Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Saskia Kuusk

Rinnavähi riskitegurid ja nende mõju rinnavähi riskile TÜ Eesti geenivaramu andmete põhjal

Bakalaureusetöö matemaatilise statisitka erialal (9 EAP)

Juhendaja: Prof. Krista Fischer, *Ph.D*

Tartu 2022

Rinnavähi riskitegurid ja nende mõju rinnavähi riskile TÜ Eesti geenivaramu andmete põhjal

Bakalaureusetöö

Saskia Kuusk

Lühikokkuvõte

Käesoleva bakalauseusetöö eesmärk on kirjeldada, kuidas sõltub rinnavähi esinemise risk geneetilistest teguritest. Tartu Ülikooli Eesti geenivaramu näitel uuritakse rinnavähi polügeense riskiskoori ja rinnavähi riski suurendavate harvade geenimutatsioonide mõju rinnavähi riskile. Selleks kasutatakse elukestuse iseloomustamiseks Kaplan-Meieri graafikuid ja Coxi võrdeliste riskide mudeleid. Töö esimeses peatükis tehakse ülevaade andmetest ja tutvustus Tartu Ülikooli Eesti geenivaramust. Teises peatükis antakse ülevaade töös kasutatavast statistilisest metoodikast. Kolmandas peatükis tuuakse välja geenivaramu andmetel tehtud analüüsi tulemused. Rinnavähi geneetilise riskiskoori ja harvade geenimutatsioonide olemasolu leiti analüüsi tulemusena tugeva mõjuga rinnavähi riskile. Geenimutatsiooniga naiste hulgas ei suudetud näidata, et riskiskooril oleks mõju rinnavähi riskile.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: elukestusanalüüs, geenivaramu, polügeenne riskiskoor, rinnavähk.

Breast cancer risk factors and their impact on breast cancer risk based on data from the UT Estonian Biobank

Bachelor thesis

Saskia Kuusk

Abstract

The aim of this bachelor's thesis is to describe, how the risk of breast cancer depends on genetic factors. The effect the polygenic risk score for breast cancer and rare gene mutations, that increase the risk of breast cancer, have on the risk of breast cancer are investigated on the example of The Estonian Biobank of the University of Tartu. Kaplan-Meier graphs and Cox proportional hazards models are used to characterize life expectancy. The first chapter provides an overview of the data and an introduction to the Estonian Biobank of the University of Tartu. The second chapter gives an overview of the statistical methods used in the thesis. The third chapter presents the results of the analysis performed on the data of the biobank. The genetic risk score for breast cancer and the presence of rare gene mutations were found to have a strong effect on the risk of breast cancer. Among women with gene mutation, it could not be shown that the risk score had an effect on the risk of breast cancer.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Survival analysis, biobank, polygenic risk score, breast cancer.

Sisukord

Sissejuhatus

Käaesoleva bakalaureusetöö eesmärk on kirjeldada, kuidas sõltub rinnavähi esinemise risk geneetilistest teguritest. Tartu Ülikooli Eesti geenivaramu näitel uuritakse rinnavähi polügeense riskiskoori ja rinnavähi riski suurendavate harvade geenimutatsioonide mõju rinnavähi riskile. Selleks kasutatakse elukestuse iseloomustamiseks Kaplan-Meieri graafikuid ja Coxi võrdeliste riskide mudeleid.

Rinnavähk on kõige sagedamini esinev pahaloomuline kasvaja naiste seas. Aastal 2020 diagnoositi maailmas 2,3 miljonil naisel rinnavähk ja 685 000 naist suri selle tagajärjel (*Breast Cancer. World Health Organisation* 2022). Igal aastal diagnoositakse Eestis rinnavähk rohkem kui 500 naisel ja 25% nendest juhtudest on haiguse avastamise hetkel vähk kõrgemasse staadiumisse arenenud. Rinnavähi ravi on haiguse varajasel avastamisel efektiivsem. Sümptomite ilmumise hetkel võib vähk olla juba hakanud inimese kehas levima. Selleks, et avastada rinnavähk juba enne sümptomite ilmnemist teostatakse sõeluuringuid. (*Rinnavähk ja sõeluuringud. Mammograaf radioloogiakliinik* 2022) Eestis kutsutakse rinnavähisõeluuringutele 50-68 aastaseid naisi iga kahe aasta tagant (*Rinnavähi sõeluuringud. Haigekassa* 2022). Selline lähenemine ei ole aga kõige tõhusam, kuna sõeluuringutele kaasatakse inimesi ainult vanuse põhjal, kuid rinnavähi riskitegureid on veel teisigi. Peale selle ei saa erinevatel põhjustel sõeluuringute käigus kontrollitud kogu riskivanusesse peetud rahvastik.

Sellisele üldisele lähenemisviisile vastandub personaalne meditsiin, mis võimaldab pakkuda indiviidile kohandatud ja sihipärast ravi. Uute tehnoloogiate, sealhulgas genoomika rakendamine suurendab meie arusaama haigustest ja seega võimet isikupäraseks raviks ja ennetuseks. (Brittain, Scott ja Thomas, 2017) Tervise Arengu instituudi projekti „Personaalmeditsiini rakendamine Eestis“ (2019–2022) tulemusena luuakse geneetiliste andmete laiemaks kasutusele võtuks vajalikud eeldused 2022. aasta lõpuks (*Personaalmeditsiin. Tervise Arengu Instituut* 2022). Tartu Ülikooli Eesti geenivaramu geeniandmete põhjal võiks olla võimalik muuta Eesti rinna-

vähi sõeluuringud efektiivsemaks, kaasates uuringusse isikuid mitte ainult vanuse, vaid ka rinnavähi riski mõjutavate geeniandmete põhjal.

