Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Laura Birgit Luitva

Jämesoolevähi riskitegurid TÜ Eesti Geenivaramu andmete põhjal

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendajad: Jaanika Kronberg, PhD

Krista Fischer, PhD

Tõnu Esko, PhD

Tartu 2020

Jämesoolevähi riskitegurid TÜ Eesti Geenivaramu andmete põhjal

Bakalaureusetöö

Laura Birgit Luitva

Lühikokkuvõte. Bakalaureusetöö eesmärk on leida jämesoolevähi mittegeneetilised riskitegurid ning uurida geneetilise riskiskoori mõju jämesoolevähile. Lisaks uuritakse, kui hästi leitud mittegeneetilised riskitegurid ja geneetiline riskiskoor jämesoolevähki prognoosivad. Töös kasutatakse Tartu Ülikooli Eesti Geenivaramu andmeid, kus on üle 48 000 geenidoonori. Andmete analüüsimisel kasutatakse elukestusanalüüsi meetodeid. Seejuures arvestatakse, et tegemist on vasakult tõkestatud ja paremalt tsenseeritud andmetega, ning ajaskaalana kasutatakse vanust. Töö teoreetilises osas antakse ülevaade jämesoolevähist, elukestusanalüüsist, geneetilisest riskiskoorist ning ROC-kõveratest. Töö praktilises osas kirjeldatakse andmeid ning leitakse jämesoolevähki prognoosivad tunnused. Seejärel jagatakse andmestik treening- ja testandmestikuks. Treeningandmestikus leitakse jämesoolevähki prognoosivad mudelid ning testandmestikus prognostilised skoorid ja ROC-kõverate abil prognooside täpsused.

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

Võtmesõnad: jämesoolevähk, elukestusanalüüs, geneetiline riskiskoor, ROC-kõver

Risk factors for colorectal cancer based on Estonian Genome Centre data

Bachelor's thesis

Laura Birgit Luitva

Abstract. The aim of this bachelor's thesis is to find non-genetic risk factors for colorectal cancer, to study the effect of polygenic risk score and evaluate how well the non-genetic risk factors and polygenic risk score predict colorectal cancer. The Estonian Genome Centre data used for analysis includes more than 48,000 gene donors. The methods of survival analysis are used in analysing the data with age being used as a time scale and with left truncation and right censoring taken into account. The theoretical part of the thesis gives an overview of colorectal cancer, survival analysis, polygenic risk score and ROC curves. The practical part gives an overview of the data and determines the risk factors for colorectal cancer. The data is divided into a training and test set. The training

set is used to find the models for predicting colorectal cancer. Prognostic scores are found in the test set, as well as the accuracy of the predictions using ROC curves.

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

Keywords: colorectal cancer, survival analysis, polygenic risk score, ROC curve

Sisukord

Sissejuhatus

Jämesoolevähk on pahaloomuline kasvaja, mille esmasjuhtude arv on aastatega aina kasvanud ning mis moodustab Eestis ligikaudu 10% kõikidest vähijuhtudest [1]. Jämesoolevähi riski mõjutab nii pärilikkus kui ka erinevad mittegeneetilised riskitegurid, mis on peamiselt seotud elustiiliga ja erinevate haigustega. Käesoleva bakalaureusetöö eesmärk on leida jämesoolevähki prognoosivad mittegeneetilised tegurid ning uurida geneetilise riskiskoori mõju jämesoolevähile. Veel pakub huvi, kui hästi leitud mittegeneetilised riskitegurid ja geneetiline riskiskoor jämesoolevähki prognoosivad. Töös kasutatakse Tartu Ülikooli Eesti Geenivaramu andmeid, kus on üle 48 000 geenidoonori.

Töö jaguneb teoreetiliseks ja praktiliseks osaks. Teoreetilisse osasse kuulub neli peatükki. Esimeses peatükis selgitatakse, mis on jämesoolevähk ning selle riskitegurid. Teises peatükis antakse ülevaade töös kasutatavatest elukestusanalüüsi meetoditest ning kolmandas peatükis tutvustatakse ülegenoomset seoseuuringut ja geneetilist riskiskoori. Viimasena antakse ülevaade prognooside täpsust iseloomustavatest ROC-kõveratest.

Praktilises osas kirjeldatakse esmalt andmeid ja kasutatud tunnuseid. Seejärel leitakse jämesoolevähki prognoosivad mittegeneetilised tegurid ning uuritakse geneetilise riskiskoori mõju jämesoolevähile. Andmestik jagatakse treening- ja testandmestikuks, et uurida, kui hästi mittegeneetilised riskitegurid ja geneetiline riskiskoor jämesoolevähki prognoosivad. Treeningandmetel leitakse jämesoolevähi riski prognoosiv mudel, mida testitakse testandmetel. Lisaks tuuakse välja ka mudel, kus ei sisaldu geneetilist riskiskoori, ning mudel, kus sisaldub vaid geneetiline riskiskoor, et võrrelda mudelite prognoosimise täpsusi.

Töös kasutatakse elukestusanalüüsi meetodeid, kus ajaskaalana kasutatakse vanust ning arvestatakse, et tegemist on paremalt tsenseeritud ning vasakult tõkestatud andmetega. Täpsemalt kasutatakse Coxi võrdeliste riskide mudeleid ning tulemuste illustreerimiseks Kaplan-Meieri hinnanguid üleelamisfunktsioonile. Prognooside täpsusi hinnatakse ROC-kõverate abil. Analüüside läbiviimisel ning tulemuste graafilisel kujutamisel kasutatakse rakendustarkvara R ning töö kirjutamisel programmi LaTeX.

Töö autor tänab juhendajat Krista Fischerit rohkete nõuannete ja selgituste eest ning juhendajaid Jaanika Kronbergi ja Tõnu Eskot TÜ Eesti Geenivaramu andmete kasutamise loa eest.

1 Jämesoolevähk

Vähk ehk pahaloomuline kasvaja on geneetiline haigus, mida põhjustavad geenides toimunud muutused ehk mutatsioonid. Mutatsioonide tagajärjel tekivad pahaloomulised kasvajarakud, mille paljunemine on kontrollimatu. Rakkude pidurdamatu paljunemise tõttu saavad vähkkasvajad levida edasi ka teistesse organitesse. Vähki põhjustavad geenimuutused võivad olla nii päritud kui ka elu jooksul tekkinud. [2]

Jämesool asub seedekulga alumises osas ning koosneb käärsoolest ja pärasoolest. Jämesoole ülesandeks on teostada vee ja mineraalsoolade ainevahetust ning samuti seedimisest tekkinud jääkainete lagundamist, hoidmist ja väljutamist. [3]

Jämesoolevähk ehk käär- ja pärasoolevähk on käärsooles või pärasooles tekkinud vähkkasvaja [3]. Täpsemalt on jämesoolevähile vastavad RHK-10 koodid C18, C19, C20 ja C21 [1]. Koodid C18–C21 tähistavad järgmiseid pahaloomulisi kasvajaid:

- C18 - käärsoole pahaloomuline kasvaja;
- C19 - pärasoole ja sigmakäärsoole ühenduskoha pahaloomuline kasvaja;
- C20 - pärasoole pahaloomuline kasvaja;
- C21 - päraku ja pärakukanali pahaloomuline kasvaja [4].

1.1 Jämesoolevähi riskitegurid

Jämesoolevähi riskifaktorite alla kuuluvad vanus, pärilikkus, mitmed haigused ja elustiiliga seotud riskitegurid. Elustiiliga seotud riskifaktoriteks on toitumisharjumused, vähene füüsiline aktiivsus ning suitsetamine. Söömisharjumustest suurendab jämesoolevähi saamise riski toitumine, mis sisaldab vähe kiudaineid, kuid palju loomset rasva ja loomset valku. [3]

On teada, et jämesoolevähki haigestumise risk on suurem nendel inimestel, kellel on esinenud soolepolüüpe ehk soole limaskesta healoomulisi kasvajaid. Riski suurendavad ka põletikulised soolehaigused: Crohni tõbi ja haavandiline jämesoolepõletik. Haigustest on riskifaktoriks veel esimest tüüpi diabeet. Samuti on jämesoolevähi risk suurem, kui lähisugulastel on esinenud jämesoolevähki, mõne muu organi vähki või polüüpe sooles. [3]

2 Elukestusanalüüs

Siinses peatükis ja järgnevates alapeatükkides kirjeldatav metoodika põhineb David Colleti raamatul „Modelling survival data in medical research“ juhul, kui ei ole märgitud teisiti [5].

Elukestusanalüüsi meetodeid kasutatakse andmete puhul, kus uuritavaks tunnuseks on ajavahemiku pikkus fikseeritud algmomendist kuni kindla sündmuse toimumiseni (lõppmomendini). Andmeid, mis sisaldavad mingi protsessi kestust, nimetatakse kestusandmeteks. Kestusandmetes sisalduvad ajavahemiku pikkused ei saa olla negatiivsed. Üldiselt ei ole kestusandmed ka sümmeetrilise jaotusega. Seega ei saa kestusandmete korral rakendada standardseid statistilisi protseduure, kuna vajalik normaaljaotuse eeldus on rikutud. Kestusandmete eripäraks on veel tsenseeritus, millest kirjutatakse järgmises alapeatükis.

Eelnevalt kirjeldatud ajavahemiku pikkust fikseeritud algmomendist kuni teatud sündmuse toimumiseni nimetatakse elukestusanalüüsis elukestuseks. See aga ei tähenda, et tegemist peab olema just elu kestusega, st huvipakkuvaks sündmuseks ei pea olema surm. Näiteks võib vaadeldavaks sündmuseks olla valu leevenemine, sümptomite esinemine, ülesande lõpetamine, elektroonilise seadme katki minemine jpm. Seega on elukestusanalüüsis kasutatavad analüüsimeetodid rakendatavad paljudes erinevates valdkondades.

Siinses töös käsitletakse huvipakkuva sündmusena jämesoolevähi diagnoosimist ning elukestusena ajavahemiku pikkust alates ajamomendist, millal indiviid liitus TÜ Eesti Geenivaramu kohordiga, kuni ajamomendini, millal indiviidil diagnoositi jämesoolevähk.

2.1 Tsenseeritus

Subjekti elukestust nimetatakse tsenseerituks, kui huvipakkuvat sündmust ei ole subjektil vaadeldud. Tavaliselt on tsenseerimine tingitud sellest, et subjektil pole sündmus enne katse lõppu toimunud. Tsenseerimisega on tegemist ka juhul, kui subjekti katsest lahkumise või mõne muu põhjuse tõttu ei ole informatsiooni sündmuse toimumise kohta. Vahel võib tsenseerimiseks lugeda ka indiviidi surma uuringu jooksul, kui surm pole huvipakkuvaks sündmuseks.

Tsenseerimist saab liigitada kolmeks: paremalt tsenseerimine, vasakult tsenseerimine ja

intervall-tsenseerimine. Käesolevas töös käsitletakse paremalt tsenseerimist, mille korral on teada viimane ajahetk, millal subjektil ei olnud veel sündmus toimunud. Seega on tegemist paremalt tsenseeritusega, kui vaadeldud elukestus on väiksem subjekti tegelikust elukestusest, mis on teadmata. Vastupidiselt on vasakult tsenseerimisel teada, et subjektil toimus sündmus enne tsenseerimist, st subjekti tegelik elukestus on väiksem kui vaadeldud elukestus. Juhul, kui on teada intervall, milles subjektil toimus huvipakkuv sündmus, siis nimetatakse vaatlust intervall-tsenseerituks.

