Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Marili Zimmermann

Elukestusanalüüs vasakult tõkestatud andmete ning ajast sõltuva argumenttunnuse korral TÜ Eesti geenivaramu kohordi näitel

Magistritöö matemaatilise statistika erialal (30 EAP)

Juhendajad: Prof. Krista Fischer, *Ph.D*
Nele Taba, *MSc*

Tartu 2018

Elukestusanalüüs vasakult tõkestatud andmete ning ajast sõltuva argumenttunnuse korral TÜ Eesti geenivaramu kohordi näitel

Käesoleva magistritöö eesmärgiks on välja selgitada, milline on kõige sobivam meetod elukestusanalüüsi läbiviimiseks, kui andmetes esineb nii paremalt tsenseeritust kui ka vasakult tõkestatust. Lisaks pakkus huvi, kas ja kuidas peaks arvestama ajas muutuvate argumenttunnustega ning milline ajaskaala on epidemioloogilise uuringu analüüsimisel parim. Leitud tulemuste põhjal rakendati kõige sobivamat meetodit TÜ Eesti geenivaramu andmete analüüsil. Esmalt tuuakse ülevaade elukestusanalüüsi olemusest ning tähtsamatest aspektidest, mida sellise analüüsi juures tuleb jälgida. Töö teises peatükis kirjeldatakse läbi viidud simulatsioonuuringut ning tuuakse välja tulemused. Kolmandas peatükis kirjeldatakse analüüsis kasutatavat Eesti geenivaramu andmestikku ning kirjeldatakse tehtud elukestusanalüüsi. Ühtlasi vaadatakse ka elukestust erinevate riskitegurite lõikes nagu sugu, haridus, kehamassiindeks, II tüüpi diabeedi diagnoos ning geneetilise riskiskoori väärtus. Riskitegurite seos elukestusega tuli analüüsi käigus selgelt välja.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika
Märksõnad: elukestusanalüüs, simulatsioon, matemaatiline statistika, geenidoonorid

Survival analysis of left truncated data and with time-dependent variable, based on Estonian Genome center's cohort

The goal of this thesis is to determine which model is the best for survival analysis with right censored and left truncated data. In addition, the thesis addressed the questions of how to add time-dependent variable to the model and which time scale should be used in epidemiological research. The right model was thereafter chosen for the analysis of the survival data from the Estonian Genome Center. Firstly, an overview of survival analysis and its important aspects is presented. The second part of this thesis describes the conducted simulation analysis and the following results. The third part of this thesis includes the conducted survival analysis using the data of The Estonian Genome Center. The results of the survival analysis show significant differences in survival times based on a number of variables, such as sex, education, body mass index, diagnosis of type II diabetes and genetic risk scores.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics
Keywords: survival analysis, simulation, mathematical statistics, gene donors

Sisukord

Sissejuhatus

Käesoleva magistritöö eesmärgiks on välja selgitada, milline on kõige sobivam meetod elukestvusanalüüsi läbiviimiseks, kui andmetes esineb nii paremalt tsenseeritust kui ka vasakult tõkestatust. Lisaks pakkus huvi, kas ja kuidas peaks arvestama ajas muutuvate argumenttunnustega ning milline ajaskaala on epidemioloogilise uuringute analüüsimisel parim. Selleks viiakse läbi mitu simulatsioonuuringut. Eesti geenivaramu andmebaasi näitel uuritakse geenidoonorite elukestust ning surma riski suurendavaid tegureid (nt suitsetamisstaatus, II tüüpi diabeedi diagnoos, geneetilise riskiskoori väärtus).

Tartu Ülikooli genoomika instituudi alla kuuluv Eesti geenivaramu on Tartu Ülikooli koosseisus olev teadus- ja arendusasutus, mille põhiülesanneteks on edendada geeniuuringute arengut, koguda teavet Eesti rahvastiku tervise ja pärilikkuse informatsiooni kohta ning rakendada geeniuuringute tulemused rahva tervise parandamiseks [1]. Eesti geenivaramu andmebaasiga on liitunud aastatel 2002-2013 ligi 52 000 inimest. Eesti geenivaramu andmebaasi lingitakse igal aastal ka Eesti haigekassa andmebaasiga ning Tervise Arengu Instituudi surma põhjuste registriga. Seetõttu saab geenidoonorite kohta teada, kas neil on enne või pärast uuringuga liitumist diagnoositud mingeid haigusi, millal diagnoos on pandud või kas geenidoonor on üldse veel elus. Kui geenidoonor on surnud, siis on teada ka surmakuupäev. Just viimane on oluline elukestusanalüüsi läbiviimisel.

Elukestusanalüüsi läbiviimiseks on välja töötatud erinevaid meetodeid. Antud töös kasutatakse elukestuse iseloomustamiseks Kaplan-Meieri graafikuid ning Coxi võrdeliste riskide mudeleid. Tähelepanu pööratakse tsenseerimisele, ajaskaalale ning vasakult tõkestatusele. Töö esimeses peatükis on toodud ülevaade antud töös kasutavatest statistilistest meetoditest. Teises peatükis on kirjeldatud läbi viidud simulatsioonuuringuid ning saadud tulemusi. Kolmandas peatükis on Eesti geenivaramu andmetel läbi viidud analüüsi kirjeldus ning saadud tulemused.

Magistritöö kirjutamiseks on kasutatud programmi LaTeX. Analüüsid on läbi viidud ning tulemused on graafiliselt esitatud statistikapaketi R abil.

Autor tänab magistritöö juhendajaid Krista Fischerit ning Nele Taba rohkete paranduste ja nõuannete eest.

1 Statistiline metoodika

Elukestusanalüüs on statistiliste meetodite klass selliste andmete analüüsiks, mis iseloomustavad aega mingi sündmuse toimumiseni. Elukestusanalüüs töötati välja suremuse analüüsimiseks ning sellel otstarbel kasutatakse seda tänini kõige rohkem. Analüüsi kasutusvaldkond on siiski laiem ja eri valdkondi hõlmav. Analüüsitavateks sündmuseks võib peale surma olla ka inimese vähki haigestumine, riistvara katkiminek, töölt lahtilaskmine, sünd, abielu ja palju muud. [2] Elukestusanalüüsi teiseks eestikeelseks nimetuseks on elulemusanalüüs [3]. Inglise keeles tuntakse antud analüüsimeetodit enamasti nime all *Survival Analysis*, kuid kasutatakse ka teisi nimesid nagu *Failure Time Analysis*, *Duration Analysis* või *Reliability Analysis* [2].

Kestusandmete või elulemusandmete (ingl.k *survival data*) all mõeldakse valimeid, kus iga vaatluse juures on oluline protsessi kestus. Huvi pakub ajavahemik kindlaks määratud algmomendist teatud sündmuse toimumiseni. [2] Seda ajavahemikku kutsutakse elukestuseks (ingl.k *survival time*)[3]. Elulemusandmete analüüsimiseks ei kasutata enamasti standardseid analüüsimeetodeid, kuna vaatlust iseloomustavaks näitajaks on ajavahemiku pikkus, mis ei ole kunagi negatiivne ega saa seetõttu olla normaaljaotusega. Tihtipeale ei ole ka elukestuse jaotus sümmeetriline, vaid pigem paremalt pikema sabaga. Lisaks seisneb selliste andmete eripära selles, et kestusandmed võivad olla tsenseeritud. [4] Elulemusanalüüs on kasutatav mitmetes eluvaldkondades, kuid kuna antud töös on elukestusena vaatluse all inimese eluiga ning sündmusena vaadeldakse inimese surma, siis kasutatakse neid sõnu selles töös sünonüümidena.

Järgnevad statistilise metoodika osa juurde kuuluvad peatükid põhinevad D. Collett'i raamatul "Modelling Survival Data in Medical Research, Third edition" [4], kui ei ole märgitud teisiti.

1.1 Olulisemad definitsioonid ja valemid

Olgu T mittenegatiivne juhuslik suurus, mis kirjeldab subjekti elukestust, olgu t juhusliku suuruse T realiseerunud väärtus ning olgu $F(t)$ ja $f(t)$ vastavalt antud juhusliku suuruse jaotusfunktsioon ja tihedusfunktsioon.

Üleelamisfunktsiooniks $S(t)$ kutsutakse tõenäosust, et elukestus on pikem kui ajamoment t. Teisisõnu pole huvipakkuv sündmus toimunud enne ajamomenti $t \geq 0$:

$$S(t) = P(T > t) = 1 - F(t) = 1 - \int_0^t f(u)\mathrm{d}u, \tag{1}$$

kus $F(t)$ on T jaotusfunktsioon ning $f(t)$ tihedusfunktsioon. Üleelamisfunktsioon $S(t)$ on monotoonselt kahanev funktsioon, mille korral $S(0) = 1$.

Teiseks elukestusanalüüsis oluliseks terminiks on riskifunktsioon. Riskifunktsioon $h(t)$ iseloomustab tõenäosust, et huvipakkuv sündmus toimub lõpmatult väikeses ajavahemikus $[t, t + \Delta]$ tingimusel, et see sündmus ei ole toimunud enne ajahetke t. Täpsemalt omab riskfunktsioon $h(t)$ kuju:

$$h(t) = lim_{\Delta t \to 0}\{\frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}\}. \tag{2}$$

Varasemalt defineeritud üleelamisfunktsioon ja riskifunktsioon on seotud järgnevalt:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\ln S(t)\}.$$

1.2 Eksponentjaotus ja Weibulli jaotus elulemusandmetele

Elukestusanalüüsis on oluline riskifunktsiooni kuju. Seetõttu valitakse ka parameetrilised jaotused elukestuse iseloomustamiseks just riskifunktsiooni kuju järgi.

Kui elukestus T on eksponentjaotusega $T \sim Exp(\lambda)$, siis üleelamisfunktsioon ja riskifunktsioon saavad piirkonnas $0 \leq t \leq \infty$ kuju:

$$\begin{aligned} S(t) &= e^{-\lambda t} \\ h(t) &= \lambda. \end{aligned} \tag{3}$$

Eksponentjaotuse korral on riskifunktsioon konstantne iga ajahetke korral, sõltumata eelnevatest ajahetkedest. Keskmine elukestus on λ^{-1} ning mediaan $t(50) = \frac{1}{\lambda}\ln 2$. Eksponentjaotuse juures on oluline tema mälupuudumise omadus: $P(T > s + t \mid T > s) =$

$P(T > t)$ iga $s, t > 0$ korral [3].

Tihtipeale pole konstantse riski eeldus täidetud, seetõttu saadakse riskifunktsioonile üldisem kuju Weibulli jaotust kasutades. Weibulli jaotusel on kaks parameetrit: $\lambda > 0$ ehk skaalaparameeter ja $\gamma > 0$ ehk kujuparameeter. Kui juhuslik suurus T on Weibulli jaotusest ehk $T \sim W(\lambda, \gamma)$, siis tema üleelamisfunktsioon ja riskifunktsioon on kujul:

$$\begin{aligned} S(t) &= \exp(-\lambda t^{\gamma}) \\ h(t) &= \lambda\gamma t^{\gamma-1}. \end{aligned} \tag{4}$$

Keskväärtus Weibulli jaotusega juhuslikul suurusel on $E(T) = \lambda^{-1/\gamma}\Gamma(\gamma^{-1} + 1)$, kus $\Gamma(x)$ on gamma funktsioon, ning mediaan $t(50) = \{\frac{1}{\lambda}\ln 2\}^{1/\gamma}$.

Vahel kasutatakse ka teistsugust parametrisatsiooni Weibulli jaotuse kirjeldamiseks, kus parameetriteks on a ja b nii, et $\gamma = a$ ja $\lambda = (1/b)^a$. Sellisel juhul avalduvad üleelamis- ja riskifunktsioon kujul:

$$\begin{aligned} S(t) &= \exp(-(t/b)^a) \\ h(t) &= (1/b)^a a t^{a-1}. \end{aligned} \tag{5}$$

Niisugust parametrisatsiooni kasutatakse näiteks tarkvara R funktsioonides *dweibull, pweibull, qweibull,* ja *rweibull,* mille abil saab leida Weibulli jaotuse tihedus- ja jaotusfunktsiooni väärtuseid, arvutada kvantiile ja genereerida juhuslikke arve. [5]

1.3 Tsenseerimine

Öeldakse, et vaatlus on tsenseeritud, kui huvipakkuvat sündmust ei ole teatud vaatluse korral vaadeldud. Tsenseerimise põhjuseks võib olla nii informatsiooni puudumine (näiteks uuritav lahkub katsest) kui ka katse lõppemine enne sündmuse toimumist antud indiviidil. Kõige sagedamini esineb paremalt tsenseerimist. Paremalt tsenseerimisega on tegemist juhul, kui uuritav ajavahemik lõppeb enne, kui huvipakkuv sündmus toimub. Teised tsenseerimistüübid on vasakult tsenseerimine (vastupidine paremalt tsenseerimisele) ja intervall-tsenseerimine (samaaegselt nii paremalt kui vasakult tsenseeritud).