Läll, 2019 doktoritöös näidati, et rinnavähi geneetilisel riskiskooril on mõju rinnavähi riskile. Käesolevas töös on kasutada rohkemate geenidoonorite andmed ja lisaks ka andmed üksikute kõrge riskiga seotud geenimutatsioonide kohta. Töö eesmärk on hinnata polügeense riskiskoori mõju suuremas andmestikus, hinnata harvade geenimutatsioonde mõju Eesti andmetes ning uurida polügeense riskiskoori ja mutatsioonide võimalikku koosmõju rinnavähi riskile. Esimeses peatükis on tehtud ülevaade andmetest ja Tartu Ülikooli Eesti geenivaramu tutvustus. Teises peatükis on antud ülevaade töös kasutatavast statistilisest meetodikast. Kolmandas peatükis on toodud geenivaramu andmetel tehtud analüüsi tulemused.

Autor tänab bakalaureusetöö juhendajat Krista Fischerit nõuannete ja paranduste eest.

1 Ülevaade andmetest

Tartu Ülikooli Eesti geenivaramuga on 2022. aasta seisuga liitunud üle 200 000 inimese. Geenidoonorite andmeid on kogutud 2002. aastast ja andmebaas sisaldab geeniandmeid ligi 20% kohta Eesti täisealisest elanikonnast. (*Üldinfo. Tartu Ülikooli Eesti geenivaramu* 2022) Geenivaramu doonorite koosseis kajastab üldiselt eesti elanikkonna soolist, vanuselist ja geograafilist jaotust (*Eesti geenivaramu. Tartu Ülikool* 2021).

Geenivaramu andmebaas sisaldab peale pärilikkusaine (DNA) andmete ka andmeid doonorite tervisliku seisundi, elustiili ja toitumise kohta. Andmebaasi lisatakse perioodilistelt Eesti Haigekassa andmetest info haiguste olemasolu ja diagnooside kuupäevade kohta. Tervise Arengu instituudi andmete põhjal uuendatakse surma põhjuseid ja -kuupäevi. Viimane uuendus mõlemast andmebaasist toimus 2022. aasta alguses.

Igale indiviidile geenivaramu kohordis on leitud kaks erinevat rinnavähi geneetilist riskiskoori ja nende kombineerimisel on loodud uus skoor rinnavähi geneetilise eelsoodumuse hindamiseks. Nendel indiviididel, kellel esineb teatud rinnavähi riski suurendav harv geenimutatsioon on märgitud, millises geenis mutatsioon esineb.

Hinnatakse, et 5 kuni 10% rinnavähi juhtudest on seotud mõne konkreetse harvaesineva päriliku geenimutatsiooniga. Levinumad ja tugevaima mõjuga on muudatused geenides BRCA1 ja BRCA2. Nende mutatsioonide kandjale on Ameerika Ühendriikide andmete puhul hinnatud riskiks elu jooksul rinnavähki haigestuda kuni 72%. Rinnavähi mõõduka või kõrge riskiga seostatakse ka pärilikke mutatsioone geenides ATM, CHEK2, NBN, NF1. (*Genetics. Breastcancer.org* 2022) Tabelis 1 on toodud mutatsioonide esinemissagedused naiste seas geenivaramu andmestikus.

Analüüsis kasutatakse andmeid nende naistest geenidoonorite kohta, kes ei ole enne geenivaramuga liitumist haigestunud rinnavähki. Rinnavähi polügeense riskiskoori mõju hindamisel jäetakse andmestikust välja ka treeningandmestikus olnud

geenidoonorid. Kuna peale 2022. aasta algust liitunuid on jälgitud väga lühikest aega, jäetakse need doonorid samuti analüüsist välja. Seega analüüsitakse käesolevas töös 134 214 geenivaramuga liitunud naise andmeid. Elukestusanalüüsi jaoks vajalik tunnus elukestuse kohta määratakse doonori vanuse põhjal liitumise hetkel ja rinnavähi diagnoosi või tsenseerimise hetkel. Jälgimisaja lõpuks on haigestumine rinnavähki või tsenseerimine. Tsenseerimishetkeks on kas inimese surm või viimane uuendus rinnavähki haigestumistest ja surmadest andmebaasis.

Geenidoonoritest 65,5% on naised ja 34,5% mehed. Naistest geenidoonorite jälgimisaeg (aeg liitumisest rinnaväki haigestumiseni või tsenseerimiseni) on keskmiselt 5,6 aastat. Naiste keskmine vanus geenivaramuga liitumisel on 45,3 aastat. Rinnavähi haigusjuht on registreeritud 2% naisdoonoritest. Nendest 34% on saanud rinnavähi diagnoosi peale geenivaramuga liitumist. Vähemalt üks rinnavähi riski suurendav geenimutatsioon esineb 437 naisdoonoril (0,33%).

Tabel 1: Geenimutatsioonide esinemissagedused naiste seas.

Mutatsioon	**sagedus**
BRCA1	140
BRCA2	76
ATM	16
CHEK2	199
NBN	5
NF1	1
KOKKU:	437

2 Statistiline metoodika

Elukestusanalüüs on matemaatilise statistika valdkond, mida kasutatakse selliste andmete uurimiseks, kus huvipakkuv tunnus on aeg mingist algmomendist kuni huvipakkuva sündmuse toimumiseni. Sageli kirjeldavad need andmed subjekti eluaega ning sellest tuleneb ka meetoodika nimetus. Peale suremuse uurimise on aga sellel palju kasutusvõimalusi erinevates valdkondades. Näiteks võib huvipakkuvaks ajavahemikuks olla ka aeg patsiendi ravi algusest kuni tervenemiseni või seadme rikketa töötamise aeg. Kuigi need ajavahemikud ei kirjelda elu kestust nimetatakse neid siiski elukestuseks. Kastutusel on ka eestikeelne oskussõna elulemusanalüüs. (Allison, 2010; Fischer, 2007)

Elukestusanalüüsis kasutatakse kestusandmeid, mille all mõeldakse valimeid, kus iga vaatlus annab informatsiooni protsessi kestuse kohta (Fischer, 2007). Nende andmete puhul pole oluline teada mitte ainult seda, millisel vaatlusel on sündmus toimunud, vaid ka millal see aset leidis (Allison, 2010). Kestusandmete üks põhiomadusi on, et uuritav tunnus on mittenegatiivne diskreetne või pidev juhuslik suurus. Teine põhiomadus, mis on ka üks peamisi probleemiallikaid elukestusanalüüsis, on tsenseeritus. See tekib juhul kui huvipakkuv sündmus ei ole kogu valimis vaadeldud. Kõige levinum tsensreerituse juht on kui huvipakkuva ajavahemiku kohta on teada ainult, et see ületab mingi teatud väärtuse (teada on vaid ajamoment, mil sündmus ei ole veel toimunud). Sellisel juhul on tegemist paremalt tsenseeritusega. (Moore, 2016) Antud töös on samuti tegemist selliselt tsenseeritud andmetega.