Olgu T juhuslik suurus, mis vastab huvipakkuva sündmuse toimumise ajale, ja olgu juhuslik suurus C tsenseerimise aeg. Tsenseeritud andmete analüüsimisel on tähtsaks eelduseks, et subjektil on sündmuse toimumise aeg T ning tsenseerimise aeg C sõltumatud juhuslikud suurused. Paremalt tsenseeritud andmete korral tähendab see, et subjekti tsenseerimise aeg C ei anna mingit informatsiooni sündmuse toimumise aja T kohta, peale teadmise, et tsenseerimise aeg on väiksem sündmuse toimumise ajast ($C < T$). Sel juhul nimetatakse tsenseeritust mitteinformatiivseks. Vastasel korral on tegemist informatiivse tsenseerimisega. Siinses töös käsitletakse vaid mitteinformatiivset tsenseerimist.

2.2 Ajaskaala valik ja vasakult tõkestatus

Nagu eespool mainitud, on uuritavaks tunnuseks ajavahemiku pikkus teatud algmomendist kuni huvipakkuva sündmuse toimumiseni. Algmomendi valik määrab ka analüüsis kasutatava ajaskaala. Ajaskaalana on võimalik kasutada kalendriaega, vanust või aega alates mingist sündmusest. Ajaskaala valik määrab, milliseid subjekte omavahel võrreldakse. Seega näiteks ajaskaalana vanust kasutades võrreldakse igal ajamomendil subjekte, kes on sellel ajamomendil samas vanuses. Ajaskaala valik on tähtis, sest vale valiku korral võib saada nihkega hinnangud. [6]

Epidemioloogilistes uuringutes on tihti mõistlik võrrelda samas vanuses inimesi, kuna paljude haiguste korral mõjutab vanus haigestumise riski. Seega kasutatakse epidemioloogilistes uuringutes ajaskaalana sageli vanust. Ka käesolevas töös kasutatakse ajaskaalana vanust. Juhul kui ajaskaalaks on valitud vanus, siis tuleb arvestada, et fikseeritud ajamomendi (vanuse) korral ei ole osa subjekte veel vaatluse alla jõudnud ning osa subjekte on vaatluse alt juba väljunud. [7]

Lisaks tuleb siinses töös arvesse võtta vasakult tõkestatust, mis tähendab, et uuringuga said liituda ainult need subjektid, kes olid uuringu alguses liitumise hetkel elus. Vasakult tõkestatust arvestamata võib saada nihkega hinnangud. Juhul kui ajaskaalana kasutatakse vanust ning arvestatakse vasakult tõkestatusega, siis võrreldakse igat subjekti samas vanuses subjektidega, kes olid selles vanuses veel uuringus. [6]

2.3 Üleelamis- ja riskifunktsioon

Olgu T juhuslik suurus, mis tähistab subjekti elukestust ja mille võimalikud väärtused on mittenegatiivsed. Olgu elukestuse T tihedusfunktsioon $f(t)$ ning jaotusfunktsioon $F(t)$, mille korral kehtib:

$$F(t) = P(T < t) = \int_0^t f(u)du.$$

Üleelamisfunktsioon $S(t)$ on defineeritud kui tõenäosus, et huvipakkuv sündmus ei toimu enne ajamomenti t:

$$S(t) = P(T \geq t) = 1 - F(t) = 1 - \int_0^t f(u)du.$$

Riskifunktsiooniks $h(t)$ nimetatakse tõenäosust, et sündmus toimub ajamomendil t tingimusel, et see ei toimunud enne seda ajamomenti. Pideva aja korral on riskifunktsioon defineeritud järgmiselt:

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right\}.$$

Viimasest võrdusest saab tuletada seose riskifunktsiooni $h(t)$ ja üleelamisfunktsiooni $S(t)$ vahel, mis avaldub kujul:

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

kus $f(t)$ on elukestuse T tihedusfunktsioon.

2.4 Kaplan-Meieri hinnang

Kaplan-Meieri hinnang on peamine meetod, millega hinnatakse üleelamisfunktsiooni tsenseeritud andmete jaoks. Olgu n subjekti ning olgu subjektidel toimunud sündmus või

tsenseerimine ajamomentidel $t_1, t_2, \ldots, t_n$. Olgu sündmused toimunud r erineval ajamomendil ($r \leq n$). Sündmuste toimumise ajad järjestatakse kasvavas järjekorras. Suuruselt j-ndat ajamomenti tähistatakse suurusega $t_{(j)}$, $j = 1, \ldots, r$, st järjestatud ajamomendid on $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Veel olgu vahetult enne ajamomenti $t_{(j)}$ vaatluse all n_j subjekti (riskigrupi suurus) ning olgu ajamomendil $t_{(j)}$ toimunud d_j sündmust.

Kaplan-Meieri hinnang üleelamisfunktsioonile avaldub kujul:

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

kus $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, \ldots, r$ ning $t_{(r+1)} = \infty$. Seega on üleelamisfunktsiooni hinnangu leidmiseks piisav arvutada korrutise tegurid vaid ajamomentidel, kus sündmus toimus. Üleelamisfunktsiooni graafiku saamiseks kantakse leitud Kaplan-Meieri hinnangud joonisele.

2.5 Võrdeliste riskide mudel

Olgu vaatluse all 2 gruppi ning olgu gruppide riskifunktsioonid $h_1(t)$ ja $h_2(t)$. Olgu gruppidel võrdelised riskid ehk riskide suhe on igal ajamomendil t konstantne. Sel juhul avaldub võrdeliste riskide mudel kujul:

$$h_1(t) = h_2(t)\psi,$$

kus ψ on konstant.

Nüüd saab teha üldistuse olukorrale, kus vaatluse all on n subjekti. Olgu i-nda subjekti riskifunktsioon $h_i(t)$ ning olgu iga subjekti korral teada argumenttunnuste $X_1, X_2, \ldots, X_p$ väärtused $x_{1i}, x_{2i}, \ldots, x_{pi}$, kus $i = 1, \ldots, n$. Tähistagu $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ i-nda subjekti argumenttunnuste väärtuste vektorit. Riski i-ndal subjektil saab leida kujul:

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

kus $h_0(t)$ on baasriskifunktsioon ning $\psi(\mathbf{x}_i)$ on funktsioon väärtustest $\mathbf{x}_i$.

Riskide suhe $\psi(\mathbf{x}_i)$ ei saa olla negatiivne, seega sobib selleks võtta eksponentfunktsioon:

$$\psi(\mathbf{x}_i) = e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}},$$

kus $\beta_1, \beta_2, \ldots, \beta_p$ on teatud parameetrid. Tähistagu $\boldsymbol{\beta}$ parameetrite β_i vektorit ning seega $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$. Nüüd saab mudeli kirjutada kujul:

$$h_i(t) = h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}_i} = h_0(t)e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}}. \tag{1}$$

Saadud mudel on elukestusanalüüsis kõige sagedamini kasutatav regressioonmudel ning seda nimetatakse võrdeliste riskide või proportsionaalsete riskide mudeliks. Parameetrilisel juhul määratakse võrdeliste riskide mudelis baasriskifunktsioon $h_0(t)$ vastava jaotuse parameetritega. Poolparameetriliste mudelite korral hinnatakse vaid parameetreid $\beta_1, \beta_2, \ldots, \beta_p$, st baasriski $h_0(t)$ ei hinnata.

2.6 Coxi võrdeliste riskide mudel

Coxi võrdeliste riskide mudel ei hinda regressioonmudelis (1) baasriski $h_0(t)$. Seega kasutab Coxi võrdeliste riskide mudel poolparameetrilist lähenemisviisi. Selleks, et tsenseeritud andmete korral hinnata võrdeliste riskide mudelis (1) parameetreid $\boldsymbol{\beta}$, tuletas D. R. Cox nn osalise tõepärafunktsiooni, milles ei sisaldu baasriski $h_0(t)$. Mainitud tõepärafunktsiooni konstrueerimisel lähtutakse sellest, et parameetrite $\boldsymbol{\beta}$ hindamisel ei ole olulised täpsed sündmuste toimumise ajad, vaid nende aegade suhteline järjestus.

Olgu n subjekti ning olgu ajahetkel t_i toimunud sündmus või tsenseerimine i-ndal subjektil $(i = 1, \ldots, n)$. Eeldatakse, et toimus r sündmust $(r \leq n)$ ning olgu järjestatud sündmuste toimumise ajad $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Riskigrupp ajamomendil $t_{(j)}$ on defineeritud järgmiselt:

$$R(t_{(j)}) = \{i : t_i \geq t_{(j)}\}.$$

Riski subjektil, kellel toimus sündmus ajahetkel $t_{(j)}$, tähistatakse suurusega $h(t_{(j)}, \mathbf{x}_j)$, kus $\mathbf{x}_j$ on argumenttunnuste $X_1, X_2, \ldots, X_p$ väärtuste vektor selle subjekti korral. Tõepärafunktsioon avaldub kujul:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{h(t_{(j)}, \mathbf{x}_j)}{\sum_{k \in R(t_{(j)})} h(t_{(j)}, \mathbf{x}_k)}.$$

Võrdeliste riskide mudeli $h(t, \mathbf{x}) = h_0(t)\psi(\mathbf{x}, \boldsymbol{\beta})$ kehtides, taandub eelmisest võrdusest

välja baasriskifunktsioon $h_0(t)$. Seega avaldub tõepärafunktsioon järgmiselt:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\psi(\mathbf{x}_j, \boldsymbol{\beta})}{\sum_{k \in R(t_{(j)})} \psi(\mathbf{x}_k, \boldsymbol{\beta})}. \tag{2}$$

Saadud tõepärafunktsioon ei sisalda baasriski $h_0(t)$ ning ei sõltu täpsetest sündmuse toimumise aegadest $t_{(j)}$. Parameetreid $\boldsymbol{\beta}$ saab hinnata tõepärafunktsiooni (2) logaritmimise ning maksimeerimise teel.

2.7 Võrdeliste riskide eelduse kontrollimine

Siinses töös kontrollitakse võrdeliste riskide eeldust kaalutud Schoenfeldi jääkide abil. Olgu n subjekti, kellest r subjektil on vaadeldud huvipakkuv sündmus ning $n-r$ subjekti on paremalt tsenseeritud. Olgu selliste andmete korral Coxi võrdeliste riskide mudelis p seletavat tunnust $X_1, X_2, \ldots, X_p$ ning parameetrite hinnangud $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. Seega i-nda subjekti riskifunktsiooni hinnang avaldub kujul:

$$\hat{h}_i(t) = e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_i} \hat{h}_0(t) = e^{\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}} \hat{h}_0(t),$$

kus $x_{1i}, x_{2i}, \ldots, x_{pi}$ on tunnustele $X_1, X_2, \ldots, X_p$ vastavad väärtused i-ndal subjektil ning $\hat{h}_0(t)$ on baasriskifunktsioon.

Esmalt vaadeldakse tavalisi Schoenfeldi jääke, mis on defineeritud nii, et igale subjektile vastab jääkide hulk, kus iga jääk vastab ühele argumenttunnusele X_j. Schoenfeldi jääk i-nda subjekti ja tunnuse X_j korral on leitav valemist:

$$r_{Pji} = \delta_i \{x_{ji} - \hat{a}_{ji}\}.$$

Valemis tähistab δ_i indikaatortunnust, mille väärtus on 1, kui i-ndal subjektil vaadeldi sündmus, ja 0 vastasel korral (i-nda subjekti elukestus oli tsenseeritud). Suurus $\hat{a}_{ji}$ avaldub kujul:

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_l}}{\sum_{l \in R(t_i)} e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_l}},$$

kus t_i on i-nda subjekti elukestus ning $R(t_i)$ on riskigrupp ajahetkel t_i. Nüüd defineeri-

takse kaalutud Schoenfeldi jäägid. Selleks tähistatakse i-nda subjekti Schoenfeldi jääkide vektorit $\mathbf{r}_{Pi} = (r_{P1i}, r_{P2i}, \ldots, r_{Ppi})'$ ning hinnatud parameetrite $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ kovariatsioonimaatriksit $\text{var}(\hat{\boldsymbol{\beta}})$. Kaalutud Schoenfeldi jäägid on defineeritud järgmiselt:

$$\mathbf{r}^*_{Pi} = r\text{var}(\hat{\boldsymbol{\beta}})\mathbf{r}_{Pi},$$

kus r on arv, mis näitab, kui mitmel subjektil toimus sündmus.