[2] Järgnevalt kirjeldatakse täpsemalt paremalt tsenseerimist, kuna see tsenseerimistüüp on selles töös üks peamisi huviobjekte.

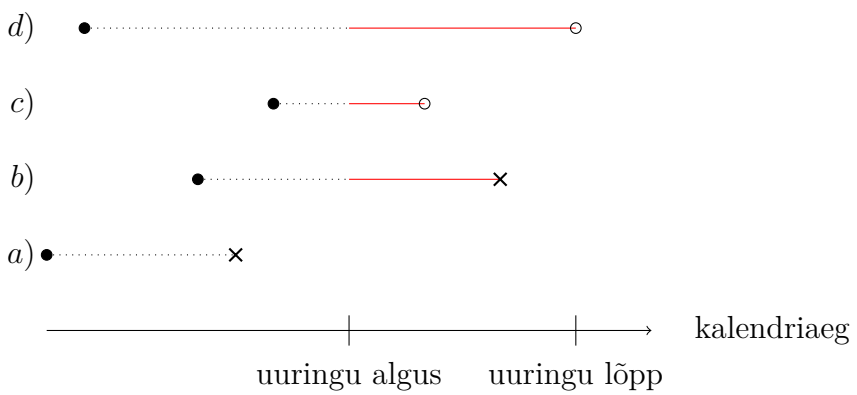
Olgu T_i ja C_i juhuslikud suurused, kus T_i on i-nda vaatluse elukestus ja C_i on i-nda vaatluse tsenseerimisaeg ehk viimane teadaolev ajahetk vaatluse kohta. Iga vaatluse jaoks vaadeldakse paari (Y_i, δ_i), kus

$$Y_i = min(T_i, C_i) \text{ ja } \delta_i = \begin{cases} 1, \text{ kui } T_i \leq C_i \\ 0, \text{ kui } T_i > C_i. \end{cases}$$

Mitmete statistiliste meetodite juures tehakse eeldus, et tegelik sündmuse aeg ja tsenseerimisaeg oleksid sõltumatud ehk tegu oleks mitteinformatiivse tsenseerimisega. See tähendab, et teades vaatluse tsenseerimisaega, ei saa vaatluse kohta öelda muud, kui et tegelik eluaeg oli tsenseerimisajast suurem, $C_i < T_i$. [6]

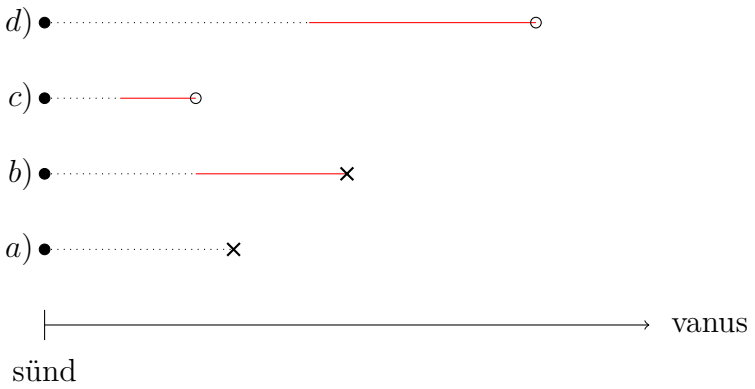
1.4 Ajatelje valik

Ajatelje valik võib meditsiinilistes uuringutes olla keeruline ning vale ajatelje kasutamine võib viia nihkega hinnanguteni. Ajaskaalaks saab kasutada nii subjekti vanust, kalendriaega kui ka aega mingist teisest sündmusest alates. Enamasti tuleks valida üks ajatelg, kuigi võimalik on mudelitesse lisada tunnuseid ka teiste ajaskaalade alusel. [2] Näiteks on ajaskaalaks uuringus oldud aeg, aga tunnusena võetakse arvesse ka vanust või uuringuga liitumise aastat. Ajatelje valik on oluline, kuna see määrab, kellega uuritavaid võrreldakse. Kui kasutatavaks ajateljeks on kalendriaeg, siis võrreldakse igat uuritavat nendega, kes on samal ajal uuringus. Kui kasutatavaks ajatelje nullpunktiks on uuringu algus, siis igal ajahetkel võrreldakse uuritavat teiste uuritavatega, kes on sama kaua uuringus olnud.



Joonis 1: Eluigade kujutamine kalendriajas, punktiirjoonega märgitud uuringule eelnev aeg, punase pidevjoonega uuringus oldud aeg, ringiga märgitud tsenseerimise hetk ja ristiga sündmuse toimumise hetk

Kui ajateljeks on vanus, siis võrreldakse uuritavaid teiste samaealiste inimestega. Kui kliinilistes uuringutes võib olla mõistlik määrata ajaskaala alguspunktiks haiguse diagnoosimine või uuringusse randomiseerimine [2], siis epidemioloogiliste uuringute puhul on soovitatavaks ajaskaalaks inimese vanus [7].



Joonis 2: Eluigade kujutamine vanuse skaalal, punktiirjoonega märgitud uuringule eelnev aeg, punase pidevjoonega uuringus oldud aeg, ringiga märgitud tsenseerimise hetk ja ristiga sündmuse toimumise hetk

Joonisel 1 ja 2 on toodud neli vaatlust, kus eluead on tähistatud joonega nii, et uuringule eelnev aeg on punktiirjoonega ning uuringus olevat aega tähistav lõik on punase täisjoonega. Joonisel 1 on ajaskaalaks kalendriaeg ning joonisel 2 vanus. Vaatluste a) ja b) korral on ristiga märgitud sündmuse toimumise hetk ning vaatlustel c) ja d) on ringiga märgitud tsenseerimise hetk. Seejuures on vaatlus c) tsenseeritud, kuna eemaldus

uuringust ning vaatlus d) on tsenseeritud, kuna uuring sai läbi ning sündmust polnud selleks hetkeks veel toimunud.

1.5 Vasakult tõkestatus

Ajaskaalana vanust kasutades tekib olukord, kus teatud vanuses pole osa inimestest veel vaatluse all, samas on osad inimesed juba vaatluse alt väljas. Seetõttu on võrdlusgrupis olevate inimeste arv vanusegrupiti erinev. Arvestama peab ka tõsiasja, et uuringusse said jõuda vaid need inimesed, kes olid elus uuringu alguses. Viimast olukorda nimetatake vasakult tõkestatuseks ning illustreerib joonisel 1 vaatlus a) [7]. Vasakult tõkestatuse mittearvestamine võib viia nihkega hinnanguteni [6]. Kui ajaskaalaks on vanus ning vasakult tõkestatust võetakse arvesse, siis igat inimest võrreldakse sama vana inimesega, kes olid ka uuringus selles vanuses.

1.6 Kaplan-Meieri hinnang

Kaplan-Meieri hinnangut kasutatakse üleelamisfunktsiooni hindamiseks. Olgu vaadeldud valimit, kus vaatlused pole tsenseeritud. Sellisel juhul on üleelamisfunktsiooni $S(t)$ hinnanguks $\hat{S}$:

$$\hat{S}(t) = \frac{\text{Inimeste arv, kellel } T \geq t}{\text{Inimeste arv andmestikus}}.$$

Üleelamisfunktsiooni hindamine valimis, kus on ka tsenseeritud vaatlused, toimub järgnevalt. Olgu k erinevat sündmuse toimumise aega, $t_1 < t_2 < \ldots < t_k$. Märkigu n_j vaatlusi, mis on ajahetkel t_j riskigrupis ehk antud vaatlustel ei ole toimunud sündmust ega tsenseerimist enne ajamomenti t_j. Vaatlused, mis tsenseeritakse täpselt ajahetkel t_j, loetakse ka tolle ajahetke riskigruppi. Märkigu d_j vaatluste arvu, mille korral toimub sündmus vaadeldaval ajahetkel t_j. Kaplan-Meieri hinnang elulemusfunktsioonile on siis defineeritud järgnevalt:

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

kus $t_1 \leq t \leq t_k$. Enne esimese sündmuse toimumist on elulemusfunktsiooni hinnang defineeritud 1-ks, $\hat{S}(t) = 1$, kui $t < t_1$. Ajahetkede $t > t_k$ jaoks, jaguneb $\hat{S}_t$ kaheks. Kui

ei leidu ühtegi tsenseeritud ajahetke, mis oleks suurem kui t_k, siis $\hat{S}(t) = 0$ kui $t > t_k$. Kui aga leidub tsenseeritud ajahetki ka pärast viimse sündmuse toimumist, siis $\hat{S}(t)$ ei ole pärast viimast tsenseerimise hetke defineeritud.

1.7 Võrdeliste riskide mudel elukestusele

Vaatluse elukestuse ja muude kirjeldavate tunnuste vaheliste seoste uurimiseks kasutatakse võrdeliste riskide mudeleid.

Olgu esmalt kaks gruppi ja olgu $h_1(t)$ ja $h_2(t)$ vastavad riskifunktsioonid ajahetkel t. Siis põhineb võrdeliste riskide mudel eeldusel, et riskide suhe kahes grupis on ajas konstantne ehk

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

kus ψ on konstant ning ajahetk $t \geq 0$. Konstanti ψ nimetatakse riskisuhteks.

Olgu tegemist n vaatlusega ning olgu $h_i(t), i = 1, 2, \ldots, n$ riskifunktsioon i-nda vaatluse jaoks. Olgu $X_1, X_2, \ldots, X_p$ ajas mittemuutuvad argumenttunnused ning neile vastavad väärtused $x_1, x_2, \ldots, x_p$, märkigu $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. Baasriskifunktsiooniks olgu $h_0(t)$, mille puhul $x_l = 0$ iga $l \in (1, .., p)$ korral. Siis saab i-ndale vaatlusele vastava riskifunktsiooni kirjutada kujul

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

kus $\mathbf{x}_i$ on i-ndale vaatlusele vastav argumenttunnuste vektor ning $\psi(\mathbf{x}_i)$ on funktsioon. Kuna riskisuhte vektor ei saa olla negatiivne, siis saab kirjutada, et $\psi(\mathbf{x}_i) = e^{\eta_i}$, kus $\eta_i = \boldsymbol{\beta}'\mathbf{x}_i$ ning $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ on argumenttunnuste regressioonikordajad.

Kokkuvõttes on võrdeliste riskide mudeli kujuks:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}}. \tag{6}$$

Parameetriliste võrdelise riski mudelite korral on $h_0(t)$ määratud jaotuse parameetritega. Poolparameetrilisel juhul jäetakse $h_0(t)$ määramata ja hinnatakse ainult parameetrid $\boldsymbol{\beta}$.

1.8 Weibulli jaotusega võrdeliste riskide mudel

Weibulli jaotusega võrdeliste riskide mudeli korral eeldatakse, et elukestused T on Weibulli jaotusega. Selle mudeli eelduste kohaselt on kujuparameeter γ kõigil indiviididel ühesugune, kuid skaalaparameeter λ_i sõltub argumenttunnuste väärtustest i-ndal indiviidil. Seega avaldub riskifunktsioon kujul

$$h_i(t) = \lambda_i \gamma t^{\gamma-1},$$

kusjuures eeldatakse, et

$$\lambda_i = e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}} = e^{\beta_0} \cdot e^{\beta_1 x_{i1}} \cdot \ldots \cdot e^{\beta_p x_{ip}}.$$

Seega avaldub baasriskifunktsioon kui

$$h_0(t) = \lambda_0 \gamma t^{\gamma-1},$$

kus $\lambda_0 = e^{\beta_0}$. Viimaste võrdluste põhjal saab öelda, et tegu on võrdeliste riskide mudeliga, sest argumendi x_{ij} suurenemisel ühe ühiku võrra muutub $h_i(t)$ väärtus e^{β_j} korda. Seejuures eeldati, et kasutati võrdustes (4) toodud parametrisatsiooni. [6]

Kui kasutada parametrisatsiooni (5), on näha, et parameeter a on kõigil indiviididel sama, kuid $b = b_i$, mis $\beta_j^* = -\frac{\beta_j}{a}$ korral avaldub kui

$$b_i = e^{\beta_0^*+\beta_1^* x_{i1}+\ldots+\beta_p^* x_{ip}}.[6]$$

Tarkvara R funktsiooni *survreg* abil hinnataksegi parameetrid $\beta_0^* \ldots, \beta_p^*$ ning parameetri *scale*, mis hindab tegelikult kujuparameetri pöördväärtust, $s = \frac{1}{a}$. [5]

1.9 Coxi võrdeliste riskide mudel

Coxi võrdeliste riskide mudelite korral on tegu poolparameetriliste mudelitega, sest baasriskifunktsioon jäetakse hindamata ning hinnatakse ainult argumenttunnuste kordajate väärtused.