Kestusandmete puhul ei ole enamasti võimalik kasutada tavalisi statistilisi meetodeid. Tihti on andmete puhul vajalik normaaljaotuse eeldus, mis eelnevalt kirjeldatud andmete puhul ei ole peaaegu kunagi täidetud, esiteks juba väljatoodud mittenegatiivsuse omaduse tõttu. Teiseks, pole kestusandmed enamikul juhtudel sümmeetrilise jaotusega. (Fischer, 2007)

2.1 Polügeenne riskiskoor

See peatükk on kirjutatud Läll, 2019 doktoritöö „Risk scores and their predictive ability for common complex diseases“ põhjal.
Laiemas mõttes saab geneetilise varieeruvuse allikad jagada kahte klassi. Neist kõige levinum ja enim uuritud geneetilise varieeruvuse allikate klass on ühenukleotiidilised polümorfismid (SNPd). Hinnatakse, et inimese genoom sisaldab vähemalt 11 miljonit SNPd (Frazer *et al.*, 2009). Tänu asjaolule, et SNPd on levinud kogu genoomis, on need kasulikud genoomsete muutuste uurimiseks.

Üksikud SNPd on valdavalt väikese mõjuga geneetilisele varieeruvusele. Seetõttu kasutatakse mingi tunnuse prognoosimiseks ühe SNPi asemel erinevate SNPide kaalutud alleelidooside summat. Alleelidoos on genotüübis valitud alleeli esinemiste arv. Genotüüp on kõigi organismi kromosoomides paiknevate geenialleelide kogum. Eeldatakse, et iga SNP jaoks on kolm võimalikku genotüüpi. Olgu näiteks genotüübid aa, Aa ja AA. Valime lugemiseks alleeli A. Alleelidoosid on siis vastavalt alleeli A esinemiste arvud 0, 1 ja 2. Mitme SNPi kombineerimisel saadud tunnust nimetatakse geneetiliseks riskiskooriks või ka polügeenseks riskiskooriks. See on osutunud kasulikuks erinevate komplekshaiguste nagu teist tüüpi diabeedi või rinnavähi geneetilise eelsoodumuse hindamiseks.

Geneetiline riskiskoor i-nda indiviidi jaoks on üldiselt defineeritud kui k SNPi kaalutud summa:

$$GRS_i = \sum_{j=1}^{k} w_j X_{ij}, \tag{1}$$

kus X_{ij} märgib j-nda SNP ja i-nda indiviidi alleelidooside arvu ning $w_j \in (-\infty, \infty)$ on j-nda SNP kaal. Geneetilise riskiskoori leidmisel parima võimaliku ennustava väärtuse saamiseks on peamine küsimus selles, kuidas valida kaasatav SNPde komplekt ja nendele vastavad kaalud. Geneetilise riskiskoori puhul on probleemiks see, et see pole üheselt määratud. Erinevaid alusuuringuid kasutades võib saada mitu skoori, mis on üksteisega vaid nõrgalt korreleeritud. Rinnavähi eelsoodumust

saab seetõttu kõige paremini hinnata mitme erineva geneetilise riskiskoori kombineerimisel üheks skooriks. Riskiskoori hinnangu täpsuse suurendamiseks on loodud mitmest olemasolevast geneetilisest riskiskoorist uus skoor MetaGRS.

Olgu $Z_{i1}, \cdots, Z_{ip}$ valemiga (1) saadud p standardiseeritud geneetilist riskikoori i-nda indiviidi jaoks. Treeningandmestikust on hinnatud vastavate skooride mõju ja nende hinnangud on tähistatud kui $\hat{\alpha}_1, \cdots, \hat{\alpha}_p$. Pearsoni korrelatsioonikordaja ρ_{jk} on arvutatud skooride $Z_{.j}$ ja $Z_{.k}$ vahel samuti treeningandmestikust. MetaGRS i-nda indiviidi jaoks on defineeritud kui

$$MetaGRS_i = \frac{\sum_{j=1}^{p} \hat{\alpha}_j Z_{ij}}{\sqrt{\sum_{j=1}^{p} \hat{\alpha}_j^2 + 2\sum_{j=1}^{p}\sum_{k=j+1}^{p} \hat{\alpha}_j \hat{\alpha}_k \rho_{jk}}},$$

kus riskiskooride mõju hinnangud on käsitletud konstantidena. Antud töös on kasutatud rinnavähi riski hindamiseks MetaGRSi, kus on kombineeritud kaks rinnavähi eelsoodumust kõige paremini ennustavat geneetilist riskiskoori.

2.2 Elukestusanalüüsi meetodid

Järgnevad elukestusanalüüsi meetodite peatükid põhinevad Collett, 2015 raamatul „Modelling survival data in medical research,“ kui ei ole märgitud teisiti.

2.2.1 Riski- ja üleelamisfunktsioonid

Elukestusanalüüsi meetodid tuginevad elukestuse jaotusele ja selle hindamiseks on kahe põhilise funktsioonina defineeritud riski- ja üleelamisfunktsioon.

Olgu T juhuslik suurus, mis kirjeldab subjekti elukestust. Selle juhusliku suuruse jaotusfunktsioon on esitatav kujul:

$$F(t) = P(T < t) = \int_0^t f(u)\, du, \tag{2}$$

kus $f(t)$ on T tihedusfunktsioon. Üleelamisfunktsioon $S(t)$ määrab tõenäosuse, et

huvipakkuv sündmus pole toimunud kuni ajahetkeni t. Võrdusest (2) saab üleelamisfunktsioonile kuju:

$$S(t) = P(T \geq t) = 1 - P(T < t) = 1 - F(t).$$

Selle funktsiooni väärtus kohal 0 on 1 ja t kasvades kas kahaneb või jääb konstantseks.