Kaalutud Schoenfeldi jääkide korral kehtib omadus $E(r^*_{Pji}) \approx \beta_j(t_i) - \hat{\beta}_j$, kus $\beta_j(t)$ on tunnuse X_j ajas muutuv kordaja ja $\beta_j(t_i)$ on kordaja väärtus ajahetkel t_i. Võrdeliste riskide eeldust saab tunnuse X_j kohta kontrollida graafikult, kuhu on kantud väärtused $r^*_{Pji} + \hat{\beta}_j$. Kui graafiku punktid asuvad horisontaalsel joonel, siis on tunnuse X_j kordaja ajas konstantne ning seega on selle tunnuse korral võrdeliste riskide eeldus täidetud. Kui graafikule sobitada sirge, siis saab testida, kas sirge tõusu erinevus nullist on statistiliselt oluline. Suure olulisuse tõenäosuse korral on sirge tõus võrdne nulliga ning seega rahuldab võrdeliste riskide eeldust.

3 Ülegenoomne seoseuuring

Ülegenoomse seoseuuringu (*Genome-Wide Association Study*, GWAS) mõiste on tekkinud alles 2007. aastal, kuid praegu on ülegenoomne seoseuuring üheks levinuimaks uuringuliigiks, mida kasutatakse geneetilise epidemioloogia valdkonnas [8]. Ülegenoomsete seoseuuringute põhiliseks eesmärgiks on välja selgitada, millised on erinevate haiguste geneetilised riskifaktorid [9].

Ülegenoomses seoseuuringus analüüsitakse DNA järjestuse erinevusi inimese genoomis ning selle põhjal leitakse huvipakkuva haiguse geneetilised riskifaktorid. Leitud riskitegurite abil on võimalik hinnata indiviidi riski haigestumiseks. Samuti kasutatakse ülegenoomsetest seoseuuringustest saadud geneetilisi riskifaktoreid, et välja töötada uusi meetodeid haiguste ravimiseks ja ennetamiseks. [9]

3.1 Põhimõisted geneetikast

DNA (desoksüribonukleiinhape) ahela moodustavad omavahel liitunud nukleotiidid. Üks nukleotiid koosneb omakorda kolmest ühendist, milleks on lämmastikalus, suhkur (desoksüriboos) ja fosforhappe jääk (fosfaatrühm). Lämmastikalused, mis DNA ehituses esinevad, on adeniin (A), guaniin (G), tümiin (T) ja tsütosiin (C). Lõiku DNA-st, kus on olemas kogu informatsioon ühe valgu moodustamiseks, nimetatakse geeniks. Geeni esinemisvormi nimetatakse alleeliks. Dialleelsuse korral on geenil vaid kaks erinevat esinemisvormi. Kui geen esineb rohkem kui kahel erineval kujul, siis on tegemist polüalleelsusega. [10]

DNA molekuli moodustavad kaks omavahel koos püsivat nukleotiidahelat, mis on keerdunud topeltspiraali kujuliselt. DNA molekulis on alati ühes ahelas oleva adeniini vastas teise ahela tümiin ning guaniini vastas tsütosiin. Kromosoom koosneb DNA molekulist. Kromosoomid asuvad raku tuumas, kus kõrgematel organismidel (nt inimestel) on kromosoomide arv $2n$, kuna igat kromosoomi on dubleeritud. Suguraku tuumas on aga ainult üks kromosoom igast sellisest kromosoomide paarist, kus kromosoomides sisalduvad geenid määravad organismil samu pärilikke tunnuseid. Selle tõttu on suguraku tuumas n kromosoomi. Sellistest kromosoomidest moodustub genoom. [10]

3.2 Ühenukleotiidne polümorfism

Kahe inimese DNA vaheline erinevus on ligikaudu 0,2% [10]. See väike erinevus põhjustab inimeste seas mitmeid erisusi, mille üheks näiteks on risk haigestuda teatud haigusesse [11]. Kõige sagedasemad järjestuse erinevused inimese genoomis on ühenukleotiidsed polümorfismid, mida kasutatakse paljudes uuringutes [12]. Ülegenoomsetes seoseuuringutes otsitakse selliseid ühenukleotiidseid polümorfisme, mis mõjutavad teatud haigusesse haigestumist [11].

Ühenukleotiidne polümorfism (*Single Nucleotide Polymorphism*, SNP) on ühe lämmastikaluse (A, G, T või C) erinevus, mis esineb samas asukohas kahe erineva populatsiooni DNA järjestustes [8]. Täpsemalt on ühenukleotiidne polümorfism genoomi kindlas asukohas selline üksiku nukleotiidi muutus, mida esineb rohkem kui ühel protsendil populatsioonist [12]. Enamasti on ühenukleotiidsed polümorfismid dialleelsed, seejuures esineb üks alleelidest populatsioonis harvemini [11]. Ühenukleotiidse polümorfismi sagedusena mõistetakse populatsioonis vähem esineva alleeli sagedust [9].

3.3 Geneetiline riskiskoor

Geneetilise riskiskoori leidmiseks kombineeritakse ühenukleotiidsete polümorfismide mõjud. Algselt kaasati geneetilisse riskiskoori vaid need ühenukleotiidsed polümorfismid, mille mõju uuritavale haigusele oli leidnud tõestust ülegenoomses seoseuuringus, võttes arvesse ka mitmest testimist ($p < 5 \cdot 10^{-8}$). Tänaseks on aga leitud, et parema prognoosivõimega skoori saamiseks võib geneetilisse riskiskoori lisada märksa rohkem ühenukleotiidseid polümorfisme. [8]

Olgu ühenukleotiidsete polümorfismide arv k ning olgu β_j j-nda ühenukleotiidse polümorfismi mõju suurus. Tähistagu X_j j-nda ühenukleotiidse polümorfismi efektialleelide arvu. Efektialleeli all mõistetakse ühenukleotiidsest polümorfismist ühte alleeli, millel otsitakse seost uuritava haigusega. Ühenukleotiidsete polümorfismide dialleelsuse tõttu on X_j võmalikud väärtused 0, 1 ja 2. [8]

Haiguse esinemise tõenäosust p hinnatakse ülegenoomses seoseuuringus logistilise regressiooni mudeliga kujul:

$$\ln\left(\frac{p}{1-p}\right) = \mu + l(Z) + \beta_j X_j,$$

kus μ on konstant, $l(Z)$ on mittegeneetiliste tegurite mõju ning $j = 1, \ldots, k$. Geneetiline riskiskoor (GRS) saadakse nüüd järgmisest valemist:

$$GRS_k = \sum_{j=1}^{k} \hat{\beta}_j X_j,$$

kus $\hat{\beta}_j$ on parameetri β_j hinnang. Geneetilist riskiskoori nimetatakse ka polügeenseks riskiskooriks (PRS). [8]

4 ROC-kõver

Olgu uuritav tunnus binaarne, mille väärtused tähistavad positiivset ja negatiivset juhtu. Kui subjektil toimus huvipakkuv sündmus, siis on tegemist positiivse juhuga ning vastasel korral negatiivse juhuga. Olgu uuritavat tunnust prognoosival mudelil pidev väljund. Sel juhul fikseeritakse lävend, mille alusel mudeli prognoosid klassifitseeritakse vastavalt uuritava tunnuse klassidesse. [13]

Leitud prognooside jagunemine on esitatud tabelis 1, kus on kasutatud järgmiseid tähistusi:

- tõene positiivne (*true positive*, TP) - õigesti positiivseks prognoositud juhtude arv;
- vale positiivne (*false positive*, FP) - valesti positiivseks prognoositud juhtude arv;
- tõene negatiivne (*true negative*, TN) - õigesti negatiivseks prognoositud juhtude arv;
- vale negatiivne (*false negative*, FN) - valesti negatiivseks prognoositud juhtude arv [14].

Tabel 1: Prognooside jagunemine

Prognoos	Tegelik väärtus	
	Positiivne	Negatiivne
Positiivne	Tõene positiivne (TP)	Vale positiivne (FP)
Negatiivne	Vale negatiivne (FN)	Tõene negatiivne (TN)

Tundlikkus näitab, kui suure osa tegelikest positiivsetest juhtudest mudel prognoosib õigesti [14]. Tundlikkus on leitav järgmisest valemist:

$$\text{tundlikkus} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Spetsiifilisus näitab, kui suure osa tegelikest negatiivsetest juhtudest mudel prognoosib õigesti [14]. Spetsiifilisus avaldub kujul:

$$\text{spetsiifilisus} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

Erinevate lävendite korral saadud tundlikkuste ja spetsiifilisuste graafilisel kujutamisel saadakse ROC-kõver (*receiver operating characteristic curve*), mis iseloomustab mudeli prognoosimise täpsust. Täpsemalt kujutatakse ROC-kõvera y-teljel tundlikkust ning x-teljel suurust $(1 - \text{spetsiifilisus})$. Sellisel graafikul viitab diagonaal läbi punktide (0;0) ja (1;1), et mudel prognoosib juhuslikult. Seega mida kaugemal on ROC-kõver diagonaalist, seda paremini mudel prognoosib. [13]

Prognoosi täpsust kirjeldab ROC-kõvera alla jääv pindala AUC (*area under the curve*). AUC näitab tõenäosust, et mudeli väljund on suurem juhuslikult valitud subjektil, kellel toimus huvipakkuv sündmus (positiivne juht), võrreldes juhuslikult valitud subjektiga, kellel ei toimunud sündmust (negatiivne juht). Eelmine väide kehtib eeldusel, et positiivne juht on seotud suurema mudeli väljundi väärtusega. Maksimaalne AUC väärtus on 1, mille korral mudel ennustab alati õigesti. Seega mida suurem on ROC-kõvera alune pindala, seda parem on prognoosimise täpsus. AUC väärtuse 0,5 korral prognoosib mudel juhuslikult. [13]

5 Andmete analüüs

5.1 Andmete kirjeldus

Käesolevas bakalaureusetöös kasutati andmeid, mis olid saadud Tartu Ülikooli Eesti Geenivaramu (TÜ EGV) andmete linkimisel E-tervise, Eesti Haigekassa ja Eesti Vähiregistri andmetega. Eluskestusanalüüsi läbiviimiseks vaadeldi siin töös vaid indiviide, kellel TÜ Eesti Geenivaramuga liitumise kuupäeval ega enne seda ei olnud diagnoositud jämesoolevähki. Kasutatud TÜ Eesti Geenivaramu andmed sisaldasid üle 48 000 geenidoonori.

E-tervise, Eesti Haigekassa ning Eesti Vähiregistri andmetega linkimisel saadi andmed jämesoolevähi diagnooside kohta. Jämesoolevähina käsitleti haiguseid, mille RHK-10 koodid olid C18, C19, C20 ja C21. See tähendas, et jämesoolevähk oli diagnoositud indiviidil, kellel oli diagnoositud vähemalt üks loetletud pahaoomulistest kasvajatest. Kuna indiviidil sai olla mitu diagnoosi, siis vaadeldi igal indiviidil vaid ühte kõige esimesena saadud diagnoosi ning selle diagnoosimise kuupäeva.

