds, D., Francke, U. ja Eriksson, N. (2012). Comparison of Family History and SNPs for Predicting Risk of Complex Disease. PLoS Genetics, 8(10), p.e1002973. doi: 10.1371/journal.pgen.1002973.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), lk 861-874. doi: 10.1016/j.patrec.2005.10.010.

Heinaru, A. (2012). Geneetika. Tartu Ülikooli Kirjastus.

Kaart, T. (2011). Loomade aretusväärtuse hindamine ja aretusprogrammid. Loengukursus. Eesti Maaülikool. Kättesaadav: http://www.eau.ee/~ktanel/VL_0192/pt4_2012.pdf [Vaadatud 12.03.2016]

Kasela, S. (2011). Ülegenoomne assotsiatsiooniuuring ja selle praktiline läbiviimine TÜ Eesti Geenivaramu andmete põhjal. Bakalaureusetöö. Tartu Ülikool.

Käärik, E. (2015). Andmeanalüüs II. Loengukursus. Tartu Ülikool. Kättesaadav: http://dspace.ut.ee/bitstream/handle/10062/35401/AndmeanaluusII.pdf [Vaadatud 27.03.2016]

Morris, A., Voight, B., Teslovich, T., Ferreira, T., Segrè, A. jt. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysio-

logy of type 2 diabetes. Nature Genetics, 44(9), lk 981-990. doi:10.1038/ng.2383.

Taskutark.ee. (2016). Mendeli seadused. Kättesaadav: http://www.taskutark.ee/m/mendeli-seadused/?auth=dGFza3V0YXJr [Vaadatud 28.04.2016].

Wassertheil-Smoller, S. (2004). Biostatistics and epidemiology. New York: Springer-Verlag.

Lisad

Lisa 1. Mõistete sõnastik

AIC - Akaike informatsioonikriteerium. Samade andmete pealt hinnatud mudelitest sobib andmetega kõige paremini väikseima AIC väärtusega mudel.

Alleel - Üks võimalikest SNP variantidest kindlas DNA lõigus.

AUC - ROC kõvera alune pindala. Näitab tõenäosust, et test määrab juhuslikult valitud, tegelikult haige, inimese haigeks suurema tõenäosusega kui juhuslikult valitud, tegelikult terve, inimese haigeks.

DNA - Päriliku informatsiooni edasikandja inimestel. Koosneb nukleotiididest tähistustega A, T, C ja G.

Fenotüüp - Indiviidi genotüübi ja keskkonnamõjude tulemusel realiseerunud omaduste kogum.

Geen - DNA segment, mis kodeerib ühte valgu molekuli. Geenid moodustuvad alleelidest, mis üks on pärit emalt ja teine isalt.

Geenimarker - SNP, mille abil identifitseeritakse haigusega seotud geene.

Genotüüp - Indiviidi geenide kogum.

GWAS - Ülegenoomne assotsiatsiooniuuring. Kasutatakse komplekshaiguste puhul SNPi efektide hindamiseks.

MAF - Konkreetse SNPi puhul populatsioonis kõige harvemini esineva alleeli sagedus.

Mendeli seadused - 1. Ristates kahte homosügootset isendit, on esimese järglaspõlvkonna isendid omavahel geneetiliselt sarnased. 2. Ristates heterosügootseid isendeid, tekib järglaspõlvkonnas lahknemine genotüübilise ja fenotüübilise tunnuse avaldu-

misel. Kodomineerivate alleelide puhul on nii genotüübiline kui fenotüübiline lahknemine suhtes 1:2:1. 3. Kaks geeni päranduvad üksteisest sõltumatult. (Taskutark.ee, 2016)

Päritavus - Koefitsient, mis näitab kui suur osa fenotüübilisest pärilikkusest on tingitud genotüübist.

SNP - Üksiku nukleotiidi muutusest põhjustatud kahe või enama geneetilise variandi olemasolu populatsioonis. Nt. osadel inimestel on DNA lõik AT**A**AC ja teistel on AT**G**AC.

Lisa 2. Programmikoodid

Simulatsioonide koodid on saadaval Dropboxi kaudu vähemalt kuni 01.07.2016 aadressil:

http://tinyurl.com/koodid-BSc-Sille-Habakukk

Hiljem võib olla vajalik kontakteeruda autoriga: sille.habakukk@eesti.ee

Lisa 3. Varasema uuringu tulemused

Tabel 3.3: Varasema uuringu tulemuste AUC väärtused (Do jt, 2012)

Haigus	Levimus	Peremudel	SNP mudel	Kasutatud SNP arv
Crohni tõbi	0.005	0.551	0.717	60
Ealine maakuli degeneratsioon	0.047	0.66	0.758	9
Soolevähk	0.051	0.526	0.564	11
Bipolaarne häire	0.051	0.637	0.55	5
2. tüüpi suhkrutõbi	0.339	0.587	0.592	21

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Sille Habakukk

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Geenimarkerite põhjal hinnatud haiguseriski võrdlus perekonnaajalooga", mille juhendajad on Krista Fischer ja Kristi Läll

    1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

    1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 29.04.2016