Riskifunktsioon iseloomustab riski, et sündmus toimub teatud ajahetkel t. Funktsioon on piirväärtuseks tõenäosusele, et sündmus toimub ajahetkel t tingimusel, et see ei toimunud enne seda momenti. Täpsemalt uuritakse sündmuse toimimist lõpmatult väikeses ajavahemikus t ja $t + \Delta t$ vahel, tingimusel et $T \geq t$. Riskifunktsiooni kuju on defineeritud järgmiselt:

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

Kehtib ka seos:

$$h(t) = \frac{f(t)}{S(t)}.$$

2.2.2 Kaplan-Meieri hinnang

Kaplan-Meieri hinnangut kasutatakse üleelamisfunktsiooni mitteparameetrilsel hindamisel ja ka graafilisel kujutamisel.

Olgu valimis n subjekti vaadeldud elukestustega $t_1, \cdots, t_n$ ja sündmus on vaadeldud nende seast r objektil, ehk $r \leq n$. Need elukestused kujutatakse variatsioonireas $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Tähistame n_j-ga subjektide arvu, kellel pole ajahetkeni $t_{(j)}$ vaadeldud sündmus, ega pole tsenseeritud (kaasaarvatud need, kellel toimub sündmus või tsenseerimine sellel ajahetkel). Ehk n_j tähistab subjektide arvu, kes ajahetkeks $t_{(j)}$ on veel vaatluse all. Tähistus d_j märgib subjektide arvu, kellel vaadeldakse sündmus ajahetkel $t_{(j)}$.

Siis tõenäosus, et subjektil vaadeldakse sündmus ajamomendil $t_{(j)}$ on hinnatav kui

$\frac{d_j}{n_j}$. See on ka riskifunktsiooni hinnang $\hat{h}_j$ vastaval ajamomendil (Fischer, 2007). Üleelamistõenäsos on seega $\frac{n_j - d_j}{n_j}$ ja sellest saame Kaplan-Meieri hinnanguks üleelamisfunktsioonile

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

kus $t_{(k)} < t < t_{(k+1)}$ ja $k = 1, \cdots, r$. Juhul kui $t < t_{(1)}$, siis $\hat{S}(t) = 1$, ja kui $t \geq t_{(r)}$, siis $\hat{S}(t) = 0$.

2.2.3 Coxi võrdeliste riskide mudel

Üks peamisi põhjuseid kestusandmete modeleerimiseks on kindlaks teha, millised tunnused ja nende kombinatsioonid mõjutavad riskifunktsiooni. Kõige sagedamini kasutatakse elukestusanalüüsis võrdeliste riskide mudelit.

Olgu kirjeldavate tunnuste väärtused määratud vektoriga $\mathbf{x} = (x_1, x_2, \cdots, x_p)'$, kus $x_j,\ j = 1, \cdots, p$ määrab tunnuste $\mathbf{X}_1, \cdots, \mathbf{X_p}$ seast $\mathbf{X}_j$ väärtuse. Riskifunktsioon subjektile, kelle puhul on kõigi vektori $\mathbf{x}$ poolt määratud tunnuste väärtus 0, on $h_0(t)$. Võrdeliste riskide mudel eeldab, et riskifunktsioon i-nda subjekti jaoks on esitatav kui

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

kus $\psi(\mathbf{x}_i)$ on vektori $\mathbf{x}_i$ funktsioon. Juhul, kui $\psi(0) = 1$, siis võib seda funktsiooni tõlgendada suhtelise riskina subjektile ajahetkel t, võrreldes subjektiga, kellel $\mathbf{x}$ on võrdne nullvektoriga. Valides $\psi(\mathbf{x}_i) = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$, kus $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ on vektor kirjeldavate tunnuste kordajatest mudelis, saab üldine võrdeliste riskide mudel kuju

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) h_0(t), \tag{3}$$

kus $x_{ki},\ k = 1, 2, \cdots, p$ on i-nda indiviidi k-nda kirjeldava tunnuse $\mathbf{X}_k$ väärtus. Coxi võrdeliste riskide mudel on mudel, kus $h_0(t)$ kuju ei määrata ja hinnatakse vaid parameetreid $\beta_1 \cdots \beta_p$. Seetõttu on tegemist poolparameetrilise meetodiga.

Järgneva tulemuse esitamiseks tuuakse uuesti välja kasutatavad tähistused. Olgu vaatluse all n subjekti, millest r subjektil on vaadeldud sündmus. Ajavahemikud sündmuse toimumiseni kujutatakse variatsioonireas $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Subjektide hulk, kellel pole ajahetkeks $t_{(j)}$ sündmuse toimumine vaadeldud ega pole tsenseeritud, on $R(t_{(j)})$. Hulka $R(t_{(j)})$ nimetatakse riskigrupiks. Osalise tõepära funktsioon mudeli jaoks võrdudses (3) on

$$L = \prod_{j=1}^{r} \frac{h_0(t)\psi(\mathbf{x}_{(i)})}{\sum_{l \in R(t_{(j)})} h_0(t)\psi(\mathbf{x_l})} = \prod_{j=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{x_{(j)}})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}'\mathbf{x}_l)},$$

kus $x_{(j)}$ on vektor subjekti kirjeldavate tunnuste väärtustest, kellel vaadeldakse sündmus j-ndana ajal $t_{(j)}$. Saab näidata, et $\boldsymbol{\beta}$ väärtused, mis maksimeerivad funktsiooni L on nihketa ja mõjusad hinnangud mudeli parameetritele.