Lisaks kasutati andmeid jämesoolevähi geneetilise riskiskoori kohta. Kasutatud geneetilise riskiskoori andmed koostas Kristi Läll. Jämesoolevähi geneetilised riskiskoorid olid geenidoonoritel leitud haiguste C18, C19 ja C20 põhjal.

5.2 Kasutatatud tunnused

Elukestusanalüüsi läbiviimiseks kasutati töös järgmisi tunnuseid: vanus TÜ Eesti Geenivaramuga liitumisel, TÜ Eesti Geenivaramuga liitumise kuupäev, surma kuupäev ning jämesoolevähi diagnoosimise kuupäev. Kasutatud mittegeneetilised tunnused olid sugu, kehamassiindeks, diagnoositud haigused ja veel mitmed tunnused, mis olid seotud toitumisega, liikumisega, suitsetamisega ja alkoholi tarbimisega. Kehamassiindeks (KMI) arvutati valemist:

$$\text{KMI} = \frac{\text{kehakaal (kg)}}{\text{pikkus}^2\ \text{(m)}}.$$

Toitumisega seotud tunnustest näitasid kohvi, tee, leiva ja saia tarbimist kirjeldavad tunnused, mitu tassi või viilu päevas tarbitakse. Ülejäänud toitumisega seotud tunnused näitasid erinevate toiduainete tarbimist nädalas (1: ei tarbi üldse, 2: tarbib 1–2 päeval,

3: tarbib 3–5 päeval, 4: tarbib 6–7 päeval). Andmetes oli selliseid tunnuseid järgmiste toiduainete tarbimise kohta: kartul, riis/makaronid, puder/müsli, piimatooted, kala, liha, lihaprodukti (vorstid/viinerid), värske juurvili, keedetud juurvili, värsked puuviljad/marjad, kompotid/keedised, maiustused, karastusjoogid ning munad.

Liikumisega seotud tunnused näitasid, mitu tundi nädalas indiviid kindla tegevusega tegeles. Nendeks tunnusteks olid jalutamine, mõõduka kiirusega kõnd, kiire kõnd ning trenni tegemine. Suitsetamist puudutavad tunnused olid järgmised: suitsetab (0: ei, 1: jah), on kunagi suitsetanud (0: ei, 1: jah), suitsetatud aastate arv ning suitsetamise ühik, mis näitas päevas suitsetatud sigarettide arvu (või muud tubakatoodet samaväärses koguses). Alkoholi tarbimist näitavaks tunnuseks oli alkoholiühik, mis näitas kümne grammi puhta alkoholi tarbimist ühes päevas.

Kasutatud tunnustel muudeti ilmselgelt vigased väärtused puuduvaks (nt juhul, kui jalutamist oli märgitud 140 tundi nädalas). Puuduvad väärtused imputeeriti mediaaniga. Seejuures binaarsetel tunnustel ei olnud puuduvaid väärtusi ning üldiselt oli tunnustel puuduvaid väärtusi alla 0,5%. Pidevatel tunnustel esines veel mõningaid suuri väärtusi, mis tekitasid kahtlust. Seega leiti nendel tunnustel 99% kvantiil, millega asendati tunnuse väärtused, mis ületasid seda piiri.

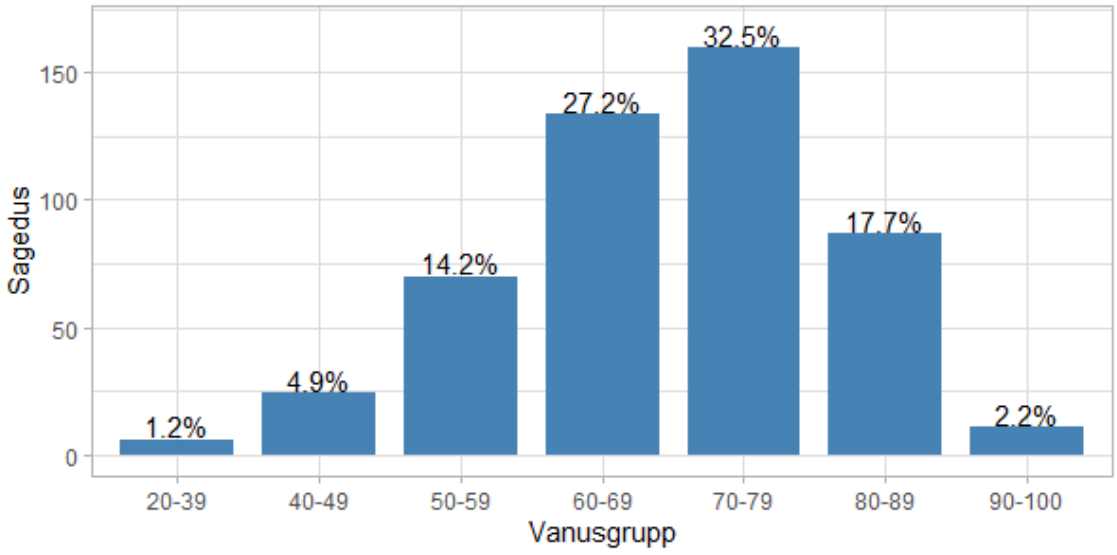
5.3 Kirjeldav statistika

Andmestikus oli geenidoonoreid kokku 48 545. Jämesoolevähi diagnoose oli andmestikus 492 ehk ligikaudu 1% geenidoonoritest oli diagnoositud jämesoolevähk. Jämesoolevähi diagnooside hulgas oli kõige rohkem C18 diagnoose, mida oli kokku 331 geenidoonoril. Seega moodustas jämesoolevähi diagnoosidest ligikaudu kaks kolmandikku C18 diagnoosid. Kõige vähem oli C21 diagnoose, mis moodustasid ligikaudu 2% jämesoolevähi diagnoosidest. Täpsem diagnooside jagunemine on esitatud tabelis 2.

Tabel 2: Haiguste C18–C21 sagedus TÜ Eesti Geenivaramu kohordis

	Diagnoos			
	C18	C19	C20	C21
Sagedus	331 (67,28%)	41 (8,33%)	110 (22,36%)	10 (2,03%)

Keskmine uuringuga liitumise vanus oli ligikaudu 44,6 aastat ning keskmiselt olid subjektid vaatluse all 10,3 aastat. Jämesoolevähki diagnoositi keskmiselt vanuses 69,7 aastat. Joonisel 1 on esitatud jämesoolevähi diagnooside arv erinevates vanusgruppides koos osakaaludega, mis näitavad kui suure osa need juhud moodustavad kõikidest jämesoolevähi diagnoosidest. Esitatud jooniselt on näha, et kõige rohkem diagnoositi jämesoolevähki vanusevahemikus 70–79 aastat ning 60–69 aastat. Vanusgrupis 60–79 eluaastat diagnoositi üle poole jämesoolevähi diagnoosidest ehk täpsemalt 59,7%. Vanusevahemikus 50–89 oli diagnoositud ligikaudu 91,6% andmestikus olevatest jämesoolevähi juhtudest.



Joonis 1: Jämesoolevähi diagnooside jagunemine vanusgruppidesse TÜ Eesti Geenivaramu kohordis

Naisi oli andmestikus ligikaudu 66,5% ning mehi 33,5%. Keskmine kehamassiindeks oli 26,3, mis viitas ülekaalule (normaalkaalu korral on kehamassiindeks vahemikus 18,5 kuni 25). Normaalkaalus oli 44,6% andmestikus olevatest geenidoonoritest. Suitsetajaid oli andmestikus 28,7% ning elu jooksul olid suitsetanud 42,4% geenidoonoritest. Tabelis 3 on

esitatud mõningate elustiiliga seotud tunnuste keskmised koos standardhälvete ja mediaanidega. Tabelist 3 on näha, et keskmiselt olid geenidoonorid suitsetanud üle kuue aasta ning päevas suitsetati keskmiselt umbes kolm sigaretti. Keskmine suitsetatud aastate arv vaid nende hulgas, kes olid kunagi suitsetanud, oli ligikaudu 16 aastat. Päevas suitsetatud sigarettide arv oli suitsetajate seas ligikaudu 11.

Tabel 3: Liikumisaktiivsusega, suitsetamisega ja alkoholi tarbmisega seotud tunnuste näitajad TÜ Eesti Geenivaramu kohordis

Tunnus	Keskmine	Standardhälve	Mediaan
Jalutamine (tunde nädalas)	4,22	4,63	3
Mõõdukas tempos kõndimine (tunde nädalas)	5,35	6,85	3
Kiires tempos jalutamine (tunde nädalas)	2,03	3,45	0
Trenni tegemine (tunde nädalas)	1,59	2,76	0
Suitsetatud aastate arv	6,66	11,24	0
Suitsetamise ühik (sigarettide arv päevas)	3,25	6,35	0
Alkoholiühik (10 g puhast alkoholi päevas)	0,35	0,70	0,1

Tabeli 3 põhjal jalutasid geenidoonorid nädalas keskmiselt veidi üle nelja tunni, kõndisid mõõdukas tempos natuke üle viie tunni ning kõndisid kiires tempos umbes kaks tundi. Trenni tehti nädalas keskmiselt üle ühe tunni, kuid tunnuse mediaan oli 0 ehk vähemalt pooled geenidoonoritest ei teinud trenni. Täpsemalt tegid trenni 37,4% geenidoonoritest, kes treenisid nädalas keskmiselt 4,3 tundi. Lisaks on tabelist 3 näha, et puhast alkoholi tarbiti päevas keskmiselt 3,5 grammi, mis on umbes üks alkoholiühik ehk 32 ml 40% kanget alkoholi kolme päeva jooksul.

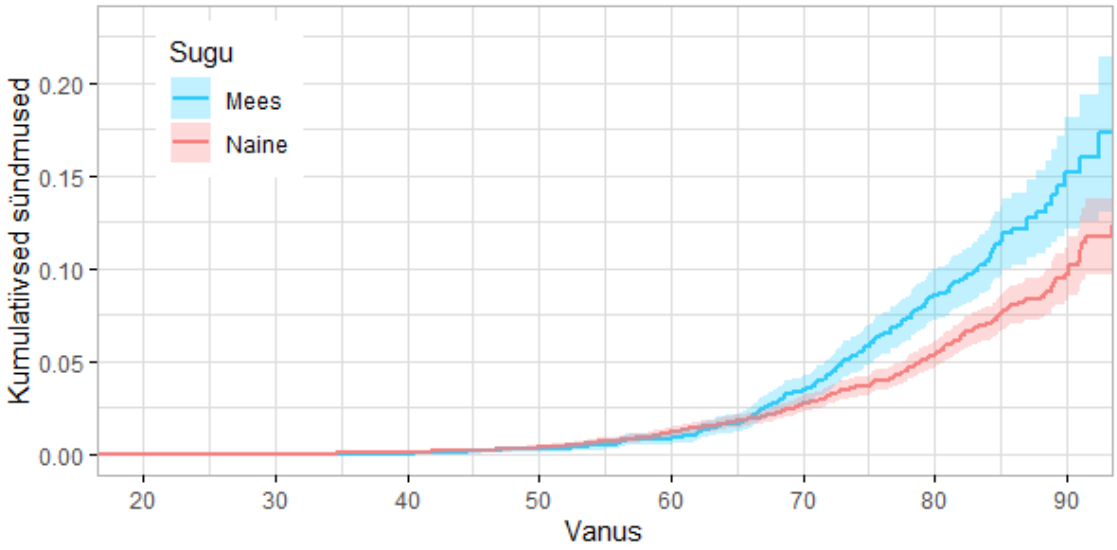
5.4 Jämesoolevähi riskitegurid

5.4.1 Mittegeneetilised riskitegurid

Esmalt vaadeldi jämesoolevähki haigestumist meestel ja naistel. Joonisel 2 on kujutatud hinnang elukestuse jaotusfunktsioonile ehk $1 - \hat{S}(t)$ sugude lõikes koos 95% usaldusintervalliga. Jaotusfunktsiooni hinnangu leidmisel kasutati Kaplan-Meieri hinnangut üle-

elamisfunktsioonile. Seejuures oli ajaskaalana kasutatud vanust ning arvestatud vasakult tõkestatusega ja paremalt tsenseeritusega.