Iga kahe vaatluse x_a ja x_b korral ei sõltu riskide suhe baasriskist ega ajast, vaid ainult argumenttunnuste väärtustest:

$$\frac{h(t|\mathbf{x}_a)}{h(t|\mathbf{x}_b)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_a)}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_b)} = \exp(\boldsymbol{\beta}'(\mathbf{x}_a - \mathbf{x}_b)).$$

Coxi võrdeliste riskide mudeli tõepärafunktsiooni kirjapanekuks tuuakse uuesti välja kasutatavad märgistused. Olgu n vaatlust ning k toimunud sündmust, mis ajaliselt on järjestatud $t_1 < t_2 < \ldots < t_k$, ning kus j-s sündmuse toimumise aeg on t_j. Olgu ajahetkel t_j riskigrupp $R(t_j)$, kuhu kuuluvad need vaatlusalused, kes on hetk enne vaadeldavat ajahetke veel vaatluse all. Sel juhul on osalise tõepära funktsioon Coxi mudeli jaoks

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{\psi(\mathbf{x}_j, \boldsymbol{\beta})}{\sum_{l \in R_{(t_j)}} \psi(\mathbf{x}_l, \boldsymbol{\beta})} = \prod_{j=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_j)}{\sum_{l \in R_{(t_j)}} \exp(\boldsymbol{\beta}'\mathbf{x}_l)}, \tag{7}$$

kus $\boldsymbol{x}_j$ on selle vaatlusaluse argumenttunnused, kelle surmaaeg t_j on toimumisaja järgi järjekorras j-s. Suurust (7) kutsutakse osalise tõepära funktsiooniks, kuna lugejas vaadatakse ainult vaatlusi, mille korral toimub sündmus, mitte tsenseerimine. Tõepärafunktsiooni (7) korral on oluline märgata, kuidas on lugeja ehk riskigrupp defineeritud, kuna riskigruppi kuuluvad vaatlused sõltuvad ajaskaala valikust, nagu on kirjeldatud peatükis 1.4.

1.9.1 Coxi võrdeliste riskide mudel ajast sõltuva muutuja korral

Siiani kirjeldatud mudelites on kõigi argumenttunnuste väärtused fikseeritud hetkel, kui indiviid liitub uuringuga, hoolimata sellest, et mõned neist (kehakaal, mingi haiguse esinemine) võivad jälgimisaja jooksul muutuda. Mõnikord võib vaja minna ka ajast sõltuva tunnuse lisamist mudelisse. Ajast sõltuv tunnus võib olla näiteks mingisse haigusesse haigestumine, mis võib inimese edasist riski muuta. Kui lisada valemisse (6) ajast sõltuv muutuja $x_{ij}(t)$ ning olgu γ_j koefitsient ajast sõltuvale tunnusele, siis saab mudel kuju:

$$h_i(t) = h_0(t)\exp(\sum_{j=1}^{p1} \beta_j x_{ij} + \sum_{j=1}^{p2} \gamma_j x_{ij}(t)), \tag{8}$$

kus $p1$ on ajast mittesõltuvate muutujate arv ja $p2$ on ajast sõltuvate muutujate arv. [6] Näiteks, kui $x_{ij}(t) = 0$ enne ajahetke τ_i ja $x_{ij}(t) = 1$ pärast seda ajahetke, siis see tähendab, et alates hetkest τ_i muutub indiviidi risk e^{γ_j} korda. Parameetritele hinnangute arvutamise teeb keerukaks see, et ajas muutuvate argumentide väärtusi on vaja teada kõikide riskigrupis olevate vaatluste jaoks igal ajamomendil kui sündmus toimub [3].Üks võimalus analüüsimaks sellist mudelit on viia andmestik kujule, kus vaadatakse ajavahemikke, mille sees ajas muutuv tunnus on konstantne.

1.10 Elukestusanalüüsi meetodite rakendamine tarkvara R abil

Tarkvaraga R tuleb elukestusanalüüsi läbiviimiseks kasutada paketti *survival.* Tsenseeritud andmete käsitlemiseks luuakse objekt *Surv(time,event)*, millel on tsenseeritud andmete korral kaks komponenti: *time* - vaatluse all oldud aeg ning *event* - sündmuse toimumise indikaator(1 - sündmus vaadeldi, 0 - indiviidi elukestvus oli tsenseeritud). Kui tegeletakse vasakult tõkestatud andmetega, siis peab kasutama kolme parameetrit *Surv(*$time_1$*,* $time_2$*, event): time*$_1$ - aeg, millal inimene uuringusse tuli, $time_2$ - uuringu lõpu aeg inimesele ning sündmuse toimumise indikaator *event*. [8]

Kaplan-Meieri hinnangu üleelamisfunktsioonile saab leida funktsiooni *survfit(formula type='kaplan-meier')* abil. Etteantava valemi võimalik kuju on *Surv(...)*$\sim x_1+..+x_p$, kus paremale poole operaatorit $\sim$ tuleb lisada argumenttunnused ning vasakul on elukestust ja sündmust kirjeldav *Surv* objekt. Näiteks kui tahetakse leida hinnangut grupi lõikes, siis tuleb lisada grupi tunnus g : *survfit(Surv(...)* $\sim g$*, type='kaplan-meier').* [8]

Coxi võrdeliste riskide mudelit saab leida funktsiooni *coxph(formula, data,...)* abil. Etteantav valem on samal kujul kui eespool mainitud *survfit()* funktsioonis. [8]

2 Simulatsioonuuringud

Antud töö koostamise käigus viidi läbi mitmeid simulatsioone, mille jaoks genereeriti andmed nii, et need sarnaneksid võimalikult palju ka edasises analüüsis kasutatava Eesti geenivaramu andmestikuga. Simulatsioonuuringus on kaks osa, millest esimeses vaadatakse Coxi võrdeliste riskide mudeli käitumist erinevate ajaskaalade, tsenseerimise ja tõkestatuse korral. Teises osas vaadati täpsemalt, kuidas kasutada Coxi mudelit ajast sõltuva muutuja korral.

2.1 Võrdeliste riskide mudelitele vastavate elukestusandmete genereerimine tarkvara R abil

Hoolimata sellest, et reaalseid andmeid analüüsitakse poolparameetriliste Coxi mudelite abil, on andmete genereerimiseks lihtsam kasutada parameetrilisi jaotusi. Käesolevas töös on esimese osa simulatsioonuuringu jaoks kasutatud Weibulli jaotust, sest see sarnaneb enamasti kõige rohkem reaalsete elukestusandmete jaotusele. Teise osa mudelite keerukuse tõttu on kasutatud kõige lihtsamat, st eksponentjaotust. Weibulli jaotusega juhuslikke arve saab genereerida kasutades R-i funktsiooni *rweibull(n,a,b)*, kus n on soovitav valimi suurus ning a ja b on jaotuse parameetrid vastavalt parametrisatsioonile (5). Järgnevalt vaadeldakse, kuidas genereerida Weibulli jaotusega elukestused T_i nii, et andmed vastaksid lihtsale võrdeliste riskide mudelile

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

kus ψ_i on etteantud konstant ning $h_1(t)$ ja $h_2(t)$ on riskifunktsioonid gruppides 1 ja 2. Lisaks soovitakse, et T_i jaotus oleks lähedane reaalses andmestikus vaadeldud elukestvusandmetele. Selleks on vaja läbida järgmised sammud:

1. Hinnatakse reaalse andmestiku jaoks sobiva Weibulli jaotuse kujuparameetri, kasutades funktsiooni *survreg*. Selleks kirjutatakse *survreg(Surv(aeg, sündmus)∼ 1, data=andmed, dist="weibull")* kus *aeg* ja *sündmus* on vaatlusaega ja sündmuse vaadeldust tähistavad tunnused. Et see on ainult vabaliikmega mudel, siis vabaliikme (*Intercept*) kordaja $\hat{\beta}_0$ hindab Weibulli jaotuse parameetri b logaritmi (parametrisatsioonis (5)), st $\hat{b} = exp(\hat{\beta}_0)$ ning väljastatud parameeter *scale* ($\hat{s}$)

hindab kujuparameetri pöördväärtust $(1/a)$, st $\hat{a} = 1/\hat{s}$. Siin kasutatakse ainult kujuparameetri hinnangut ja määratakse b järgmise sammu abil.

2. Hinnatakse skaalaparameeter b_i nii, et indiviididel grupist 1 oleks elukestvuse mediaan võrdne etteantud konstandiga m:

$$b_1 = m \cdot (\ln 2)^{1/a}.$$

3. Hinnatakse skaalaparameeteri b_2 grupis 2: $b_2 = b_1 \cdot \psi^{-1/a}$

4. Genereeritakse Weibulli jaotusega elukestvused t: *t= rweibull(n, a,* $b_1 \cdot g + b_2 \cdot (1-g)$*)*, kus g indikaator, mis saab väärtuse 1, kui vaatlus kuulub gruppi 1 ning 0 vastasel juhul.

5. Et simuleerida realistlikke tsenseeritud ja vasakult tõkestatud andmeid, genereeritakse ka sünniaastad ühtlasest jaotusest vahemikus 1920-1990 ning liitumisajad ühtlasest jaotusest vahemikus 2003-2010 ja määratakse nn analüüsi läbiviimise hetk (01.01.2017). Siis leiti iga isiku jaoks:

   (a) kas ta on kaasatud uuringuvalimisse ehk kas surm ei saabunud enne võimalikku liitumisaega

   (b) kas tal toimus jälgimisajal uuritav sündmus ehk kas surmaaeg on enne 2017. aastat

   (c) kui pikk oli jälgimisaeg sellel isikul

Simulatsioonuuringu esimeses osas kasutatud R-i kood on leitav Lisast 4 ning teise osa kood Lisast 5.

2.2 Ajaskaala, tsenseerimise ja tõkestatuse arvestamine ning selle mõju tulemustele

Simulatsioonuuringute esimeses osas oli valimi mahuks 20 000 vaatlust ning simulatsioone korrati 1000 korda. Simulatsioonid viidi läbi nelja erineva ettemääratud riskisuhte koefitsiendi $\psi = e^{\beta}$ korral: 2; 1,2; 1,05 ning 1. Andmete genereerimine toimus järgnevalt:

1. Eesti geenivaramu andmete põhjal hinnati Weibulli jaotuse üks parameetritest, kujuparameeter a =3,5.

2. Leiti Weibulli jaotuse skaalaparameeter b_1 juhul kui inimese eluea mediaan on 55, $b_1 = m \cdot (\ln 2)^{-1/a} = 55 \cdot (\ln 2)^{-1/3,4}$.

3. Genereeriti 20 000 inimese eluead nii, et esimese 10 000 jaotuseks on $T_i \sim W(a, b_1)$ ning teise 10 000 korral $T_i \sim W(a, b_2)$, kus $b_2 = b_1/\psi^{1/a}$.

4. Igale inimesele genereeriti sünniaasta ühtlasest jaotusest vahemikus 1920-1990 ning uuringuga liitumisaasta ühtlasest jaotusest vahemikus 2003-2010.

5. Analüüsi tegemise ajaks määrati 1. jaanuar 2017, mille põhjal määrati tsenseerimine.

Lisas 1 leitavas Tabelis 8 on toodud protsendiliselt liitumiseni elanud inimesed, uuringus surnud inimesed ning tsenseeritud inimeste osakaal. Coxi võrdeliste riskide mudelid hinnati viie erineva juhu jaoks ning vaadati, kas hinnang kahe grupi (esimesed 10 000 ja teised 10 000 vaatlust) vahelisele riskisuhtele on nihketa, st kas $E(\hat{\psi}) = \psi$, kusjuures keskmist $E(\hat{\psi})$ hinnati 1000 simulatsiooni keskmisena. Lisaks vaadati keskmist ruutviga, usaldusintervalli katvust ning mudeli võimsust. Mudelid olid järgnevad:

- Mudel 1 (“teoreetiline”) - kasutatakse inimese eluiga sünnist surmani, ilma tsenseerimiseta mudel, ajaskaalaks on inimese vanus ning vasakult tõkestatust ei esine.

- Mudel 2 (“tsenseeritud”) - paremalt tsenseeritud mudel. Kui inimene oli elus aastal 2017, siis tema tsenseeriti ning jälgimisajaks on vanus tsenseerimise hetkel. Kui inimene oli selleks hetkeks surnud, siis vaadati kogu eluaega. Ajaskaalaks on inimese vanus, vasakult tõkestatust ei esine.

- Mudel 3a (“vale”) - Andmestikust eemaldati inimesed, kes surid enne liitumisaega, kuid mudelit sellekohaselt ei korrigeerita (ehk vasakult tõkestatust arvesse ei võeta, kuigi see on andmetes olemas). Ajaskaalaks on inimese vanus ning arvestatakse paremalt tsenseerimist.