2.2.4 Elukestusanalüüsi meetodite rakendamine R abil

Elukestusanalüüsi meetodite rakendamiseks R tarkvaras võib kasutada paketti *survival*. Esiteks tuleb luua spetsiaalne struktuur tsenseeritud kestusandmete käsitlemiseks. Selleks on funktsioon $Surv(time,\ event)$, kus *time* on subjekti aeg, ehk kui kaua oli subjekt vaatluse all ja *event* on sündmuse toimumise identikaator, kus 1 tähistab sündmuse toimumist. (Moore, 2016; Therneau, 2007)

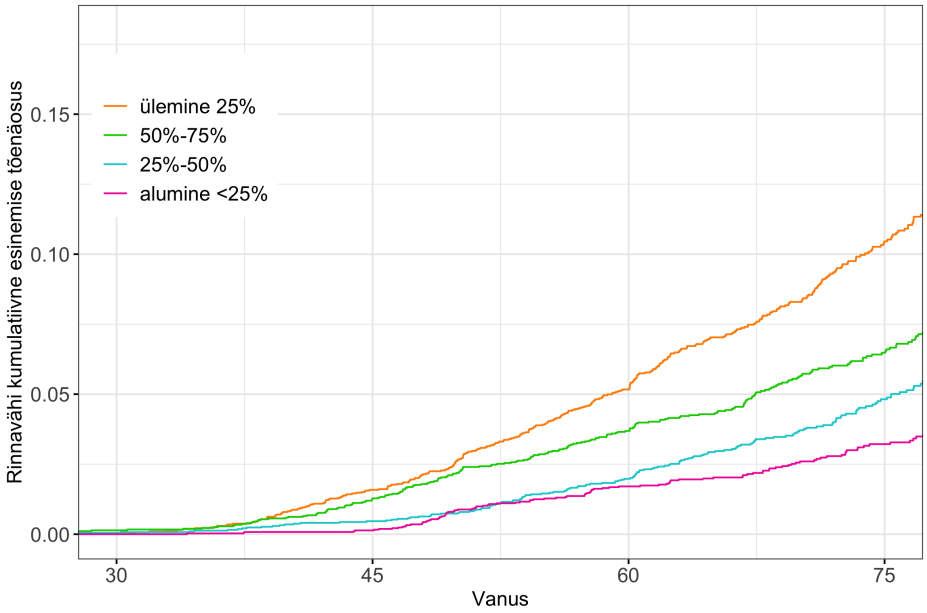
Kaplan-Meieri hinnangu leidmiseks üleelamisfunktsioonile on kasutusel funktsioon $survfit(formula \sim x,\ data, \cdots)$, kus *formula* on *Surv*-objekt ja $\sim$ operaatorist paremal on plussmärgiga eraldatud tunnused. (Therneau, 2007)

Coxi võrdeliste riskide mudeli leidmiseks kasutatakse funktsiooni $coxph(formula \sim x, data, \cdots)$. Funktsiooni kompomendid on samad, mis eelneval *survfit* funktsioonil. (Therneau, 2007)

3 Andmeanalüüs

3.1 Risk polügeense riskiskoori kategooriates

Polügeense riskiskoori mõju hindamiseks rinnavähi riskile leiti Kaplan-Meieri hinnang üleelamisfunktsioonle ja sobitati andmetele võrduse 3 põhjal Coxi võrdeliste riskide mudel. Polügeenne riskiskoor jagati nelja riskiskoori kategooriasse vastavalt kvartiilidele. Joonisel 1 on kujutatud Kaplan-Meieri hinnangu põhjal saadud hinnang rinnavähi kumulatiivsele esinemise tõenäosusele $(1 - S(t))$.



Joonis 1: Rinnavähi kumulatiivne esinemissagedus rinnavähi polügeense riskiskoori kategooriates naiste seas vanuses 30 kuni 75 aastat.

Jooniselt 1 on näha, et naistel rinnavähi riskiskoori kõrgemas kvartiilis on 75. eluaastaks hinnatud kumulatiivseks rinnavähi esinemise tõenäosuseks 10,5% (95% UI 9,3 kuni 11,6%). Naistel rinnavähi riskiskoori kolmandas kvartiilis on selle tõenäosuse hinnanguks 6,5% (95% UI 5,6 kuni 7,4%), teises ja esimeses kvartiilis vastavalt 4,8% (95% UI 4,0 kuni 5,6%) ja 3,2% (95% UI 2,6 kuni 3,9%)

Tabelis 2 on kujutatud eksponentfunktsiooni väärtused Coxi mudeliga hinnatud parameetritest, ehk riskisuhted, mis iseloomustavad mitu korda erineb vastava kategooria risk riskist nn baastasemel. Baastasemeks on kõige alumine riskikategooria.

Tabel 2: Polügeense riskiskoori kategooriatele vastavad riskisuhted.

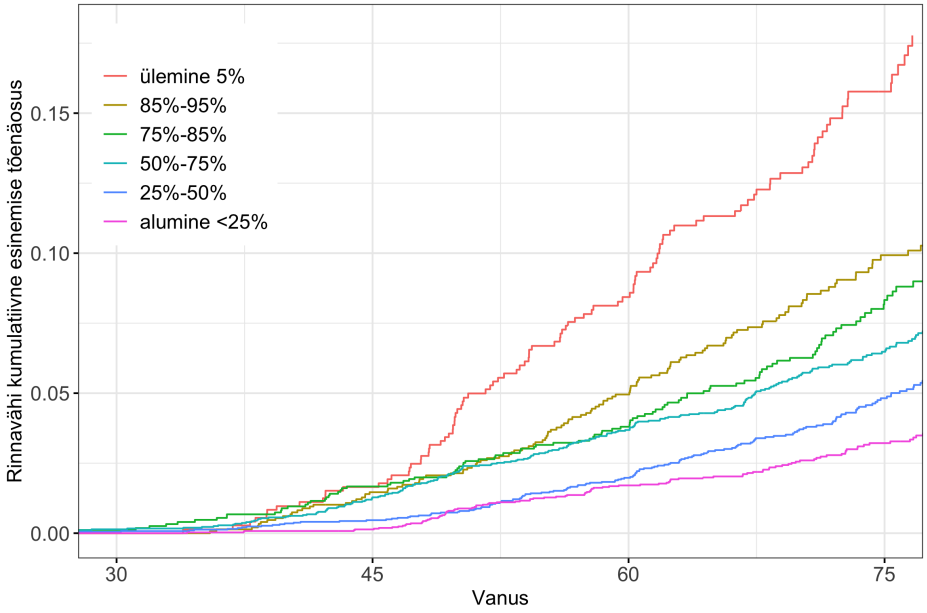
Riskikategooria	**exp($\hat{\beta}$)**	**95% usaldusintervall**	**p-väärtus**
25 kuni 50%	1,51	(1,18; 1,92)	0,0009
50 kuni 75%	2,14	(1,71; 2,68)	$4,96 \cdot 10^{-11}$
ülemine 25%	3,51	(2,84; 4,34)	$< 2 \cdot 10^{-16}$