Graafiku tegemiseks kasutati R-i paketist *survival* funktsioone *survfit* ja *Surv*. Vasakult tõkestatuse arvestamiseks anti funktsioonile *Surv* ette kolm parameetrit kujul *Surv(time, time2, event)*. Viimases tähistab *time* ajamomenti, kus subjekt liitus uuringuga, ja *time2* tähistab ajamomenti, kus subjekt lahkus uuringust. Kohal *event* on indikaatortunnus, mille väärtus on 1, kui subjektil vaadeldi huvipakkuv sündmus, ning 0 vastasel korral (subjekti elukestus oli paremalt tsenseeritud). [15]



Joonis 2: Hinnang elukestuse jaotusfunktsioonile sugude lõikes TÜ Eesti Geenivaramu kohordis

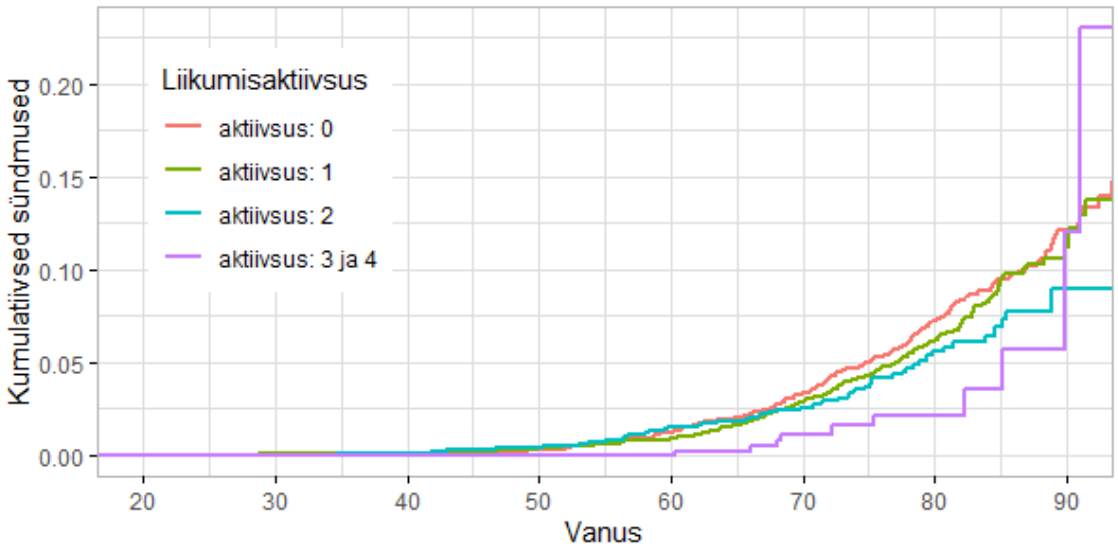
Jooniselt 2 on näha, et tõenäosus saada jämesoolevähi diagnoos enne teatud vanust hakkas meestel ja naistel erinema umbes 70. eluaastast. Jooniselt 2 selgub ka, et meestel oli suurem tõenäosus haigestuda jämesoolevähki kui naistel. Näiteks oli tõenäosus saada jämesoolevähk enne 80. eluaastat naistel hinnanguliselt 5,3% ning meestel hinnanguliselt 8,6%.

Liikumisharjumuste mõju uurimiseks defineeriti liikumise tunnuseid kombineeriv tunnus „aktiivsus“. Tunnus oli defineeritud nii, et selle väärtus suurenes iga kord ühe võrra järgmistel juhtudel:

- vähemalt kaks tundi nädalas tehti trenni;
- kõnniti kiires tempos rohkem kui kolm tundi nädalas;

- kõnniti mõõduka kiirusega vähemalt seitse tundi nädalas;
- jalutati rohkem kui üheksa tundi nädalas.

Seega olid tunnuse väärtused 0, 1, 2, 3 ja 4, kus 0 korral ei teinud indiviid ühtegi eelnevalt nimetatud tegevustest ning 4 korral tegi kõiki. Väärtustele vastavate juhtude arv oli järgmine: väärtust 0 esines 17889 korral, väärtust 1 esines 19294 korral, väärtust 2 esines 8248 korral, väärtust 3 esines 2613 korral ning väärtust 4 esines 501 korral. Joonisel 3 on esitatud hinnang elukestuse jaotusfunktsioonile aktiivsuse tasemete lõikes. Elukestuse jaotusfunktsiooni hinnangu leidmiseks kasutati jällegi Kaplan-Meieri hinnangut üleelamisfunktsioonile.



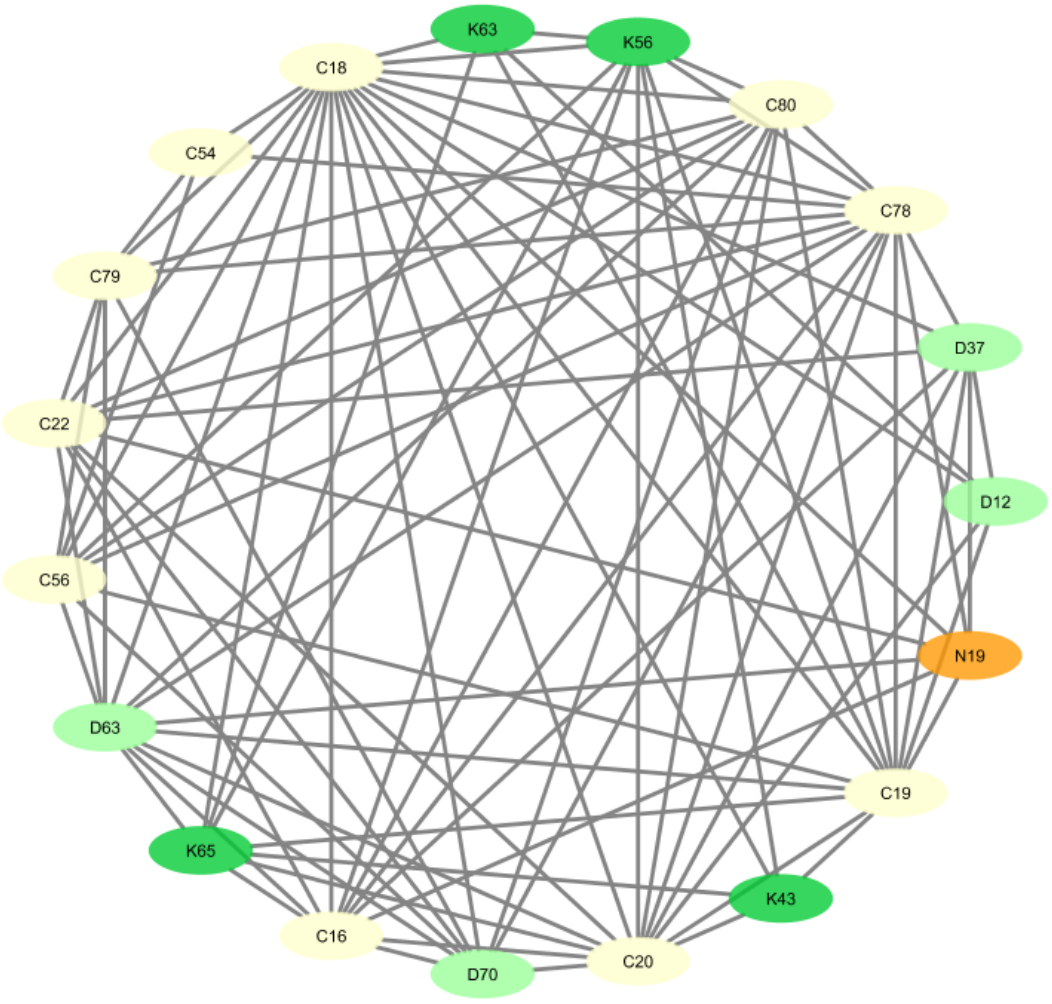
Joonis 3: Hinnang elukestuse jaotusfunktsioonile liikumisaktiivsuse tasemete lõikes TÜ Eesti Geenivaramu kohordis

Jooniselt 3 on näha, et vähem aktiivsetel inimestel diagnoositi jämesoolevähki varem. Tõenäosus saada jämesoolevähi diagnoos enne 80. eluaastat oli kõige vähem aktiivsetel hinnanguliselt 7,2%, aktiivsuse taseme 1 korral hinnanguliselt 6,1%, aktiivsuse taseme 2 korral hinnanguliselt 5,6% ning kõige aktiivsematel (väärtuste 3 ja 4 korral) hinnanguliselt 2,1%.

Selleks, et täpsemalt leida jämesoolevähki prognoosivad mittegeneetilised tunnused, moodustati Coxi võrdeliste riskide mudel. Mudelisse prooviti lisada kõiki tunnuseid, mida nimetati peatükis 5.2. Mudeli tegemiseks kasutati R-i paketist *survival* funktsioone *Surv*

ja *coxph*. Mudelis kasutati ajaskaalana vanust ning mudeli tegemisel arvestati paremalt tsenseeritusega ja vasakult tõkestatusega.

Liikumisega seotud tunnuseid prooviti mudelisse lisada nii eraldi kui ka neid kõiki kombineeriva tunnusena. Haiguste diagnoosidest vaadeldi täpsemalt esimest tüüpi diabeeti (E10), teist tüüpi diabeeti (E11) ning mittenakkuslikke peen- ja jämesoolepõletikke: Crohni tõbi (K50), haavandiline jämesoolepõletik (K51) ning muud mittenakkuslikud maopeensoolepõletikud ja koliidid (K52). Lisaks võeti uurimise alla haigused, mis olid seotud kõige sagedasema diagnoosiga ehk haigusega C18. Joonisel 4 on esitatud C18-ga seotud haigusi sisaldav komorbiidsusvõrgustik, mille koostas juhendaja Jaanika Kronberg.



Joonis 4: Komorbiidsusvõrgustik C18-ga seotud haigustest (autor: Jaanika Kronberg)

Joonisel 4 kujutatud haigustest valiti mudeli tegemiseks välja kõik C18-ga seotud haigused, mis ei olnud pahaloomulised kasvajad. Seega jäeti välja kõik haigused, mille RHK-10 kood

algas tähega „C“. Samuti jäeti välja haigus D63, millel ei olnud andmestikus ühtegi juhtu. Mudelisse prooviti lisada kõiki alles jäänud haiguseid, milleks olid D12, D37, D70, K43, K56, K63, K65 ja N19.

Jämesoolevähki prognoosivate tunnuste leidmiseks lisati Coxi võrdeliste riskide mudelisse tunnuseid ükshaaval. Iga lisamise järel kontrolliti, kas on ebaolulisi tunnuseid, mida tuleks eemaldada. Tulemused on tabelis 4, kus on esitatud jämesoolevähki mõjutavad mittegeneetilised tunnused koos riskide suhte, riskide suhte 95% usaldusintervalliga ning olulisuse tõenäosusega.