- Mudel 3b (“vale 2”) - Sarnane eelneva mudeliga, kuid ajaskaalaks kasutati uuringus oldud aega.

- Mudel 4 (“õige”)- Arvestatakse nii vasakult tõkestatust kui ka paremalt tsenseerimimist, ajaskaalaks on inimese vanus.

Mudelite valikul lähtuti sellest, et näha hinnangu käitumist erinevate aspektide, nagu tsenseerimine, ajaskaala ja tõkestatus, lõikes. Mudelite 3a ja 3b erinevus on vaid ajaskaalas, kusjuures mudelis 3b kasutatud ajaskaala, ehk uuringus oldud aeg, on kliinilistes uuringutes kõige laialdasemalt kasutatav ajaskaala. Mudelis 3a kasutatud ajaskaala ehk subjekti vanus on soovitatud just epidemioloogiliste uuringute juures.

Tabel 1: Simulatsioonide tulemused erineva riskide suhte ψ korral

	mudel	$\hat{\boldsymbol{E}}(\hat{\psi})$	**keskmine ruutviga**	**95% usaldus-vahemiku katvus**	**võimsus**
	1 - teoreetiline	1,999	0,0009	0,939	1
	2 - tsenseeritud	2,001	0,0013	0,947	1
ψ=2	3a - vale	2,539	0,3003	0	1
	3b - vale 2	1,394	0,3697	0	1
	4 - õige	1,999	0,0060	0,943	1
	1 - teoreetiline	1,200	0,0003	0,943	1
	2 - tsenseeritud	1,201	0,0005	0,947	1
ψ=1,2	3a - vale	1,274	0,0078	0,642	1
	3b - vale	1,096	0,0126	0,315	0,68
	4 - õige	1,203	0,0020	0,956	0,998
	1 - teoreetiline	1,049	0,0002	0,952	0,931
	2 - tsenseeritud	1,049	0,0004	0,952	0,731
ψ=1,05	3a - vale	1,065	0,0017	0,938	0,368
	3b - vale	1,024	0,0020	0,911	0,085
	4 - õige	1,049	0,0015	0,955	0,237
	1 - teoreetiline	1,000	0,0002	0,945	0,055
	2 - tsenseeritud	1,000	0,0004	0,940	0,060
ψ=1	3a - vale	1,001	0,0014	0,947	0,053
	3b - vale	1,001	0,0015	0,949	0,051
	4 - õige	1,001	0,0015	0,945	0,055

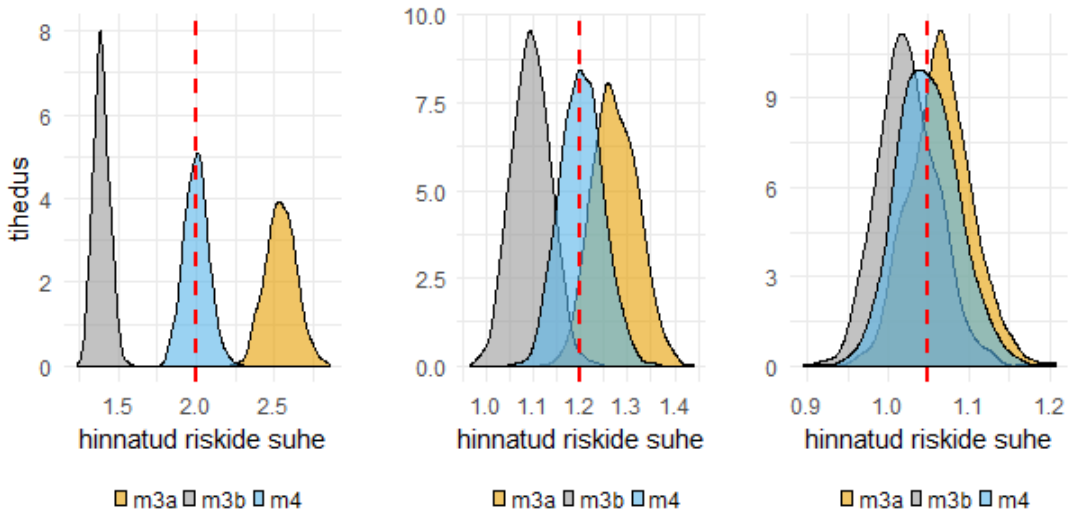
Simulatsioonide esimese osa tulemused on toodud tabelis 1. Esimene ja teine mudel ehk “teoreetiline” ning “tsenseeritud” mudel andsid oodatavalt õiged riskisuhted.

Samas ei esine tihti uuringuid, kus kõigi valimisse kuuluvate inimeste eluead on teada. Uuringud, kus ainult paremalt tsenseerimist peab arvesse võtma on näiteks kliinilised uuringud (sh ravimiuuringud), kus valim koostatakse teatud diagnoosiga indiviididest ning jälgimisaeg algab diagnoosi hetkest. Epidemioloogilistes uuringutes, kuhu kaasatakse terveid täiskasvanud, on aga vasakult tõkestatus tihti vältimatu. Mudelite 3a ja 3b korral eemaldati andmestikust inimesed, kes surid enne võimalikku liitumisaega. Samas ei ole mudeleid 3a ja 3b vasakult tõkestatuse ehk inimeste eemaldamise suhtes

korrigeeritud. Mudeli 3a korral on elukestusena kasutatud inimese vanust ning mudeli 3b korral on kasutatud inimese uuringus oldud aega. Mõlemad "valed" mudelid andsid nihkega hinnangu, esimesel juhul oli nihe ülespoole, teisel hinnati riskisuhet tegelikust väiksemaks. Mudelis 4 on arvesse võetud nii vasakult tõkestatust kui ka paremalt tsenseeritust, ajaskaalaks on inimese vanus. Antud mudeli korral on hinnang riskide suhtele nihketa. Samas on võimsus viimase mudeli korral väiksem kui "vale" mudeli korral, eriti just juhul, kui riskide suhe on väike.

Mudelite võimsuste erinevused tulevad kõige paremini välja väikese ψ väärtuse korral. On näha, et "õige" mudeli korral on võimsus 0,237 aga parema võimsuse 0,368 annab "vale" mudel.

Kui riskide suhe määrata 1-ks ehk kahel grupil erinevust ei ole, siis annavad kõik mudelid nihketa hinnangu. See on oluline teadmine edaspidiseks, kuna siis saab riskide erinevuste olemasolu kindlaks määramiseks kasutada ka mudeleid, mis tekitavad nihkega hinnangu, kuid on parema võimsusega, näiteks "vale" mudel 3a-d. Nihe tekib hinnangule ainult juhul, kui riskide suhe ei ole üks.



Joonis 3: Tihedusgraafikud hinnatud ψ-le, õige väärtus on tähistatud punase katkendliku joonega; m3a on "vale" mudel, m3b "vale 2" mudel, m4 on "õige" mudel.

Joonisel 3 on toodud tihedusgraafikud kolme ψ väärtuse ja kolme mudeli korral. Õiged ψ väärtused on kujutatud vertikaalse punase joonega. Ka siit on näha, et "valed" mudelid annavad nihkega hinnanguid. Kuid samas on siit ka näha, et mudelil 3a-l on ühe lähedal

või sellest vasakul kõige väiksem osa, seega tuleb selle mudeli võimsus kõige suurem.

Simulatsioonuuringute esimese osa tulemusena selgus, et väga tähtis on elukestusanalüüsi juures mõelda nii tsenseerimise, ajaskaala kui ka tõkestatuse peale. Kokkuvõttes annab kõige täpsemaid tulemusi vasakult tõkestatud andmete korral "õige" mudel ning seda ka kasutatakse peamiselt ka edasises analüüsis. Edaspidiseks jäetakse siiski meelde, et "vale" mudel andis parema võimuse, kuid nihkega hinnangu.

2.3 Ajas muutuvad riskifaktorid ning nende kasutamine argumenttunnustena

Simulatsioonuuringu teises osas kasutati eluea simuleerimiseks eksponentjaotust. Eesmärgiks oli simuleerida olukorda, kus inimese eluiga sõltub mingist teatud tegurist, mis võib muutuda inimese eluea jooksul. Täpsemalt eeldatakse, et indiviidi risk ajahetkel t sõltub tunnuse $x_i(t)$ väärtusest vastavalt järgmisele mudelile:

$$h_i(t) = \psi^{x_i(t)} \cdot h_0(t),$$

kus tunnus $x_i(t)$ vastab sellele, kas indiviid on saanud ajahetkeks t teatud kroonilise haiguse diagnoosi, ehk:

$$x_i(t) = I[t \geq t^*],$$

kus t_i^* on diagnoosi saamise aeg indiviidil i. Inimeste arvuks valiti seekord 100 000 inimest, simulatsioone korrati 1000 korda ning kolme erineva riskisuhte ψ jaoks (2, 1,2 ning 1,05).

Järgnenud andmete genereerimisel kasutati loogikat:

1. Genereeriti t^* normaaljaotusest, $t^* \sim N(55, 8)$, kasutades R-i funktsiooni *rnorm*.
2. Genereeriti igale indiviidile eluiga T_1, mis vastab olukorrale, kus $x_i(t) = 0$, ehk indiviidil ei teki või ei ole veel tekkinud uuritavat haigust. Eeldatakse, et T_1 on eksponentjaotusega parameetriga $\lambda_1 = 1/70$ ja seetõttu kasutame R-i funktsiooni *rexp*. Seega $h_0(t) = \lambda_1$.

3. Kõigile inimestele leiti ühtlasest jaotusest tõenäosus, et tal tekib riskitegur, tingimusel, et ta elas ajahetkeni t^*, $p_i = P(x_i(t) = 1, \text{ kui } t^* < T_1)$.

4. Järgnevalt on vaja genereerida eluead indiviididele, kellel $t^* < T_1$, ning $p_i > 0{,}5$ ehk siis need, kes said riskiteguri (nt diabeedi diagnoosi). Nende jaoks on elukestus leitav kui

$$T_2 = t^* + \Delta T_2,$$

kus ΔT_2 on määratud diagnoosijärgse riski poolt. Eksponentjaotuse "mälupuudumise" omaduse tõttu saadakse:

$$P(T_2 > t | T_2 > t^*) = P(\Delta T_2 > t - t^*) = e^{-\lambda_2 (t - t^*)},$$

kus λ_2 vastab indiviidi diagnoosijärgsele riskile. Seega, eeldades, et $\lambda_2 = \psi\lambda_1$, genereerime eksponentjaotusega diagnoosijärgse eluea ΔT_2, võttes parameetriks λ_2. Seega defineeriti nende isikute elueaks $t^* + \Delta T$.

5. Igale inimesele genereeriti sünniaasta ühtlasest jaotusest vahemikus 1920-1990 ning liitumisaasta ühtlasest jaotusest vahemikus 2003-2010.

6. Samuti eeldati, et andmed on teada kuupäeva 01.01.2017 seisuga, st pärast seda surnud isikud loeti tsenseerituks.

Ollakse teadlikud, et viimati genereeritud andmestik ei peegelda täielikult Eesti Geenivaramu andmestikku, sest inimeste eluead ei ole tegelikkuses eksponentjaotusega. Kuid selliselt genereeritud andmestiku abil saab siiski vaadata ajas muutuvate riskifaktorite kasutamist ning leitud hinnangute täpsust.

Esimese mudeli korral oli eesmärgiks uurida, kuidas töötab mudel, kui vaadata riskiteguri olemasolu ainult liitumishetkel. Inimesed jagati kahte gruppi. Kui inimesele genereeritud tõenäosus riskiteguri muutumiseks oli suurem kui 0,5 ning ta elas aastani t^* ning uuringuga liitumine toimus pärast riskiteguri saamist, lisati ta gruppi 2; ülejäänud inimesed lisati gruppi 1. Gruppi 1 kuuluvatel inimestel kasutati elueana esimesena genereeritud suurust. Teise gruppi kuuluvate inimeste korral oli eluiga alates hetkest t^* kõrge riskiga ehk teisena genereeritud suurus. Liitumise hetkel oli riskitegur keskmiselt 16-20% liitujatel. Täpsemalt on keskmised näitajad II osas simuleeritud andmete kohta toodud Lisas 1 leitavas Tabelis 8. Analüüsimisel kasutati vanuseskaalat ning arvestati

paremalt tsenseerimist ning vasakult tõkestatust. Tulemustest Tabelis 2 on näha, et see mudel suutis riskisuhted väga hästi ära hinnata.