Kõigis riskiskoori kategooriates on rinnavähi risk statistiliselt oluliselt erinev riskiskoori madalaimast kategooriast. Naistel riskiskoori jaotuse kõrgeimas kvartiilis on 3,51 (UI 95% 2,84 kuni 4,34) korda suurem risk haigestuda rinnavähki kui naistel madalaimas kvartiilis. Kolmandas ja teises kvartiilis on risk haigestumiseks vastavalt 2,14 (95% UI 1,71 kuni 2,68) ja 1,51 (95% UI 1,18 kuni 1,92) korda suurem.

Kõige kõrgemas riskiskoori kategooras on usaldusintervalli laius kõige suurem. Täpsemaks hindamiseks on ülemine kvartiil jagatud omakorda kolmeks protsentiiliks. Joonisel 2 on selliselt saadud riskiskoori kategooriatega Kaplan-Meieri hinnang rinnavähi kumulatiivsele esinemise tõenäosusele. Nõnda tuleb välja kuivõrd selle kvartiili siseselt rinnavähi risk erineb. Järsemat kasvu rinnavähi kumulatiivse esinemise tõenäosuse hinnangus on näha polügeense riskiskoori kõrgema 5% kategoorias.

Kõrgema 5% hulka kuuluvatel naistel on hinnatud rinnavähi kumulatiivseks esinemise tõenäosuseks 75. eluaastaks 15,8% (95% UI 12,8 kuni 18,7%). Protsentiilide vahemikku 85 kuni 95% jäävatel naistel on hinnatud selleks tõenäosuseks 9,9% (95% UI 8,2 kuni 11,7%) ja vahemikku 75 kuni 85% jäävatel 8,3% (95% UI 6,7 kuni 10%). Kolmandas kvartiilis on rinnavähi kumulatiivse esinemise tõenäosuse hinnang 6,5% (95% UI 5,6 kuni 7,4%), teises ja esimeses kvartiilis vastavalt 4,8% (95% UI 4 kuni 5,6%) ja 3,2% (95% UI 2,6 kuni 3,9%).

Tabelis 3 on toodud Coxi mudeli hinnangud riskusuhtele viies erievas riskikategoorias võrreldes kõige madalama kategooriaga.



Joonis 2: Rinnavähi kumulatiivne esinemissagedus rinnavähi polügeense riskiskoori kategooriates naiste seas vanuses 30 kuni 75 aastat.

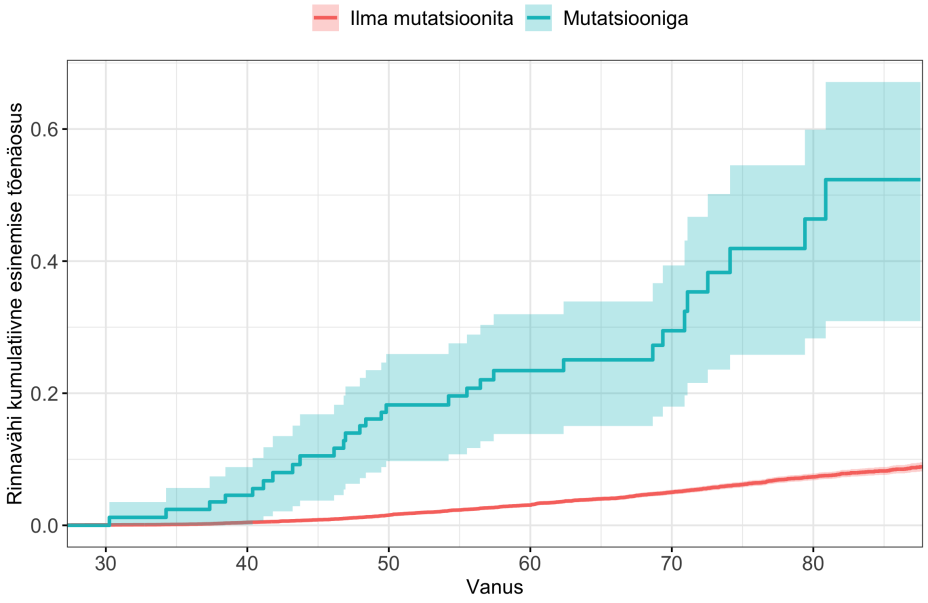
Tabel 3: Polügeense riskiskoori kategooriate riskisuhted, kus ülemine kvartiil on jaotatud kaheks väiksemaks protsentiiliks.

Riskikategooria	**exp($\hat{\beta}$)**	**95% usaldusintervall**	**p-väärtus**
25 kuni 50%	1,51	(1,18; 1,92)	0,0009
50 kuni 75%	2,14	(1,71; 2,68)	$4,96 \cdot 10^{-11}$
75 kuni 85%	2,70	(2,08; 3,51)	$7,12 \cdot 10^{-14}$
85 kuni 95%	3,30	(2,58; 4,24)	$< 2 \cdot 10^{-16}$
ülemine 5%	5,63	(4,34; 7,30)	$< 2 \cdot 10^{-16}$

Naistel polügeense riskiskoori kõrgeima 5% hulgas on 5,63 (95% UI 4,34 kuni 7,30) korda suurem risk haigestuda rinnavähki, kui naistel kõige madalamas riskiskoori kategoorias. Võrreldes naistega, kelle riskiskoor on alla mediaani, on kõrgeima 5% hulka kuuluvatel naistel 4.51 (95% UI 3,63 kuni 5,60) korda suurem risk haigestuda rinnavähki. Kogu ülejäänud kohordiga võrreldes (naised riskiskooriga alla 95% protsentiili) on kõrgeimas 5% protsentiilis 3,05 (95% UI 2,51 kuni 3,71) korda suurem risk.