Tabel 4: Jämesoolevähi mittegeneetilised riskitegurid TÜ Eesti Geenivaramu andmete põhjal

Tunnus X_i	$e^{\hat{\beta}_i}$	95% usaldusintervall	p-väärtus
Sugu	0,68	(0,57; 0,82)	$4{,}2{\cdot}10^{-5}$
E10	1,86	(0,99; 3,49)	0,052
K50	9,39	(3,01; 29,31)	0,00011
D12	3,77	(1,68; 8,46)	0,0013
Tee joomine	1,09	(1,02; 1,17)	0,017
Piimatooted	0,57	(0,38; 0,86)	0,0070
Aktiivsus	0,84	(0,75; 0,93)	0,0015

Tabelist 4 on näha, et esimest tüüpi diabeedi (E10) olulisuse tõenäosus oli natuke üle olulisuse nivoo $\alpha = 0,05$. Kuna aga nende vahe ei olnud väga suur ning esimest tüüpi diabeet oli üheks teadaolevaks riskiteguriks, siis otsustati see ikkagi mudelisse sisse jätta. Seega analüüsist selgus, et jämesoolevähki mõjutavad mittegeneetilised tegurid olid sugu, tee joomine, piimatoodete tarbimine, liikumisaktiivsus ning haigused E10, K50 ja D12. Suitsetamise kui ühe teadaoleva jämesoolevähi riskifaktori mõju ei õnnestunud ühegi suitsetamisega seotud tunnuse korral näidata. Haigused E10, K50 ja D12 tähistasid järgmisi haigusi:

- E10 - insuliinisõltuv suhkurtõbi (esimest tüüpi diabeet);
- K50 - Crohni tõbi (regionaalne ehk segmentaarne peensoolepõletik);
- D12 - käärsoole, pärasoole, päraku ja pärakukanali healoomuline kasvaja [4].

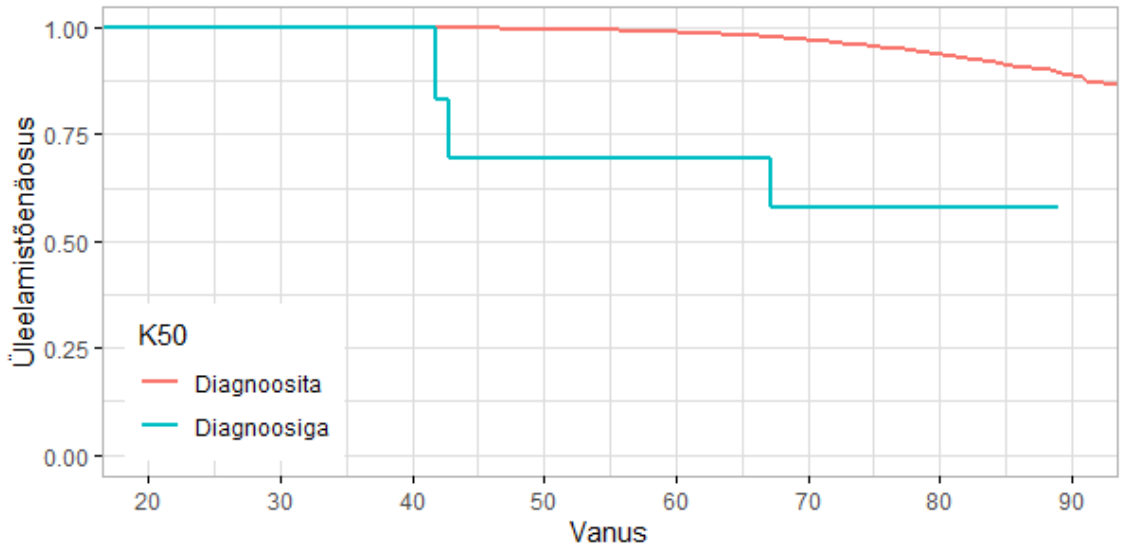
Faktortunnustel oli baastasemeks soo korral mees, haiguste E10, K50, D12 korral diagnoosi puudumine ning piimatoodete korral piimatoodete mitte tarbimine. Tee joomist ja aktiivsust käsitleti pidevate tunnustena. Tabelist 4 on näha, et jämesoolevähi riski suurendas esimest tüüpi diabeet, Crohni tõbi, käärsoole, pärasoole, päraku ja pärakukanali healoomuline kasvaja ning tee joomine. Jämesoolevähi riski aga vähendasid piimatoodete tarbimine ning liikumisaktiivsus. Samuti oli naistel väiksem risk jämesoolevähki haigestumiseks kui meestel.

Selgus, et naistel oli hinnanguliselt $\frac{1}{0,68}$ ehk 1,5 korda väiksem risk kui meestel. Aktiivsuse ühe taseme erinevuse korral oli rohkem aktiivsetel $\frac{1}{0,84}$ ehk 1,2 korda väiksem jämesoolevähi risk. Võrreldes kõige vähem aktiivsetega (aktiivsuse väärtus 0) oli kõige aktiivsematel (aktiivsuse väärtus 4) $\frac{1}{0,84^4}$ ehk ligikaudu 2,0 korda väiksem risk jämesoolevähki haigestumiseks. Samuti leiti, et piimatoodete tarbimine vähendas riski ligikaudu $\frac{1}{0,57}$ ehk 1,7 korda võrreldes piimatoodete mittetarbimisega. Mitmetes uuringutes on samuti leitud piimatoodete tarbimise mõju jämesoolevähile [16]. Arvatakse, et jämesoolevähi riski võib vähendada näiteks piimatoodetes sisalduv kaltsium [16].

Tulemuste põhjal oli aga jämesoolevähi risk seda suurem, mida rohkem igapäevaselt joodi teed. Näiteks oli päevas ühe tassi tee joomise korral 1,1 korda suurem risk kui tee mitte joomise korral ning kahe tassi tee joomise korral $1{,}1^2$ ehk ligikaudu 1,2 korda suurem risk kui tee mitte joomise korral. Antud seose korral ei leitud uuringuid, mis võiksid tulemust kinnitada. Seega oli tegemist pigem juhusliku leiuga. Esimest tüüpi diabeeti (E10) põdevatel inimestel oli jämesoolevähi risk 1,9 korda suurem kui esimest tüüpi diabeedita inimestel. Nendel, kellel oli diagnoositud käärsoole, pärasoole, päraku ja pärakukanali healoomuline kasvaja (D12), oli risk 3,8 korda suurem kui D12 diagnoosita inimestel. Crohni tõbe (K50) põdenud inimestel oli hinnanguliselt 9,4 korda suurem risk võrreldes seda haigust mitte põdenud inimestega.

Joonisel 5 on esitatud Kaplan-Meieri hinnangud üleelamisfunktsioonile Crohni tõve diagnoosiga ja diagnoosita inimestel. Esitatud Kaplan-Meieri kõveralt, kus on kujutatud Crohni tõbe põdenud geenidoonorite üleelamisfunktsiooni hinnang, on astmete arvu järgi näha, et Crohni tõbe põdenud geenidoonoritest oli jämesoolevähi diagnoosi saanud kolm geenidoonorit. Crohni tõbe oli kokku diagnoositud 29 geenidoonoril. Väheste juhtude arv selgitab ka riskide suhte laia 95% usaldusintervalli (3,0; 29,3) (vt tabel 4). Joonise 5 põh-

jal oli tõenäosus, et jämesoolevähki ei diagnoosita enne 80. eluaastat, Crohni tõbe mitte põdenud inimestel hinnanguliselt 93,7% ning Crohni tõbe põdenud inimestel hinnanguliselt 57,9%.



Joonis 5: Kaplan-Meieri hinnang üleelamisfunktsioonile K50 diagnooside lõikes TÜ Eesti Geenivaramu kohordis

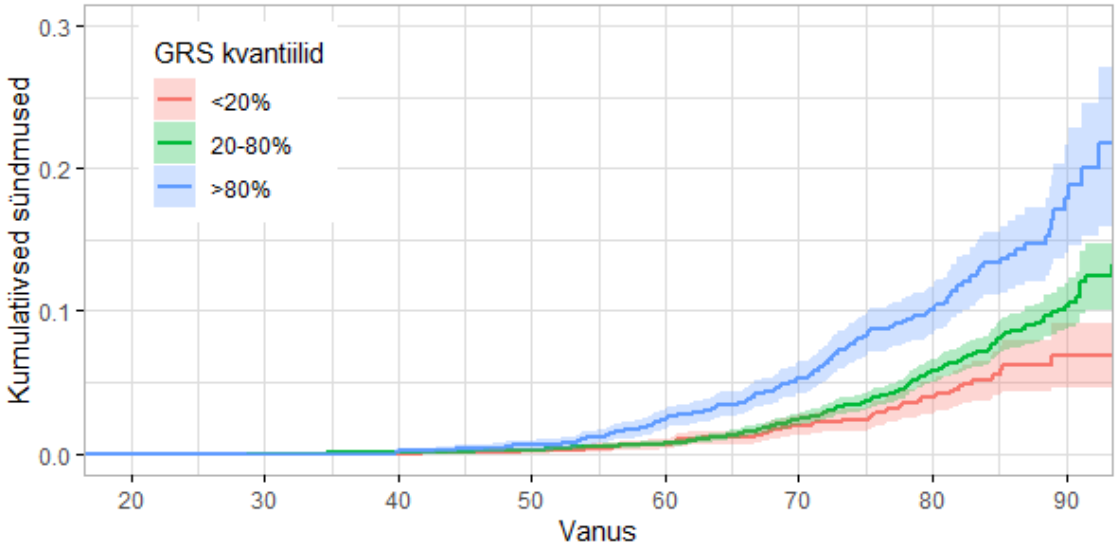
Mudeli võrdeliste riskide eelduse kontrollimiseks kasutati R-i paketist *survival* funktsiooni *cox.zph*, mis kontrollib eelduse täidetust kaalutud Schoenfeldi jääkide abil [15]. Testi tulemused on esitatud tabelis 5. Kõikide tunnuste korral oli olulisuse tõenäosus suurem, kui olulisuse nivoo ($\alpha = 0,05$). Seega oli kõikide tunnuste korral täidetud võrdeliste riskide eeldus ehk riskide suhe oli iga tunnuse korral ajas konstantne.

Tabel 5: Võrdeliste riskide eelduse kontrollimisel saadud olulisuse tõenäosused

Tunnus X_i	p-väärtus
Sugu	0,14
E10	0,76
K50	0,062
D12	0,62
Tee joomine	0,38
Piimatooted	0,48
Aktiivsus	0,30

5.4.2 Geneetiline riskiskoor

Joonisel 6 on esitatud hinnang elukestuse jaotusfunktsioonile geneetilise riskiskoori tasemete lõikes koos 95% usaldusintervalliga. Täpsemalt vaadeldi madalat, keskmist ja kõrget geneetilist riskiskoori. Madala geneetilise riskiskooriga olid need geenidoonorid, kellel geneetiline riskiskoor oli madalaima 20% hulgas. Kõrge geneetilise riskiskooriga olid aga need geenidoonorid, kellel geneetiline riskiskoor oli kõrgeima 20% seas. Ülejäänutel geenidoonoritel oli seega keskmine geneetiline riskiskoor.