Teise mudeli juures võeti arvesse ka olukorda, kus inimesel sai tekkida riskitegur uuringus oldud aja jooksul. Analüüsimaks sellist olukorda, kasutati andmetel formaati, kus inimese eluiga jaotatakse mitmeks aja lõiguks, kus ühe ajavahemiku sees on inimesel ainult üks risk. Seega on inimeste kohta, kellel ei teki üldse riskitegurit, või inimeste kohta, kelle oli riskitegur olemas juba uuringusse saabumise hetkel, andmestikus üks rida nagu varasemalt. Kuid inimestel, kellel tekib uuringus olemise hetkel riskitegur, vaadatakse eluiga kahes perioodis ning nende kohta on andmestikus kaks rida. Esimeses reas on eluiga kuni riski tekkimise hetkeni ning see vaatlus tsenseeritakse. Teises reas on inimesel jälgimisaja alguspunktiks mitte enam uuringusse astumise vanus, vaid vanus, kui ta sai riskiteguri, ning lõpp-punktiks on kas surmavanus või tsenseerimisvanus uuringu lõppedes. Teise rea korral kirjeldab eluiga kõrge riskiga genereeritud suurus eksponentjaotusest. Näide selliselt simuleeritud andmestikkust on toodud Lisas 2.

Tabel 2: Simulatsioonide tulemused erineva riskide suhte ψ ja ajast sõltuva muutuja korral

	mudel	$\hat{\boldsymbol{E}}(\hat{\boldsymbol{\psi}})$	**keskmine ruutviga**	**95% usaldus-vahemiku katvus**	**võimsus**
ψ=2	mudel1	2.005	0,0032	0.951	1
	mudel2	1.999	0,0024	0.963	1
ψ=1,2	mudel1	1.198	0,0014	0.946	1
	mudel2	1.198	0,0011	0.945	1
ψ=1,05	mudel1	1.051	0,0011	0.951	0,363
	mudel2	1.050	0,0009	0.948	0,412

Tulemused tabelis 2 näitavad, et nii riskiteguri kui konstandi ning riskiteguri kui ajast sõltuva tunnuse kaasamisel tuleb riskide suhte hinnang nihketa. Võimsus tuli parem teise mudeli korral ning selle korral oli ka keskmine ruutviga natuke väiksem. Kuigi hetkel ei tundu olevat suurt erinevust kahe viimase mudeli korral, siis tuleb meeles pidada, et simuleeriti olukorda, kus subjektid on vaatluse all olnud lühikest aega. Olukorda, kus uuring oleks pikem ning seetõttu ka rohkem indiviide saaksid riskiteguri uuringus olemise ajal, väärib autori hinnangul edasist uurimist järgmistes uurimistöödes.

3 Eesti Geenivaramu andmete analüüs

Tartu Ülikooli genoomika instituudi alla kuuluva Eesti geenivaramu andmebaasis on 2017. aasta seisuga üle 51 000 geenidoonori andmed. Eesti geenivaramuga sai liituda aastatel 2002-2013. Igal aastal lingitakse Eesti geenivaramu andmebaasi juurde info Eesti Haigekassa andmetest haiguste kohta ning Tervise Arengu Instituudi surma põhjuste registrist inimeste surmade ja surmakuupäevade kohta. Analüüsis kasutatav andmestik lingiti viimati 2017. aasta keskel.

Geenidoonori elukestuse määramisel kasutatakse tema vanust liitumisel ning jälgimisaja lõpus. Jälgimisaja lõpuks on kas inimese surm või tsenseerimine 2017. aasta seisuga, kumb mainitutest toimus varem. Lisaks on geenidoonori kohta teada tema liitumisaasta, sugu, haridustase, suitsetamisstaatus ning kehamassiindeks (edaspidi ka KMI). Haigekassa andmebaasiga linkimise tulemusena on geenidoonori kohta teada ka see, kas tal on erinevaid haigusi või meditsiinilisi häireid. Täpsemalt on teada, kas geenidoonoril on diagnoositud diabeet (I või II tüüp) liitumisel või on see hiljem avaldunud, kas tal on olnud infarkti või insult. Geenidoonoritele tehtud ülegenoomse assotsiatsiooniuuringu käigus leiti ka mitmed riskiskoorid [9]. Üheks selliseks on riskiskoor diabeedile, mis on ka antud töös kasutusel.

Diabeet on energiaainevahetuse häire, mille puhul ei tooda kõhunääre piisavalt insuliini, insuliini toime on nõrgenenud ja/või eritumine puudulik. Insuliin on eluks hädavajalik hormoon, mis tekib kõhunäärmes ning aitab omastada toitaineid. [10] Häiritud energiaainevahetus väljendub vere suurenenud suhkrusisaldusena. Olenevalt tekkepõhjusest rühmitatakse diabeeti erinevateks vormideks; sagedaseimad on I ja II tüüpi diabeet. I tüüpi diabeedi põhjuseks on insuliini tootvate beetarakkude hävimine põletiku tulemusel. II tüüpi diabeet avaldub insuliini mõju nõrgenemisena ja/või insuliini eritumise häirena. [11] Diabeet ning veresuhkru ebaregulaarsed näitajad soodustavad ka silma-, südame-, veresoonkonna- või närvihaiguste teket. [10] Kuna II tüüpi diabeeti põevad enamasti vanemad ja ülekaalulised inimesed ning see vorm on paljuski seotud elustiiliga, siis vaadatakse täpsemalt just II tüüpi diabeedi mõju elukestusele.

Uuringus jälgimise all on geenidoonorid 2017. aasta keskpaiga seisuga olnud keskmiselt 8,82 aastat (standardhälbega 2,7). Kõige pikemalt on jälgitud geenidoonorit 14,7 aastat. Eesti geenivaramuga liitunud inimestest on linkimise hetkeks elus 91, 7%. See tähendab,

et sündmusi ehk surma on toimunud siiski suhteliselt vähe ning enamik inimesi on tsenseeritud 2017. aasta seisuga. Lisas 6 on toodud näited analüüsis kasutatavast R-i koodist.

3.1 Elukestust mõjutavad mittegeneetilised tegurid

Mittegeneetilistest teguritest oli geenidoonori kohta teada tema kõrgeim omandatud haridustase, suitsetamise staatus ja kehamassiindeks. Kuigi geenidoonori täidetud küsimustikus oli haridustase täpsemalt kirjas, kategoriseeriti see ümber kolmeks tasemeks: põhi-, kesk- ja kõrgharidus. Kehamassiindeks arvutati kaalu (kg) ja pikkuse (m) ruudu suhtena ning loeti normaalseks, kui see jäi vahemikku 19-25.

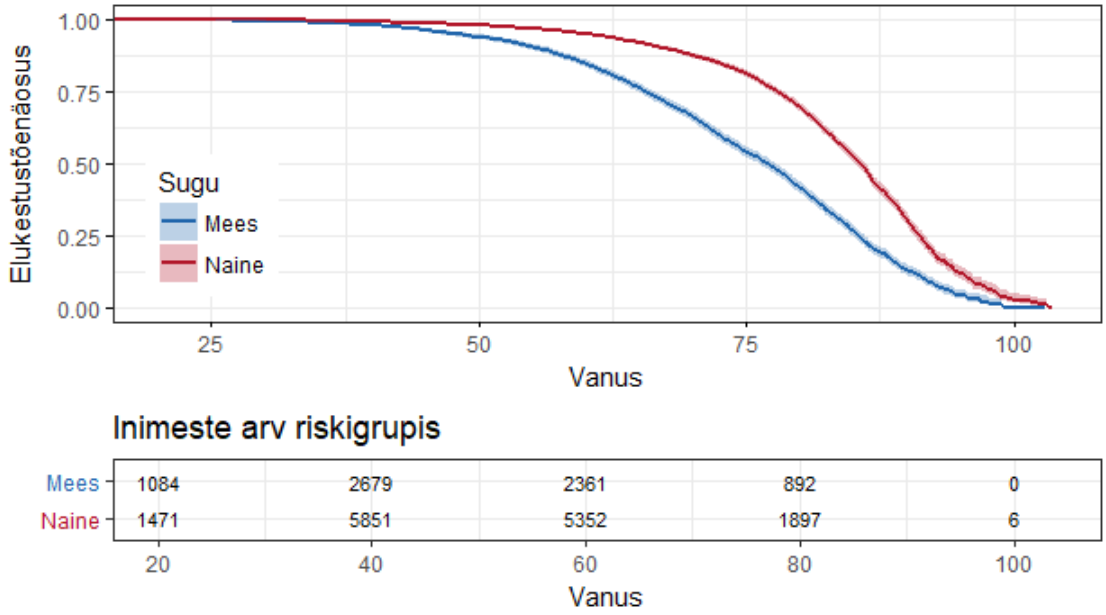
Analüüsitavas andmestikus (vt ka Tabel 3) on mehi 34,3% ja naisi 65,7%. Keskmine vanus liitumisel oli 44,6 aastat, olles meestel natuke madalam (43,4 aastat) ning naistel natuke kõrgem (45,2 aastat). Keskmine vanus surmahetkel on meestel märgatavalt väiksem kui naistel, vastavalt 69,7 aastat ja 74,9 aastat.

Keskmine kehamassiindeks on naistel ja meestel sarnane, üle 26, mis ületab normaalseks peetavat kehamassiindeksi piiri. Sealhulgas ületab 22% geenidoonoritel kehamassiindeks rasvumisele viitava 30 piiri. Alakaalulisi geenidoonoreid (KMI < 19) on vaid 3,8%. Vähemalt keskharidusega on geenidoonoritest 82% ning 24,5% on kõrgharidusega. Tabelist 3 on näha ka suitsetajate osakaal andmestikust. Meestest on praegusi suitsetajaid 39,6% ning 20,8% on endisi suitsetajaid. Naiste hulgas on suitsetajate osakaal väiksem, 22,7% praegusi ja 10% endisi suitsetajaid.

Tabel 3: Vanuse, KMI, hariduse ja suitsetamise näitajad sugude lõikes

	Mees n=17702	**Naine** n=33886	**Kokku** n=51588
keskmine vanus liitumisel (standardhälve)	43,6 (17,9)	45,4 (17,2)	44,7 (17,5)
keskmine vanus surmahetkel (standardhälve)	69,7 (13,6)	74,9 (13,4)	72,4 (13,8)
keskmine KMI (standardhälve)	26,4 (4,7)	26,3(5,7)	26,3 (5,4)
vähemalt keskharidusega, %	77,9	84,3	82,0
kõrgharidusega, %	21,1	26,3	24,5
praeguseid suitsetajaid, %	39,6	22,7	28,5
endisi suitsetajaid, %	20,8	10,0	13,7

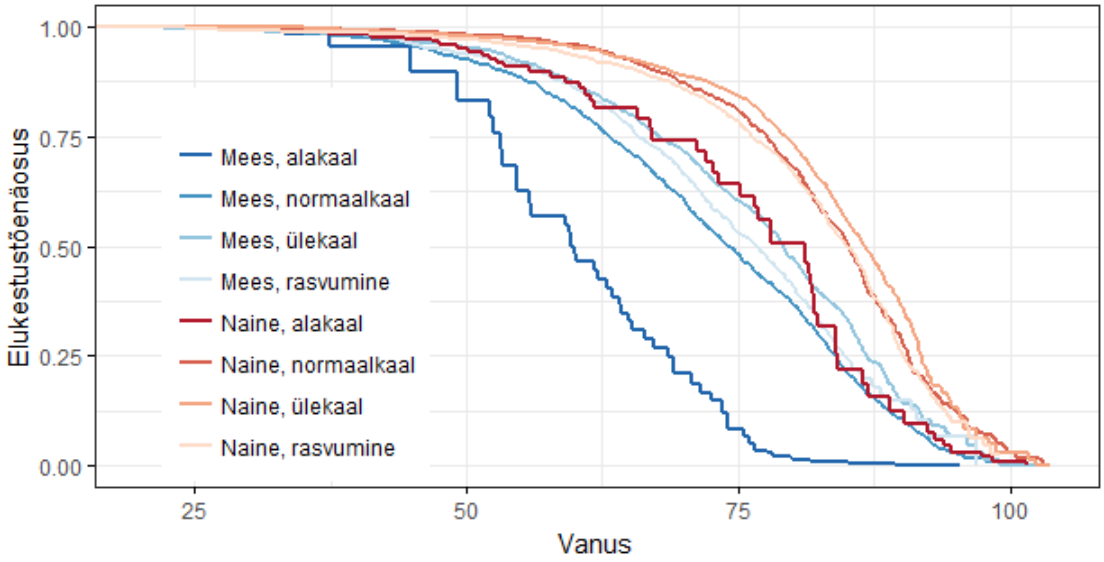
Kaplan-Meieri hinnangud üleelamisfunktsioonile sugude lõikes on nähtavad jooniselt 4. On näha, et meeste ja naiste elukestused on küllaltki erinevad ning naised elavad kauem kui mehed. Joonise juurde kuuluvas tabelis on toodud riskigrupi suurused viie erineva vanuse korral. Riskigrupis on inimesed, kes olid uuringus vaatluse all selles vanuses.