3.2 Risk harvade geenimutatsioonide olemasolul

Mutatsioonide mõju hindamiseks rinnavähi riskile leiti Kaplan-Meieri hinnang rinnavähi kumulatiivsele esinemise tõenäosusele mutatsiooniga ja mutatsioonita grupis. Joonisel 3 on kujutatud vastavad kõverad.

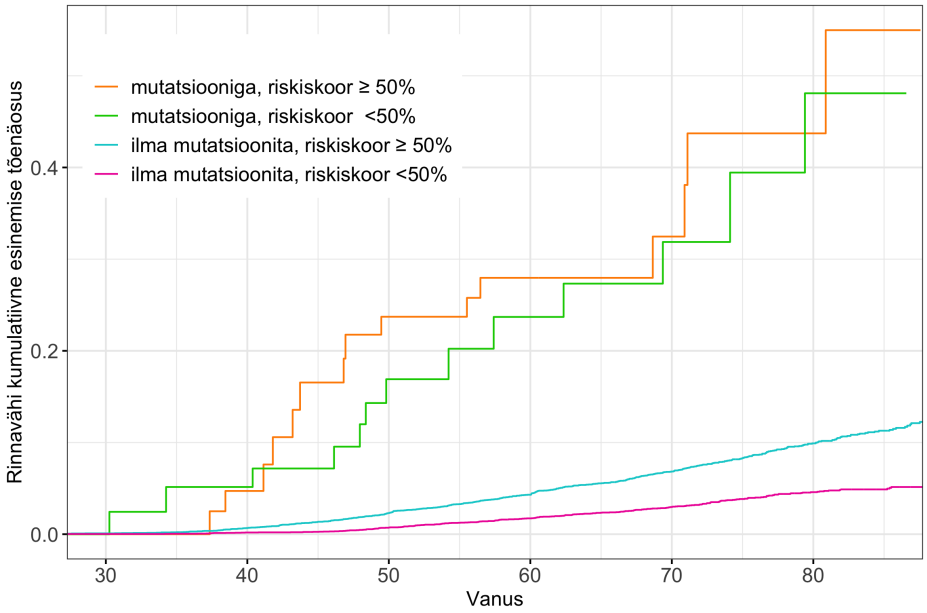


Joonis 3: Rinnavähi kumulatiivne esinemissagedus mutatsiooni esinemise põhjal vanuses 30 kuni 85 aastat.

Naistel, kellel esineb teatud rinnavähi riski suurendav geenimutatsioon on 85. eluaastaks rinnavähi kumulatiivne esinemise tõenäosuseks hinnatud 52,3% (95% UI 30,9 kuni 67,1%). Ilma geenimutatsioonita naistel on selleks tõenäosuseks 8,2% (95% UI 7,6 kuni 8,8%).

Jooniselt 3 paistab välja, kui tugevalt mõjutab mutatsiooni olemasolu rinnavähki haigestumise riski. Kuna polügeense riskiskoori arvutamisel ei võeta arvesse geenimutatsioone, jagatakse kohort neljaks kategooriaks, et näha kuidas riskiskoor ja geenimutatsioon koos rinnavähi riski mõjutavad. Mutatsiooniga ja ilma mutatsioonita grupid jagati omakorda kaheks vastavalt sellele, kas rinnavähi riskiskoor jääb

üle mediaani või allapoole sellest. Joonsiel 4 on näha saadud Kaplan-Meieri hinnangud kumulatiivsele rinnavähi esinemise tõenäosusele nendes kategooriates.



Joonis 4: Rinnavähi kumulatiivne esinemissagedus mutatsiooni esinemise ja rinnavähi polügeense riskiskoori kategooria põhjal vanuses 30 kuni 85 aastat 95% usaldusintervallidega.

Vaadates tekkinud kõveraid ilma geenimutatsioonita kategooriates, on naistel kelle geneetiline riskiskoor on üle mediaani, suurem rinnavähi kumulatiivne esinemise tõenäosus. Ilma mutatsioonita ja riskiskooriga üle mediaani jäävatel naistel on 85. eluaastaks selle tõenäosuse hinnanguks 11,3% (95% UI 10,3 kuni 12,3%) ja ilma mutatsioonita, aga riskiskooriga alla mediaani on selleks 4,9% (95% UI 4,3 kuni 5,5%).

Võrreldes aga mutatsiooniga kategooriaid ei paista polügeense riskiskoori mõju nii suur olevat. Kaks kõverat jooksevad üksteise lähedalt ja on näha ka lõikumisi. Mutatsiooniga grupis nendel, kellel geneetiline riskiskoor on üle mediaani, on 85. eluaastaks hinnatud rinnavähi kumulatiivseks esinemise tõenäosuseks 55,0% (95% UI 21,4 kuni 74,2%) ja riskiskooriga alla mediaani on selleks hinnatud 48,1% (95%

UI 18,6 kuni 66,9%). On näha, et usaldusintervallid on suures osas kattuvad, mille tõttu olulist erinevust kumulatiivses rinnavähi esinemise tõenäosuses ei ole. Usaldusintervallid langevad vähemalt osaliselt kokku kogu graafiku ulatuses.

Geenimutatsioonide ja polügeense riskiskoori koosmõju hindamiseks sobitati veel Coxi võrdeliste riskide mudel, kus sõltumatuteks muutujateks on mutatsiooni olemasolu, polügeenne riskiskoor ja nende koosmõju. Tulemused on näha tabelis 4. Baastase on ilma mutatsioonita tase.

Tabel 4: Riskisuhted mutatsiooni ja polügeense riskiskooriga.