Joonis 6: Hinnang elukestuse jaotusfunktsioonile geneetilise riskiskoori tasemete lõikes TÜ Eesti Geenivaramu kohordis

Jooniselt 6 selgub, et tõenäosus jämesoolevähi diagnoosi saamiseks enne teatud vanust hakkas kõrge geneetilise riskiga inimestel teistest erinema umbes 55. eluaastast. Samuti on näha, et kõrgema geneetilise riski korral oli ka suurem tõenäosus haigestuda jämesoolevähki. Näiteks oli tõenäosus saada jämesoolevähi diagnoos enne 80. eluaastat madala geneetilise riski korral hinnanguliselt 3,9%, keskmise geneetilise riski korral hinnanguliselt 5,8% ning kõrge geneetilise riski korral hinnanguliselt 10,0%

Geneetilise riskiskoori mõju täpsemaks uurimiseks tehti Coxi võrdeliste riskide mudel kasutades R-i paketti *survival* ning selle funktsioone *Surv* ja *coxph*. Ajaskaalana kasutati jällegi vanust ning arvestati paremalt tsenseeritud ja vasakult tõkestatud andmetega. Analüüsi tulemused on esitatud tabelis 6.

Tabel 6: Geneetilise riskiskoori mõju jämesoolevähile

Tunnus X_i	$e^{\hat{\beta}_i}$	95% usaldusintervall	p-väärtus
GRS	5,31	(3,78; 7,47)	$< 2 \cdot 10^{-16}$

Analüüsi tulemustest selgus, et geneetilisel riskiskooril oli oluline mõju jämesoolevähile. Kõrgema geneetilise riskiskoori korral oli ka suurem risk haigestuda jämesoolevähki. Kui võrrelda ühe ühiku võrra erinevate geneetiliste riskiskooridega inimesi, siis on suurema geneetilise riskiga inimesel 5,3 korda suurem risk haigestuda jämesoolevähki.

Ka selle mudeli korral kontrolliti võrdeliste riskide eeldust. Saadud olulisuse tõenäosus oli 0,32. Seega oli võrdeliste riskide eeldus geneetilise riskiskoori korral täidetud, mis tähendab, et riskide suhe oli ajas konstantne.

5.5 Jämesoolevähi riski prognoosivad mudelid

Prognostiliste skooride välja töötamiseks jagati andmed juhuslikult kaheks: treeningandmestikuks ja testandmestikuks. Treeningandmestikku võeti juhuslikult 80% andmetest ning testandmestikku ülejäänud 20%. Sellisel jagamisel jäi treeningandmestikku 38836 vaatlust, mille hulgas oli 394 jämesoolevähi diagnoosi, ning testandmestikku jäi 9709 vaatlust, mille hulgas oli 98 jämesoolevähi diagnoosi.

Treeningandmete pealt tehti kolm Coxi võrdeliste riskide mudelit: vaid mittegeneetiliste riskiteguritega mudel, vaid geneetilise riskiskooriga mudel ja mudel, kus olid mittegeneetilised riskitegurid koos geneetilise riskiskooriga. Mudelite tegemiseks kasutati jällegi R-i paketist *survival* funktsioone *Surv* ja *coxph*. Samuti kasutati mudelites ajaskaalana vanust ning arvestati vasakult tõkestatusega ja paremalt tsenseeritusega.

Lisas 1 on esitatud treeningandmete põhjal saadud ainult mittegeneetitlisi tegureid sisaldav mudel (vt tabel 11) ning vaid geneetilist riskiskoori sisaldav mudel (vt tabel 12). Tabelis 7 on treeningandmetelt saadud mudel, kuhu lisati nii eelnevalt leitud jämesoolevähki mõjutavad mittegeneetilised tunnused kui ka geneetiline riskiskoor.

Tabel 7: Mittegeneetilisi riskitegureid ja geneetilist riskiskoori sisaldav mudel jämesoolevähi prognoosimiseks

Tunnus X_i	$e^{\hat{\beta}_i}$	95% usaldusintervall	p-väärtus
Sugu	0,70	(0,57; 0,85)	0,00047
E10	2,16	(1,12; 4,20)	0,022
K50	16,13	(5,16; 50,45)	$1{,}8{\cdot}10^{-6}$
D12	3,29	(1,36; 7,98)	0,0085
Tee joomine	1,10	(1,02; 1,19)	0,013
Piimatooted	0,63	(0,39; 1,02)	0,058
Aktiivsus	0,87	(0,77; 0,98)	0,025
GRS	5,20	(3,55; 7,61)	$< 2 \cdot 10^{-16}$

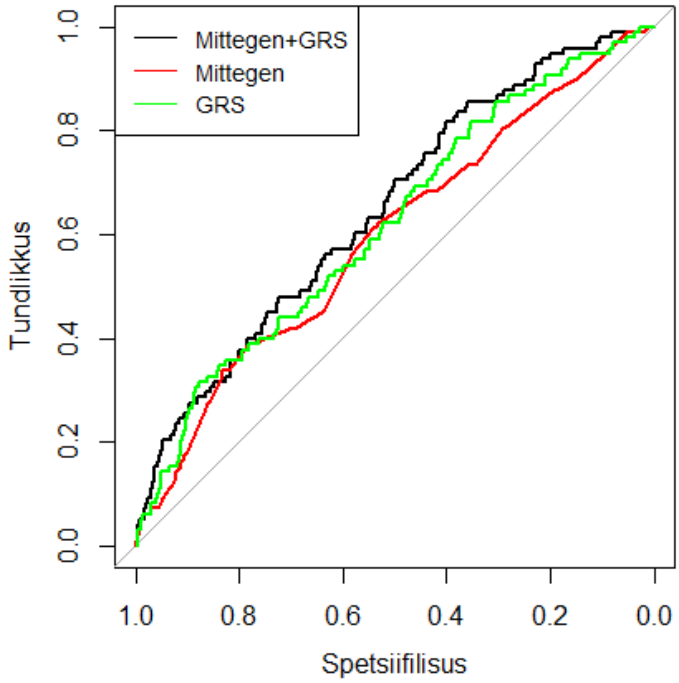
Testandmestikus leiti kõigi kolme mudeli korral prognostilised skoorid. Seejärel vaadeldi, kas prognostilised skoorid annavad statistiliselt olulise seose jämesoolevähiga. Selleks tehti kõikide mudelite prognoosidega Coxi võrdeliste riskide mudel. Tabelis 8 on esitatud analüüsist saadud prognostiliste skooride olulisuse tõenäosused iga mudeli korral. Tulemustest selgus, et prognostilistel skooridel oli iga mudeli korral statistiliselt oluline seos jämesoolevähiga. Kõige väiksem olulisuse tõenäosus oli riskiskooridel, mis olid leitud nii mittegeneetilisi tegureid kui ka geneetilist riskiskoori sisaldavast mudelist.

Tabel 8: Prognostiliste skooride olulisuse tõenäosused

Prognostilistele skooridele vastav mudel	p-väärtus
Mittegeneetilised tegurid + GRS	$4{,}2{\cdot}10^{-9}$
Mittegeneetilised tegurid	0,00032
GRS	$6{,}3{\cdot}10^{-6}$

5.5.1 Mudelite prognoosimise täpsused

Mudelite prognoosimise täpsuste uurimiseks ja võrdlemiseks vaadeldi vastavaid ROC-kõveraid. ROC-kõverate leidmiseks kasutati R-i paketist *pROC* funktsiooni *roc*. ROC-kõverad leiti kõigi kolme mudeli korral ning tulemused on kujutatud joonisel 7.



Joonis 7: ROC-kõverad

Jooniselt 7 on näha, et kõik ROC-kõverad on diagonaalist kõrgemal. Sellest saab järeldada, et mudelid ei prognoosinud juhuslikult. Kõige suurem ROC-kõvera alune pindala oli mittegeneetilisi tegureid ja geneetilist riskiskoori sisaldava mudeli korral, mis tähendab, et selle mudeli prognoosimise täpsus oli kõige suurem. ROC-kõverate alla jäävate pindalade suurused on esitatud tabelis 9.

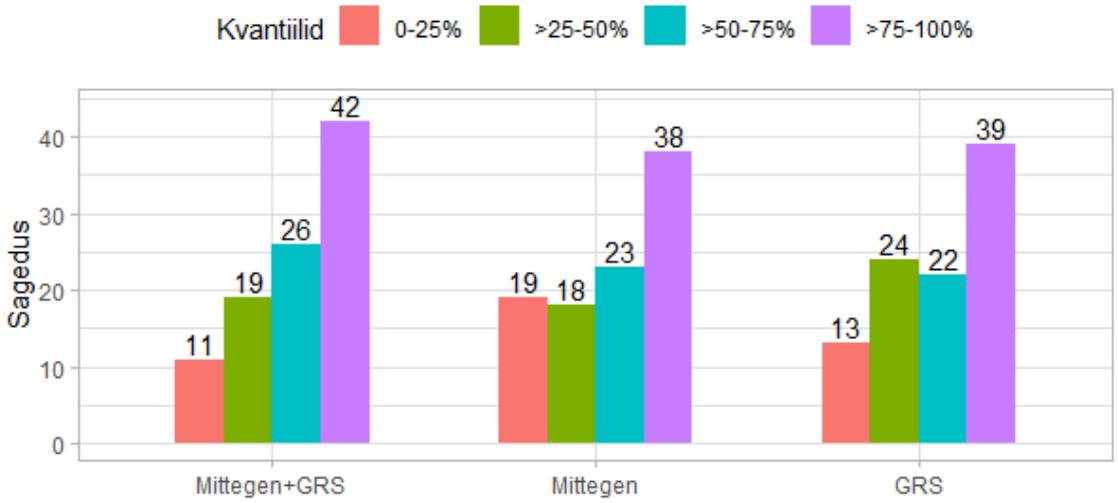
Tabel 9: AUC väärtused

Mudel	AUC
Mittegeneetilised tegurid + GRS	0,648
Mittegeneetilised tegurid	0,597
GRS	0,619

Mittegeneetilisi tegureid ja geneetilist riskiskoori sisaldava mudeli korral oli AUC väärtus ligikaudu 0,648. See tähendab, et tõenäosusega 64,8% oli juhuslikult valitud jämesoolevähi diagnoosiga indiviidil kõrgem skoor kui juhuslikult valitud tervel inimesel. Vaid mittegeneetilisi tunnuseid sisaldava mudeli korral oli AUC väärtus ligikaudu 0,597 ning

vaid geneetilist riskiskoori sisaldava mudeli korral 0,619. Seega geneetilise riskiskoori lisamine mittegeneetilisi tegureid sisaldavasse mudelisse parandas prognoosimise täpsust ümardatult 5%.

Seejärel jagati testandmestikus olevad geenidoonorid prognostiliste skooride suuruse järgi nelja riskirühma nii, et igas rühmas oli võrdne arv inimesi. Seega esimesse rühma kuulusid 25% madalaima skooriga geenidoonorid ning viimasesse rühma kuulusid 25% kõrgeima skooriga geenidoonorid. Joonisel 8 on esitatud jämesoolevähi diagnoosi saanute jagunemine nimetatud nelja gruppi prognostiliste skooride järgi.



Joonis 8: Jämesoolevähi haigete jagunemine riskigruppidesse erinevate mudelite prognostiliste skooride põhjal

Jooniselt 8 on näha, et jämesoolevähiga inimestest oli iga mudeli korral kõige rohkem inimesi viimases riskigrupis ehk kõige kõrgema 25% skooriga inimeste hulgas. Võrreldes teiste mudelitega oli mittegeneetilisi tunnuseid ja geneetilist riskiskoori sisaldava mudeli korral kõige vähem haigeid liigitatud madalaimasse riskigruppi. Samuti oli selle mudeli korral kõrge riskiga haigete arv kõige suurem. Mittegeneetilisi tunnuseid ja geneetilist riski sisaldava mudeli korral oli kõrge riskiga gruppi prognoositud inimeste arv peaaegu neli korda suurem kui madala skoori saanud jämesoolevähi haigete arv.