Joonis 4: Eesti geenivaramu geenidoonorite elukestus sugude lõikes (Kaplan - Meieri kõver 95% usaldusvahemikuga)

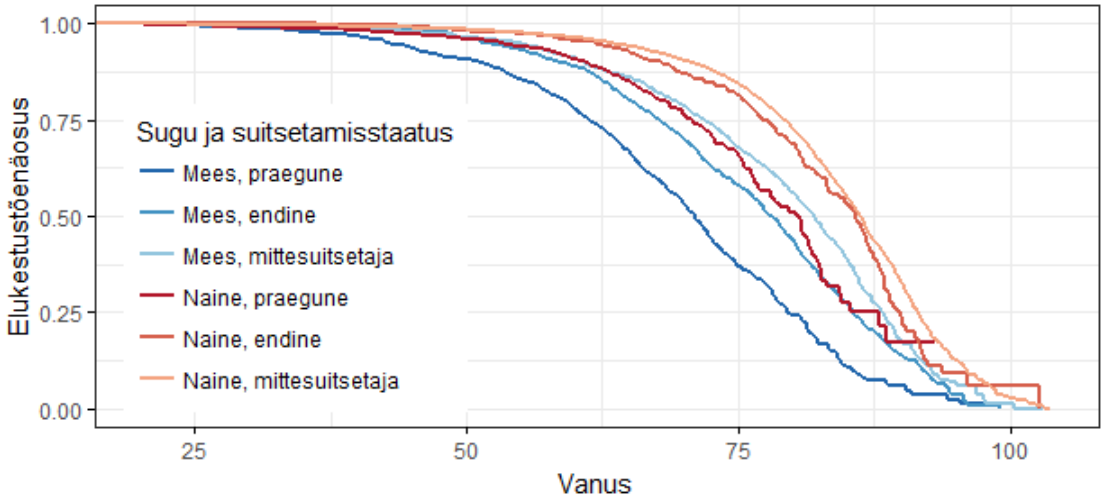
Lisas 3 on toodud joonis 9, kus riskigrupi juures ei ole vasakult tõkestatust arvesse võetud, ehk iga inimest võrreldakse sellega, kes on elanud vähemalt sama vanuseni. Jooniselt 9 on näha, et siis on naiste ja meeste elukestusjooned natuke rohkem lähestikku, aga riskigrupid palju suuremad.

Joonisel 5 on lisaks soole vaadatud ka kehamassiindeksi lõikes hinnangut üleelamisfunktsioonile. Teistest madalama eluea poolest eristuvad alakaalulised mehed. Ka alakaaluliste naiste eluiga on tunduvalt madalam võrreldes ülejäänud naiste elueaga. Alakaalulisust ning seetõttu ka varajast surma võib põhjustada mingi tegur, mida antud andmestikust näha ei saa, näiteks haigused. Kõige kõrgem elukestus on joonise järgi ülekaalulistel, kuid mitte rasvunud geenidoonoritel. Kui meestel on elukestust kujutavad jooned selgelt eristatavad, siis naistel on eristunud vaid alakaalulised naised.



Joonis 5: Eesti geenivaramu geenidoonorite elukestus sugude ja kehamassiindeksi lõikes (Kaplan - Meieri kõver)

Suitsetamise mõju elukestusele on näha jooniselt 6. Selgelt eristuvad teistest madala elukestuse poolest liitumishetkel suitsetavad mehed ja naised.



Joonis 6: Eesti geenivaramu geenidoonorite elukestus sugude ja suitsetamisstaatuse lõikes (Kaplan - Meieri kõver)

Mittegeneetiliste riskitegurite täpsemaks uurimiseks sobitati andmetele Coxi võrdeliste riskide mudel. Simulatsioonide põhjal kasutatakse mudelit, mis võtab arvesse nii tsenseerimist kui ka tõkestatust ning ajaskaalana kasutatakse vanust. Mudelisse lisati

faktoritena sugu, haridustase, suitsetamisstaatus ning kehamassiindeks. Kui kehamassiindeks lisada mudelisse pideva tunnusena, siis ei tulnud see statistiliselt oluline. Seetõttu vaadati KMI-d grupeeritud tunnusena, kus väärtusteks alakaal (KMI < 19), normaalkaal ($19 \leq$ KMI < 25), ülekaal ($25 \leq$ KMI < 30) ning rasvumine (KMI > 30).

Tabel 4: Riskisuhte erinevus

	$\hat{\psi} = e^{\hat{\beta}}$	**95% usaldusintervall**	**p-väärtus**
naine	0,53	(0,49; 0,57)	$< 2 \cdot 10^{-16}$
kõrgharidus	0,72	(0.66; 0.79)	$2,6 \cdot 10^{-12}$
põhiharidus	1,40	(1.30; 1.50)	$< 2 \cdot 10^{-16}$
praegune suitsetaja	2,17	(1.99; 2.36)	$< 2 \cdot 10^{-16}$
endine suitsetaja	1,21	(1.11; 1.32)	$1,5 \cdot 10^{-5}$
alakaal	1,96	(1.61; 2.37)	$8,7 \cdot 10^{-12}$
ülekaal	0,85	(0.79; 0.91)	$1,9 \cdot 10^{-5}$
rasvumine	1,03	(0.95; 1.12)	0,42

Tabelis 4 on toodud analüüsi tulemused. Baastasemeks on soo puhul mees, hariduse puhul keskharidus, suitsetamisstaatuse puhul mittesuitsetajad ning kehamassiindeksi puhul normaalkaalus olev inimene. Naisel on suremise risk 2 korda väiksem kui mehel. Võrreldes keskharidusega inimesega on kõrgharidusega inimese risk surra 1,4 korda väiksem ning põhihardusega inimese risk 1,4 korda suurem. Mittesuitsetajaga võrreldes on risk surra praegusel suitsetajal 2,17 korda suurem ning endistel suitsetajatel 1,21 korda suurem. KMI järgi alakaalus inimese risk surra on peaaegu 2 korda suurem kui normaalkaalus inimesel ning ülekaalus inimesel on risk 1,18 korda väiksem kui normaalkaalus inimesel. Rasvunud ja normaalkaalus inimese korral on riskide suhe 1 lähedane, teisisõnu on risk surra samasugune. Kokkuvõttes on kõige madalam suremuse risk naisel, kellel on kõrgharidus, kes pole kunagi suitsetanud ning on normaalkaalus. Kõige suurem risk surra on alakaalus põhiharidusega suitsetaval mehel.

3.2 Geneetiliste riskiskooride seos elukestusega

Geneetilised riskiskoorid inimesele on saadud ülegenoomse assotsiatsiooniuuringu käigus. Riskiskoori väärtused on skaleeritud, see tähendab, et keskmine on null ning standardhälve üks. Kõrge II tüüpi diabeedi geneetiline riskiskoor ei näita kindlat II tüüpi

diabeedi teket, vaid geneetilise soodumuse olemasolu II tüüpi diabeedi tekkeks. Analüüsitavast andmestikust oli 47 647 geenidoonoril olemas geneetilise riskiskoori väärtus II tüüpi diabeedile. Geenidoonoritest, kelle geneetiline riskiskoor on leitud, on II tüüpi diabeedi diagnoosiga (liitumishetkel või hiljem tekkinud) 9,4%. Kõrge riskiskooriga geenidoonoritest (riskiskoor on kõrgema 20% sees) on diabeedihaigeid 13,9% ning madala riskiskooriga geenidoonoritest (riskiskoor on madalama 20% sees) on diabeet 5,6%-il inimestest. Seega on madala ja kõrge riskiskooriga haigestunute vahe rohkem kui kahekordne. [12]

Geneetilised riskiskoorid on inimesel muutumatud kogu elu jooksul ning on teoreetiliselt teada juba sündides. Seega vaadatakse geneetilise riskiskoori mõju elukestusele mudelis, kuhu lisatakse argumenttunnustena vaid sugu ja riskiskoori väärtus diabeedile (lühemalt grs t2d).

Tabel 5: II tüüpi diabeedi geneetilise riskiskoori mõju elukestusele vasakult tõkestatust arvestavas mudelis.

	$\hat{\psi} = e^{\hat{\beta}}$	**95% usaldusintervall**	**p-väärtus**
naine	0,47	(0,44; 0,50)	$< 2 \cdot 10^{-16}$
grs t2d	1,03	(0,99; 1,06)	$0,083$

Tabelis 5 toodud riskide suhe on 1,03, kuid olulisuse nivool $\alpha = 0{,}05$ ei ole see statistiliselt oluline. Simulatsioonuuringu esimesest osast saadi teada, et suurema võimsusega, kuid nihkega hinnangu saab, kui vasakult tõkestatust mudelis mitte arvesse võtta. See muudab riskigruppi valemis (7), kuhu ei kuulu ainult vaatluse all olevad samas vanuses inimesed, vaid kõik inimesed, kes on võrreldava vanuseni elanud. Ajaskaalana kasutatakse vanust, seega proovitakse mudelit, mis simulatsioonuuringute esimeses osas kandis nime "vale." Argumenttunnused jäävad samaks.

Tabel 6: II tüüpi diabeedi geneetilise riskiskoori mõju elukestusele vasakult tõkestatust mittearvestavas mudelis

	$\hat{\psi} = e^{\hat{\beta}}$	**95% usaldusintervall**	**p-väärtus**
naine	0,48	(0,45; 0,51)	$< 2 \cdot 10^{-16}$
grs t2d	1,04	(1,01; 1,07)	$0,0134$

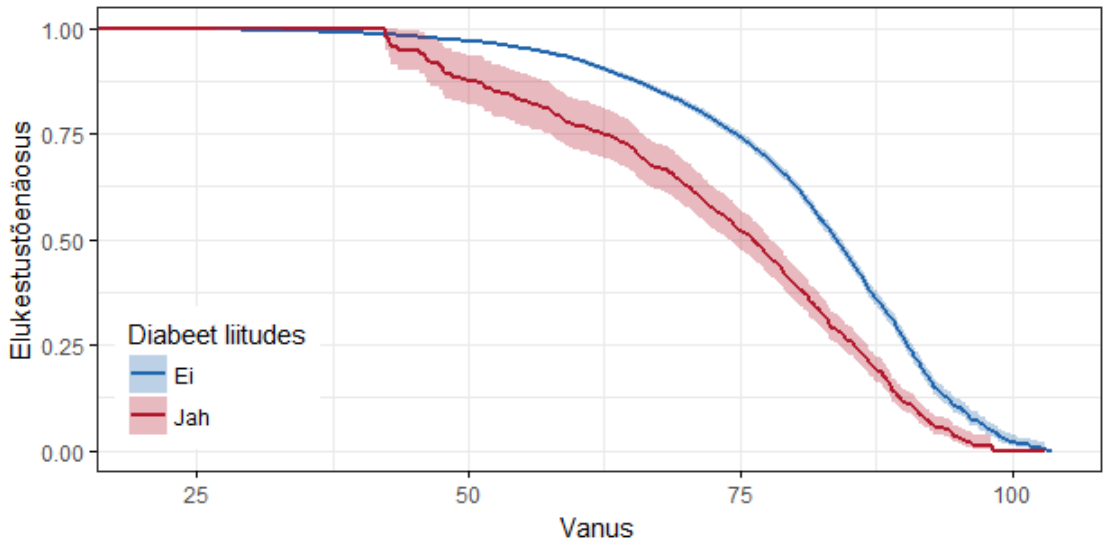
Viimati kirjeldatud viisil tuli Coxi võrdeliste riskide mudelis ka geneetilise riskiskoori

väärtus statistiliselt oluline (vt Tabel 6). Selle põhjal saab öelda, et juba sünnihetkel on II tüüpi diabeedi geneetilisel riskiskooril mingi mõju inimese elukestusele. Täpset mõju suurust ei saa aga antud juhul öelda. Teada on vaid see, et riskiskoori ühele ühikule vastav riskisuhe on väiksem kui 1,04 aga mõju on olemas. Nii väikese mõju avastamine on antud hetkel Eesti geenivaramu andmestikus keeruline, kuna uuringus oldud aeg on lühike ning väga vähe inimesi on surnud.