Riskikategooria	$\hat{\beta}$	**exp($\hat{\beta}$)**	**95% usaldusintervall**	**p-väärtus**
mutatsioon	2,25	9,50	(6,51; 13,88)	$< 2 \cdot 10^{-16}$
metaGRS	0,79	2,20	(1,98; 2,44)	$< 2 \cdot 10^{-16}$
mutatsioon:metaGRS	-0,69	0,50	(0,28; 0,90)	0,0206

Mõlemad sõltumatud muutujad ja nende koosmõju tulid statistiliselt olulised. Mutatsiooni olemasolu suurendab rinnavähi riski 9,5 korda. Geneetilise riskiskoori parameetri väärtuseks on hinnatud 0,79, polügeense riskiskoori ja mutatsiooni koosmõju parameetri hinnang on -0,69. Nende parameetrite väärtused on absoluutväärtuselt väga sarnased ja mudelist tulebki selle tõttu välja, et mutatsiooniga grupis ei mõjuta geneetiline riskiskoor rinnavähi riski nii tugevalt. Põhjuseks, miks ei suudetud tuvastada riskiskoori mõju mutatsiooniga naiste hulgas võib olla ka suhteliselt väike mutatsiooniga naiste arv.

Kokkuvõte

Bakalaureusetöö eesmärk oli kirjeldada, kuidas sõltub rinnavähi esinemise risk geneetilistest teguritest. Selleks uuriti Tartu Ülikooli Eesti geenivaramu näitel rinnavähi polügeense riskiskoori ja rinnavähi riski suurendavate harvade geenimutatsioonide mõju rinnavähi riskile. Bakalaureusetöös kasutati elukestuse iseloomustamiseks Kaplan-Meieri graafikuid ja Coxi võrdeliste riskide mudeleid.

Täpsemalt uuriti polügeense riskiskoori mõju rinnavähi riskile erinevates riskikategooriates. Esialgu neljas kategoorias riskiskoori kvartiilide põhjal ja siis kuues kategoorias, kus neljas kvartiil jagati omakorda kolmeks protsentiiliks. Kaplan-Meieri hinnangu põhjal saadi graafikud rinnavähi kumulatiivsele esinemise tõenäosusele nendes kategooriates. Samuti sobitati andmetele Coxi võrdeliste riskide mudel, iseloomustamaks mitu korda erineb vastava kategooria risk riskist kõige madalamas riskiskoori kategoorias. Riski iseloomustamiseks harvade geenimutatsioonide olemasolul leiti Kaplan-Meieri hinnang rinnavähi kumulatiivsele esinemise tõenäosusele mutatsiooniga ja mutatsioonita grupis. Rinnavähi geneetilise riskiskoori ja harvade geenimutatsioonide olemasolu koosmõju hindamiseks jagati kohort neljaks kategooriaks. Mutatsiooniga ja mutatsioonita grupid jagati omakorda kaheks vastavalt sellele, kas rinnavähi riskiskoor jääb üle mediaani või allapoole sellest. Leiti Kaplan-Meieri hinnangud kumulatiivsele rinnavähi esinemise tõenäosusele nendes kategooriates. Koosmõju hindamiseks sobitati veel Coxi võrdeliste riskide mudel, kus sõltumatuteks muutujateks olid mutatsiooni olemasolu, polügeenne riskiskoor ja nende koosmõju.

Leiti, et polügeensel riskiskooril on oluline mõju rinnavähi riskile, kinnitades Läll, 2019 tulemuste kehtivust ka suuremas valimis. Rinnavähi riski suurendavate harvade geenimutatsioonide olemasolu korral on risk selgelt suurem kui ilma mutatsioonita, aga kõrge risksikooriga naistel. Käesolevas töös ei õnnestunud näidata, et riskiskooril võiks olla mõju mutatsiooni kandvate naiste jaoks – see mõju kas puudub või ei olnud andmetes selle tõestamiseks piisavalt võimsust.

Kasutatud allikad

Allison, Paul D (2010). *Survival analysis using SAS: a practical guide*. Sas Institute.

Breast Cancer. World Health Organisation (2022). URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (vaadatud 02.05.2022).

Brittain, Helen K, Richard Scott ja Ellen Thomas (2017). "The rise of the genome and personalised medicine". *Clinical Medicine* 17.6, lk. 545.

Collett, David (2015). *Modelling survival data in medical research*. CRC press.

Eesti geenivaramu. Tartu Ülikool (2021). URL: https://genomics.ut.ee/et/sisu/eesti-geenivaramu-0 (vaadatud 22.04.2022).

Fischer, Krista (2007). *Elukestusanalüüs. Loengukonspekt*. Tartu: Tartu Ülikool, tervishoiu instituut.

Frazer, Kelly A, Sarah S Murray, Nicholas J Schork ja Eric J Topol (2009). "Human genetic variation and its contribution to complex traits". *Nature Reviews Genetics* 10.4, lk. 241–251.

Genetics. Breastcancer.org (2022). URL: https://www.breastcancer.org/risk/risk-factors/genetics (vaadatud 09.05.2022).

Läll, Kristi (2019). "Risk scores and their predictive ability for common complex diseases". Doktoritöö. Tartu Ülikool.

Moore, Dirk F (2016). *Applied survival analysis using R*. Springer.

Personaalmeditsiin. Tervise Arengu Instituut (2022). URL: https://www.tai.ee/et/personaalmeditsiin (vaadatud 09.05.2022).

Rinnavähi sõeluuringud. Haigekassa (2022). URL: https://www.haigekassa.ee/soeluuring (vaadatud 02.05.2022).

Rinnavähk ja sõeluuringud. Mammograaf radioloogiakliinik (2022). URL: https://mammograaf.ee/patsiendile/rinnavahk-ja-soeluuringud/ (vaadatud 02.05.2022).

Therneau, Terry (2007). "Survival analysis, including penalised likelihood". *R help guide*.

Üldinfo. Tartu Ülikooli Eesti geenivaramu (2022). URL: https://geenidoonor.ee/geenivaramu (vaadatud 22.04.2022).

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Saskia Kuusk,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Rinnavähi riskitegurid ja nende mõju rinnavähi riskile TÜ Eesti geenivaramu andmete põhjal," mille juhendaja on Krista Fischer, 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.

Saskia Kuusk
10.05.2022