Lisaks uuriti, kas geneetiline riskiskoor, mis oli tehtud haiguste C18–C20 põhjal, prognoosis haigust C21. Vaid geneetilist riskiskoori sisaldava mudeli põhjal selgus, et C21 diagnoosiga inimeste seast sattus üks indiviid kõige kõrgema riskiga gruppi ning kah-

te kõige madalama riskiga gruppi sattus mõlemasse samuti üks indiviid. Kokku oli C21 diagnoose testandmestikus kolm. Seega ei saanud juhtude vähesuse tõttu kindlaks teha, kas kasutatud geneetiline riskiskoor prognoosis ka haigust C21.

Edasi vaadeldi täpsemalt, kuidas geneetilise riskiskoori lisamine mittegeneetilisi tunnuseid sisaldavasse mudelisse muutis jämesoolevähi haigete jagunemist riskigruppidesse. Tulemused on esitatud tabelis 10, kust selgus, et geneetilise riskiskoori lisamisel oli varasemalt kõige madalama riskiga 19-st jämesoolevähiga inimesest liigitatud kõrgema riskiga gruppi 13 inimest. Varasemalt kõige kõrgema riskiga grupis olnud 38-st inimesest jäid 25 inimest samasse riskirühma ning ülejäänud liigitati madalama riskiga rühmadesse. Keskmiste riskirühmade korral jäid 17 inimest samasse riskirühma, seitse inimest paigutati varasemast madalama riskiga rühma ning 17 inimest paigutati varasemast kõrgema riskiga rühma.

Tabel 10: Jämesoolevähi haigete riskigruppide muutus geneetilise riskiskoori lisamisel mittegeneetiliste teguritega mudelisse

Mittegeneetilised tegurid	Mittegeneetilised tegurid+GRS			
	0–25%	>25–50%	>50–75%	>75–100%
0–25%	6	3	7	3
>25–50%	0	10	3	5
>50–75%	3	4	7	9
>75–100%	2	2	9	25

Seega mõlema mudeli korral liigitati jämesoolevähiga diagnoositud 98-st inimesest 48 haiget samasse riskirühma. Ülejäänute hulgast paranes prognoos 30-l haigel, kellest 17 paigutati kõige kõrgema riskiga gruppi. Samas paigutati varasemast madalama riskiga gruppi 20 haiget, kellest viis liigitati kõige madalama riskiga gruppi. Kokkuvõttes muutis geneetilise riskiskoori mudelisse lisamine jämesoolevähi haigete prognoose veidi paremaks. Geneetilise riskiskoori kaasamisel vähenes jämesoolevähi haigete arv kõige madalama riskiga grupis ning samal ajal suurenes haigete kuuluvus kõrgema riskiga gruppidesse.

Kokkuvõte

Bakalaureusetöö eesmärk oli leida jämesoolevähi mittegeneetilised riskitegurid ning uurida geneetilise riskiskoori mõju jämesoolevähile. Lisaks sooviti uurida, kui hästi mittegeneetilised riskitegurid ja geneetiline riskiskoor jämesoolevähki prognoosivad. Jämesoolevähina käsitleti haiguseid, mille RHK-10 koodid olid C18, C19, C20 ja C21. Mittegeneetiliste tegurite seast uuriti sugu, kehamassiindeksit, elustiiliga seotud tegureid (liikumisaktiivsus, suitsetamine, alkoholi tarbimine, söömisharjumused) ning veel mitmeid haiguseid, mis võisid mõjutada jämesoolevähi riski. Töös kasutati Tartu Ülikooli Eesti Geenivaramu andmeid, kus oli üle 48 000 geenidoonori ning 492 jämesoolevähi diagnoosi. Kasutatud andmed olid paremalt tsenseeritud ning vasakult tõkestatud. Geneetilise riskiskoori andmed olid leitud haiguste C18, C19 ja C20 põhjal.

Jämesoolevähi riskitegurid leiti Coxi võrdeliste riskide mudeli abil, kus ajaskaalana kasutati vanust ning arvestati paremalt tsenseeritud ning vasakult tõkestatud andmetega. Selgus, et mõju jämesoolevähile oli sool, piimatoodetel, tee joomisel, liikumisaktiivsusel, esimest tüüpi diabeedil, Crohni tõvel ning käärsoole, pärasoole, päraku ja pärakukanali healoomulisel kasvajal. Nimetatud haiguste korral oli diagnoosi saanutel suurem jämesoolevähi risk. Riski suurendas ka tee joomine, kuid leitud seose korral oli tõenäoliselt tegemist juhusliku leiuga. Riski vähendasid piimatoodete tarbimine ja suurem liikumisaktiivsus ning samuti oli väiksem risk naistel. Suitsetamise kui ühe teadaoleva jämesoolevähi riskifaktori mõju siinses töös ei õnnestunud näidata. Lisaks tõestati, et ka geneetilisel riskiskooril oli mõju jämesoolevähile ehk kõrgema geneetilise riski korral oli kõrgem jämesoolevähi risk.

Prognostiliste skooride leidmiseks ja testimiseks jagati andmed treening- ja testandmestikuks. Treeningandmestikus leiti jämesoolevähki prognoosivad mudelid Coxi võrdeliste riskide mudelite abil. Täpsemalt tehti järgmised mudelid: mittegeneetilisi tegureid ja geneetilist riskiskoori sisaldav mudel, vaid mittegeneetilisi tunnuseid sisaldav mudel ning vaid geneetilist riskiskoori sisaldav mudel. Saadud mudelitega leiti testandmestikus prognostilised skoorid. Kõikide kolme mudeli prognoosidel tõestati statistiliselt oluline seos jämesoolevähiga. ROC-kõveraid uurides selgus, et kõige paremini prognoosis jämesoolevähki mittegeneetilisi tegureid ja geneetilist riskiskoori sisaldav mudel. Selle mudeli korral oli tõenäosusega 64,8% juhuslikult valitud jämesoolevähi haigel kõrgem skoor kui juhuslikult valitud tervel inimesel. Selgus, et geneetilise riskiskoori mudelisse lisamine parandas

prognoosimise täpsust ümardatult 5%. Samuti vähenes koos geneetilise riskiskooriga mudeli korral kõige madalama riskiga riskigruppi kuuluvus ning suurenes kõrgema riskiga riskigruppidesse kuuluvus jämesoolevähi haigete hulgas.

Töös kasutatud geneetiline riskiskoor oli leitud haiguste C18–C20 põhjal ehk kasutatud polnud haigust C21. Kuna C21 diagnoose oli andmestikus kümme ning testandmestikus kolm, siis ei saanud juhtude vähesuse tõttu leida kinnitust, kas kasutatud geneetiline riskiskoor prognoosib ka haigust C21. Lisaks olid geenidoonorid keskmiselt vaatluse all umbes kümme aastat ning jämesoolevähi diagnoosid moodustasid andmestikust vaid ühe protsendi. Diagnooside vähesuse tõttu tuleks täpsemate tulemuste saamiseks uuringut tulevikus korrata. Samuti tuleks edaspidi riskide täpsemaks hindamiseks kasutada nn konkureerivate riskide mudelit, sest muul põhjusel surnute tsenseerimise tõttu ei ole selles töös leitud riskihinnangud päris täpsed.

Kasutatud kirjandus

[1] Tervise Arengu Instituut. (2018). *Jämesoolevähi elulemus Eestis paraneb, ent kaugmetastaasidega juhtude osakaal endiselt suur*. Vaadatud 15.03.2020 https://www.terviseinfo.ee/et/uudised/4984-jamesoolevahi-elulemus-eestis-paraneb-ent-kaugmetastaasidega-juhtude-osakaal-endiselt-suur

[2] Eesti Vähiliit. (i.a). *Vähi teke ja areng*. Vaadatud 25.02.2020 https://cancer.ee/info-vahist

[3] Tartu ülikooli Kliinikum. (2015). *Käär- ja pärasoole vähk (jämesoole vähk)*. Vaadatud 25.02.2020 https://www.kliinikum.ee/ho/info-haiguste-kohta/2-uncategorised/89-kaeaer-ja-paerasoole-vaehk-jaemesoole-vaehk

[4] Med24. (i.a). *RHK-10*. Vaadatud 11.03.2020 https://www.med24.ee/andmebaasid/rhk10

[5] Collett, D. (2003). *Modelling survival data in medical research*. CRC press.

[6] Zimmermann M. (2018). *Elukestusanalüüs vasakult tõkestatud andmete ning ajast sõltuva argumenttunnuse korral TÜ Eesti geenivaramu kohordi näitel. Magistritöö*. Tartu: Tartu Ülikool.

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

[8] Englas M., Jakobson S., Pilt E., Rahkama T., Suurväli P., Selgall da Silva M. (2019). *Eesti Vabariigi preemiad 2019*. Tallinn: Eesti Teaduste Akadeemia.

[9] Bush, W. S., Moore, J. H. (2012). Genome-wide association studies. *PLoS Computational Biology*. 8(12), doi: 10.1371/journal.pcbi.1002822

[10] Kaart T., Möls T. (2010) *Populatsioonigeneetika genotüüpide tasemel. Loengukonspekt*. Vaadatud 10.03.2020 http://www.eau.ee/~ktanel/MTMS_02_007/loeng_01_2010web.pdf

[11] Sukhumsirichart, W. (2018). Polymorphisms. *Genetic Diversity and Disease Susceptibility*. IntechOpen. doi: 10.5772/intechopen.76728

[12] Kim, S., Misra, A. (2007). SNP genotyping: technologies and biomedical applications. *Annual Review of Biomedical Engineering* 9, 289-320, doi: 10.1146/annurev.bioeng.9.060906.152037

[13] Kumar, R., Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian pediatrics*, 48(4), 277-287, doi: 10.1007/s13312-011-0055-4

[14] Kaart, T. (2012). *Binaarsete tunnuste analüüsimeetodid. Õpiobjekt.* Vaadatud 30.04.2020 `http://www.eau.ee/~ktanel/bin_tunnuste_analyys/bin_tunnuste_analyys.pdf`

[15] Therneau, T. M., Lumley, T., Atkinson E., Crowson C. (2020). *Package 'survival'.* Vaadatud 30.04.2020 `https://cran.r-project.org/web/packages/survival/survival.pdf`

[16] Norat, T., Riboli, E. (2003). Dairy products and colorectal cancer. A review of possible mechanisms and epidemiological evidence. *European Journal of Clinical Nutrition*, 57(1), 1-17, doi: 10.1038/sj.ejcn.1601522

Lisad

Lisa 1. Jämesoolevähi riski prognoosivad mudelid

Tabel 11: Mittegeneetilisi riskitegureid sisaldav mudel jämesoolevähi prognoosimiseks

Tunnus X_i	$e^{\hat{\beta}_i}$	95% usaldusintervall	p-väärtus
Sugu	0,69	(0,56; 0,85)	0,00034
E10	2,00	(1,03; 3,89)	0,040
K50	13,12	(4,19; 41,06)	$9{,}7 \cdot 10^{-6}$
D12	3,96	(1,63; 9,59)	0,0023
Tee joomine	1,10	(1,02; 1,19)	0,015
Piimatooted	0,64	(0,40; 1,03)	0,068
Aktiivsus	0,87	(0,77; 0,98)	0,023

Tabel 12: Geneetilist riskiskoori sisaldav mudel jämesoolevähi prognoosimiseks

Tunnus X_i	$e^{\hat{\beta}_i}$	95% usaldusintervall	p-väärtus
GRS	5,19	(3,54; 7,60)	$< 2 \cdot 10^{-16}$

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Laura Birgit Luitva,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Jämesoolevähi riskitegurid TÜ Eesti Geenivaramu andmete põhjal“, mille juhendajad on Jaanika Kronberg, Krista Fischer ja Tõnu Esko, 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.

Laura Birgit Luitva

18.05.2020