3.3 Diabeedi diagnoosi mõju elukestusele

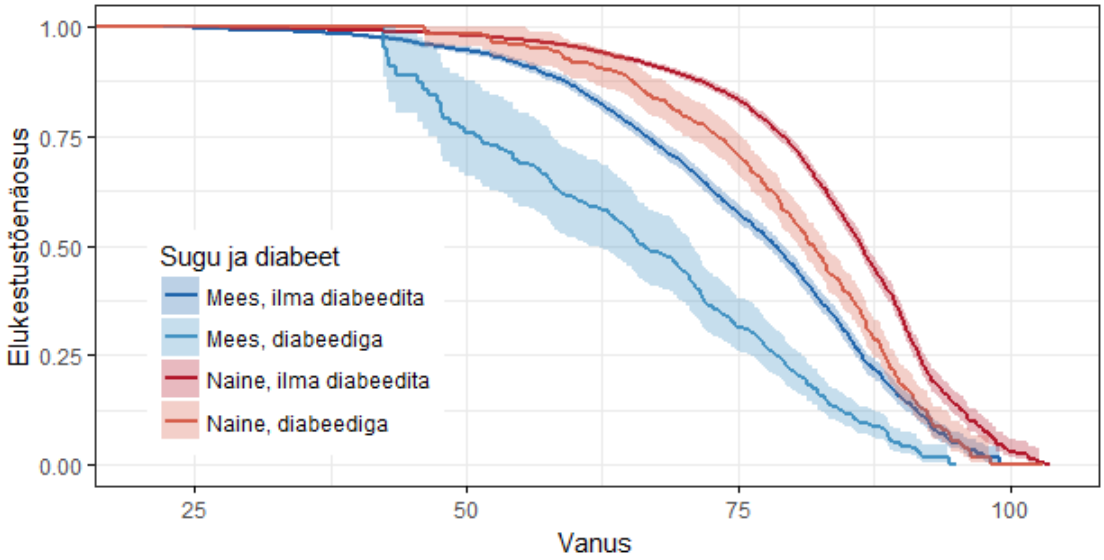
Järgnevas peatükis analüüsitavast Eesti geenivaramu andmestikust on välja jäetud geenidoonorid, kellel on haigekassa andmete järgi diagnoositud nii I kui ka II tüüpi diabeet, kuna meditsiiniliselt sellist olukorda olla ei tohiks. Põhjuseid topelt diagnoosi jaoks on mitmeid, näiteks võib olla esialgne diagnoos vale ning see parandatakse pärast õige diagnoosi saamist.

Joonisel 7 on toodud geenidoonorite elukestused diabeedi diagnoosi lõikes, kusjuures arvesse on võetud ainult II tüüpi diabeedi diagnoosi uuringusse tulemise hetkel. Selle joonisel abil saab öelda, et diabeedidiagnoosiga inimestel muutub riski joon just vanuses 45-50, mis on ka vanuseks, kus enamasti II tüüpi diabeet diagnoositakse. Heledama värviga on joonisele kantud usaldusintervall Kaplan-Meieri hinnangule.



Joonis 7: Liitumishetkeks saadud diabeedi mõju geenidoonorite elukestvusele (Kaplan - Meieri kõver 95% usaldusvahemikuga)

Järgnevalt toodud Joonsel 8 on vaadatud nii kõiki diabeedi diagnoosiga inimesi, sealhulgas siis ka neid, kes said diagnoosi uuringus olemise hetkel. Lisaks on elukestusi vaadatud ka soo lõikes, kuna eelnevalt on näha olnud selle tunnuse suur mõju elukestusele. Siit on selgelt näha, et diabeeti haigestunutel on elukestused väiksemad.



Joonis 8: Liitumishetkeks või uuringu käigus saadud diabeedi mõju geenidoonorite elukestvusele (Kaplan - Meieri kõver 95% usaldusvahemikuga)

Edasi lisati diabeedi diagnoos ka mudelisse. Et olemas oli info ka selle kohta, kas geenidoonoril oli infarkti või insuldi diagnoos enne uuringuga liitumise hetke, siis lisati ka need tunnused mudelisse, kuna need võivad riskile suurt mõju avaldada. Coxi võrdeliste riskide mudeli tulemused on toodud tabelis 7. Diabeedi korral on tegu nii nendega, kellel oli diabeet uuringu alguses, kui ka nendega, kes on selle diagnoosi saanud uuringus olemise ajal. Seega on diabeet ajast sõltuv tunnus.

Tabel 7: Riskisuhte erinevus ajas muutuva tunnuse korral

	$\hat{\psi} = e^{\hat{\beta}}$	**95% usaldusintervall**	**p-väärtus**
naine	0,54	(0,51; 0,58)	$< 2 \cdot 10^{-16}$
kõrgharidus	0,74	(0,67; 0.81)	$1,8 \cdot 10^{-11}$
põhiharidus	1,40	(1,30; 1.50)	$< 2 \cdot 10^{-16}$
praegune suitsetaja	2,18	(2,00; 2.38)	$< 2 \cdot 10^{-16}$
endine suitsetaja	1,16	(1,06; 1.27)	$< 2 \cdot 10^{-16}$
alakaal	1,97	(1,62; 2.39)	$1,16 \cdot 10^{-11}$
ülekaal	0,81	(0,75; 0.87)	$3,77 \cdot 10^{-8}$
rasvumine	0,92	(0,85; 1.00)	$0,0446$
diabeet	1,68	(1.54; 1.82)	$< 2 \cdot 10^{-16}$
infarkt	1,44	(1,30; 1.59)	$1,82 \cdot 10^{-12}$
insult	1,52	(1,34; 1.73)	$1,31 \cdot 10^{-10}$

Tulemuste põhjal saab öelda, et II tüüpi diabeediga inimestel on risk 1,67 korda suurem kui diabeedita inimestel. Ka infarkti või insuldi saanud inimestel on risk surra suurem, vastavalt 1,44 ja 1,52 korda, võrreldes nendega, kellel pole ühte või teist meditsiinilist häiret olnud. Analüüsitud mudeli järgi on kõige väiksem risk kõrgharidusega mittesuitsetaval naisel, kellel KMI on vahemikus 25-30, kellel pole diabeeti ning pole saanud infarkti ega insulti.

Kokkuvõte

Magistritöö eesmärgiks oli välja selgitada, milline on kõige sobivam meetod elukestusanalüüsi läbiviimiseks, kui andmetes esines nii paremalt tsenseeritust kui ka vasakult tõkestatust. Lisaks pakkus huvi, kas ja kuidas peaks arvestama ajas muutuva argumenttunnusega ning millist ajaskaalat epidemioloogiliste uuringute analüüsimisel kasutada. Magistritöös viidi läbi mitu simulatsioonuuringut. Simulatsioonide tulemusena selgus, et ajaskaalana tuleb kasutada vanuselist skaalat ning vasakult tõkestatuse mittearvestamise korral saadakse nihkega hinnangud riskisuhtele. Siiski leiti, et väga väikese riskisuhte avastamiseks võib olla parem, kui vasakult tõkestatust arvesse ei võeta, kuna sellisel juhul tuli mudeli võimsus suurem. Simulatsioonuuringu teises osas vaadati ajast sõltuva argumenttunnuse kasutamist. Kui kasutada ainult informatsiooni riskiteguri olemasolu kohta uuringuga liitumise hetkel, tuli mudeli võimsus väiksem, kui mudelil, mis arvestas ka uuringus olemise ajal riskiteguri saamist ehk haiguse avaldumist.

Simulatsioonuuring tulemusi rakendati Tartu genoomika instituudi alla kuuluva Eesti geenivaramu andmestikule. Andmestikus oli üle 51 000 inimese. Inimese elukestuse määramisel kasutati tema vanust geenivaramuga liitumise hetkel ning tema vanust surma või tsenseerimise hetkel. Taustatunnustena teati iga inimese kohta tema sugu, vanust, kehamassiindeksit, haridustaset ja suitsetamisstaatust. Lisaks vaadati täpsemalt ka II tüüpi diabeedi diagnoosiga inimeste elukestust ning ka II tüüpi diabeedi geneetilise riskiskoori seost elukestusega. Töös leiti, et vaadeldavatest tausttunnustest peaaegu kõik suurendavad või vähendavad inimese elukestust suuremal või vähemal määral. II tüüpi diabeedi diagnoosiga inimeste risk surra oli 1,68 korda suurem võrreldes inimesega, kellel ei ole diagnoosi. Suitsetava inimese risk on aga 2,18 korda suurem, kui mittesuitsetaval inimesel. Geeniuuringute koha pealt on oluline, et suudeti tõestada geneetilise riskiskoori ja elukestuse vahelist seost.

Selles töös vaadeldud inimesed olid uuringus olnud veel küllaltki lühikest aega ning vaid 8,2% inimestest on surnud. Kuna sündmuse toimumise arv ehk antud juhul surmade hulk on elukestusanalüüsis oluline, siis tuleks analüüsi korrata kindlasti 10 või 20 aasta pärast. Paremate järelduste tegemisteks tuleks ka riskitegurite hulka suurendada, näiteks vaadelda rohkemaid haigusi.

Viited

[1] Põhikiri. Tartu Ülikooli Eesti geenivaramu, URL (vaadatud: 10.05.2018) `http://www.geenivaramu.ee/et/pohikiri`

[2] Allison, P. D. (2010) *Survival Analysis Using SAS®: A Practical Guide, Second Edition*, SAS® Press

[3] Fischer, K (2007) Elukestusanalüüs. Loengukonspekt. Tartu: Tartu Ülikool, tervishoiu instituut.

[4] Collett, D. (2015) *Modelling Survival Data in Medical Research Third Edition*, CRC Press

[5] The Weibull Distribution. R Documentation, URL (vaadatud: 11.05.2018) `http://stat.ethz.ch/R-manual/R-devel/library/stats/html/Weibull.html`

[6] Tableman, M., Kim, J. S. ja Portnoy, S. (2004) *Survival Analysis Using S*, CRC Press

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

[8] Package "survival" (2018), URL (vaadatud: 15.05.2018) `https://cran.r-project.org/web/packages/survival/survival.pdf`

[9] Läll, K., Mägi, R., Morris, A., Metspalu, A. ja Fischer, K. (2016). Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics In Medicine*, 19, p.322–329, doi:10.1038/gim.2016.103.

[10] Mis on suhkruhaigus?. Suhkruhaigus.ee, URL (vaadatud: 07.05.2018) `http://www.suhkruhaigus.ee/-mis-on-suhkruhaigus`

[11] Mis on diabeet?. Eesti Diabeediliit, URL (vaadatud: 07.05.2018) `http://www.diabetes.ee/mis-on-diabeet`

[12] Geenivaramu uudne riskiskoor ennustab diabeeti täpsemalt kui varasemad. Tartu Ülikooli Eesti geenivaramu, URL (vaadatud: 07.05.2018) `https://www.geenivaramu.ee/et/uudised/geenivaramu-uudne-riskiskoor-ennustab-diabeeti-tapsemalt-varasemad`

Lisad

Lisa 1 - Keskmised näitajad simulatsiooniuuringutes saadud andmestike kohta

Tabel 8: Keskmised näitajad simulatsiooniuuringu I osas simuleeritud andmete kohta

	$\psi = 2$	$\psi = 1,2$	$\psi = 1,05$	$\psi = 1$
elas liitumiseni, %	47,8	52,8	54,2	54,7
surid uuringus (liitunutest), %	29,8	27,2	26,4	26,2
tsenseeriti (liitunutest), %	70,2	72,8	73,6	73,8

Tabel 9: Keskmised näitajad simulatsiooniuuringu II osas simuleeritud andmete kohta

	$\psi = 2$	$\psi = 1,2$	$\psi = 1,05$
elas liitumiseni, %	47,8	49,5	49,8
riskitegur oli enne liitumist (liitunutest), %	16,7	19,4	20,0
riskitegur tekkis pärast liitumist (liitunutest), %	7,7	7,5	7,4
ei saanud riskitegurit (liitunutest), %	75,6	74,1	72,6

Lisa 2 - Näide simuleeritud andmestikust, analüüsides ajas muutuvat riskitegurit

Tabel 10: Andmestiku kuju

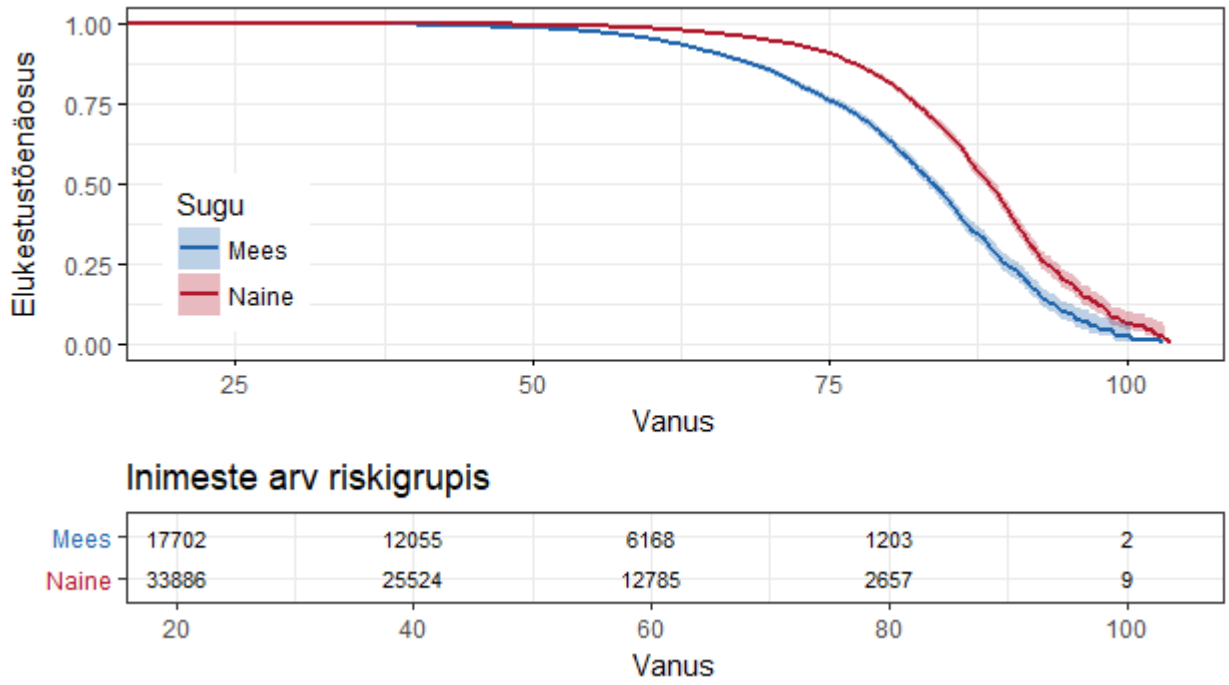
id	x	$x2$	t^*	p_i	sünniaeg	liitumisaeg	riskitegur liitudes	vanus perioodi alguses	vanus perioodi lõpus	sündmus	riskigrupp
7	117,8	8,2	71,6	0,4	1953,9	2005,2	0	51,2	63,0	0	1
21	81,9	18,9	44,1	0,53	1951,0	2009,4	1	58,1	62,9	1	2
31	65,3	1,3	59,3	0,9	1957,5	2004,9	0	47,5	59,3	0	1
31	65,3	1,3	59,3	0,9	1957,5	2004,9	0	59,3	59,5	0	2
144	80,5	8,8	63	0,6	1943,1	2003,7	0	60,6	63	0	1
144	80,5	8,8	63	0,6	1943,1	2003,7	0	63	71,8	1	2

Tabelis 10 olevate tunnuste seletused:

- id - inimese id
- $x \sim Exp(\lambda_1 = 1/70)$ - genereeritud eluiga sünnist
- $x \sim Exp(\lambda_2 = \psi/70)$ - genereeritud eluiga alates riski tekkimisest
- $t^* \sim N(55, 8)$ - genereeritud vanus, kus tekib riskitegur
- $tn \sim U(0, 1)$ - genereeritud tõenäosus, kui $p_i > 0,5$ ning $x > t^*$, siis tekib riskitegur hetkel t^*
- liitumisaeg - uuringuga liitumisaeg, genereeritud $U(2003, 2010)$
- riskitegur liitudes - kas inimesel oli uuringuga liitumise ajal vaadeldav riskitegur
- vanus perioodi alguses - perioodi algus on kas uuringuga liitumise hetk või riskiteguri tekkimise hetkel (juhul, kui see tekib uuringu ajal)
- vanus perioodi lõpus - vanus kas surres, uuringu lõppedes või riskiteguri tekkimise hetkel
- sündmus - saab väärtuse 1, kui inimene suri perioodi lõpuks, vastasel juhul 0 (inimene tsenseeriti)
- riskigrupp - kumba riskigruppi kuulub inimene antud perioodis

  Heledamaks halliks on tabelis 10 tehtud väärtused, mis genereeriti, kuid kasutada ei olnud vaja (kuna p_i<0.5).

Lisa 3 - Kaplan-Meieri graafik sugude lõikes



Joonis 9: Kaplan-Meieri kõver sugude lõikes vasakult tõkestatust arvestamata

Lisa 4 - Tarkvara R kood simulatsioonuuringu esimese osa kohta

```
library(survival)
library(dplyr)
set.seed(34556) #analuusi reprodutseerimiseks
#Hindan ühe parameetri päris andmetelt:
aaaa=survreg(Surv(f1$loppvanus)~1, dist="weibull")
a <- 1/aaaa$scale       # u 3.5
b <- exp( coef(aaaa) ) #

####Funktsioon####
#funktsioonile annan ette simulatsioonide arvu, kujuparameeri ja soovitud riski suhte
simulatsioonid <- function(s_arv, shape=3.5, riski_kordaja){
  tulemused=data.frame(matrix(NA, nrow = s_arv, ncol = 19)) #anmestik, kuhu salvestada
      simulatsioonide tulemused
  for (i in 1:s_arv){
    n=10000 #vaatluste arv ühes grupis
    med_vanus=55#vajalik kahe marameeriti omavahelisek sidumiseks
    scale=1/(log(2)**(1/shape)/med_vanus)  #skaalaparameeter
    risk=riski_kordaja**(1/shape) #soovitud risk R-i parametrisatsioonis
    #eluea simuleerimine, 2 gruppi, ühel risk suurem
    x=rweibull(n, shape, scale)
    x2=rweibull(n, shape, 1/risk*scale)
    #sünniaastad ühtlasest jaotusest
    synniaastad=runif(2*n, min=1920, max=1990)
    cutoff=rep(2017, 2*n)#analüüsi aeg
    andmed=cbind(c(x,x2), synniaastad, cutoff)
    andmed=as.data.frame(andmed)
    andmed$surmaaastad=andmed$synniaastad+andmed$V1
    andmed$grupp=c(rep(1,n), rep(2,n)) #teadaolevad riskigrupid
    #Tsenseerime andmed seisuga cutoff, inimese lopp_vanus + kas on surnud
    andmed$lopp_vanus=ifelse(andmed$synniaastad+andmed$V1<andmed$cutoff, andmed$V1,
        andmed$cutoff-andmed$synniaastad )
    andmed$cens=ifelse(andmed$synniaastad+andmed$V1<andmed$cutoff, 0, 1)
    #liitumisaastad ühtlasest jaotusest
    andmed$liitumiaastad=runif(2*n, min=2003, max=2010)
    #kes surid enne võimalikku liitumist?
    andmed$surm_enne=ifelse(andmed$synniaastad+andmed$V1<andmed$liitumiaastad, 1, 0)
    #vanused liitumisel
    andmed$liit_vanus=andmed$liitumiaastad-andmed$synniaastad
    #võtame välja inimesed kes surid enne võumalikku liitumist
    andmed2=andmed[andmed$surm_enne==0,]

    ###### Coxi mudelid ####
    #õige mudel - kaasatud kõik inimesed ilma tsenseerimise jms.
    surv_obj1=Surv(time=andmed$V1, event=rep(1,nrow(andmed)))
    mudel1=coxph(surv_obj1~grupp , data=andmed)
    #paremalt tsentseeritud mudel - kaasatud kõik inimesed
    surv_obj2=Surv(time=andmed$lopp_vanus, event=1-andmed$cens, type='right')
    mudel2=coxph(surv_obj2~factor(grupp), data=andmed)
    #Paremalt tsenseerimine, kaasatud liitunud
    #a) vanus sünnist
```

```
    surv_obj3a=Surv(time=andmed2$lopp_vanus, event=1-andmed2$cens, type='right')
    mudel3a=coxph(surv_obj3a~grupp , data=andmed2)
    #b) aeg liitumisest (ehk lõppvanus-liitvanus)
    surv_obj3b=Surv(time=(andmed2$lopp_vanus-andmed2$liit_vanus), event=1-andmed2$cens
        , type='right')
    mudel3b=coxph(surv_obj3b~factor(grupp), data=andmed2)
    #võtame arvesse vasakult tõkestatust
    surv_obj4=Surv(time=andmed2$liit_vanus, time2=andmed2$lopp_vanus, event=1-andmed2$
        cens)
    mudel4=coxph(surv_obj4~grupp, data=andmed2)

    #salvestame iga simulatsiooni tulemused
    tulemused[i,1]=summary(mudel1)$coefficients[2]#exp(coef)
    tulemused[i,2]=summary(mudel1)$coef[3] #se(soef)
    tulemused[i,3]=summary(mudel1)$coef[5]
    ...
    tulemused[i,14]=summary(mudel4)$coef[3]
    tulemused[i,15]=summary(mudel4)$coef[5]
    tulemused[i,16]=sum(andmed$cens==1)/nrow(andmed)
    tulemused[i,17]=sum(andmed2$cens==1)/nrow(andmed2) #inimeste arv, kes elasid
        liitumiseni ja siis tsenseeriti
    tulemused[i,18]=nrow(andmed2) #inimeste arv, kes elas liitumiseni
    tulemused[i,19]=sum(andmed2$cens==0)/nrow(andmed2) #inimeste arv, kes surid
        uuringus olemise ajal

    colnames(tulemused) <- c('m1_c', 'm1_s', 'm1_p', 'm2_c', 'm2_s', 'm2_p', 'm3a_c',
        'm3a_s', 'm3a_p', 'm3b_c', 'm3b_s', 'm3b_p', 'm4_c', 'm4_s', 'm4_p', 'tsens1',
        'tsens2', 'liitus', 'surid_uuringus')
  }
  return (tulemused)
}

###Väljakutsumine####
tul1=simulatsioonid(1000, 3.5, 2)
tul2=simulatsioonid(1000, 3.5, 1.2)
tul3=simulatsioonid(1000, 3.5, 1.05)
tul0=simulatsioonid(1000, 3.5, 1)
```

Lisa 5 - Tarkvara R kood simulatsioonuuringu teise osa kohta

```
#### Simulatsiooonid ajast sõltuva muutuja korral
set.seed(34556)
#funktsioonile annan ette simulatsioonide arvu, kujuparameeri ja riski suhte

simulatsioonid_aeg <- function(s_arv, risk){
   tulemused=data.frame(matrix(NA, nrow = s_arv, ncol = 11)) #anmestik, kuhu
       salvestada simulatsioonide tulemused
  for (i in 1:s_arv){
    n=100000 #vaatluste arv
    #eluiga exp jaotusest - alguses kõigil sama risk, aga u vanuses 55, osadel risk
        suureneb
    x0=rexp(n, rate=1/70)
    x02=rexp(n, rate=risk*1/70)
    x=ifelse(x0<140, x0, 140) #vältimaks liiga pikke eluaegu
    x2=ifelse(x02<140, x02, 140)
    x=ifelse(x<0.01, 0.01, x) #vältimaks juhtu, kus ajaintervall on liiga lühike ning
        mudel annab veateate
    x2=ifelse(x2<0.01, 0.01, x2)
    t_tarn=rnorm(n, mean=50, sd=8) #riski muutumiskoht
    tn=runif(n, min=0, max=1)  # tõenäosus
    #sünniaastad ühtlasest jaotusest
    synniaastad=runif(n, min=1920, max=1990)
    liitumisaastad=runif(n, min=2003, max=2010)
    cutoff=rep(2017, n)#analüüsi aeg
    id=c(1:n) #vaatlust id
    grupp2=ifelse(tn >0.5 & x > t_tarn & synniaastad+t_tarn < cutoff, 2, 0) #need
        kellel tekib riskitegur
    riskitegur_liitudes=ifelse(grupp2==2 & synniaastad+t_tarn < liitumisaastad, 1,0) #
        eraldi muutuja veel, et kellel on riskitegur juba liitudes
    andmed1=data.frame(cbind(id, x, x2, t_tarn, tn, synniaastad, cutoff,  grupp2,
        liitumisaastad, riskitegur_liitudes))#genereeritud andmes koos
    dliit1a=andmed1[riskitegur_liitudes==1,] #need, kellel oli riskitegur liirudes
    dliit0a=andmed1[riskitegur_liitudes==0,] #need, kellel oli riskitegur liirudes

    #### MUDEL 1 #### - Kas liitumisel oli diabeet või mitte?
    #diabeet liitumisel, eemaldame need, kes ei saanud liituda
    #eluaeg alates hetkest t_tarn on x2
    dliit1 <- dliit1a %>% mutate(x3=t_tarn+x2) %>% subset(synniaastad+x3-
        liitumisaastad > 0.01 ) %>%
      mutate(liit_vanus=liitumisaastad-synniaastad, lopp_aastad=ifelse(synniaastad+x3<
          cutoff, synniaastad+x3,cutoff),
             lopp_vanus=lopp_aastad-synniaastad, cens=ifelse(synniaastad+x3<cutoff
                 ,0,1), event=1-cens)
    #diabeet ei ole liitumisel, eemaldame need, kes ei saanud liituda
    #eluaeg on kogu aef x
    dliit0 <- dliit0a %>% mutate(x3=x) %>% subset(synniaastad+x3 -liitumisaastad >
        0.01 ) %>%
      mutate(liit_vanus=liitumisaastad-synniaastad, lopp_aastad=ifelse(synniaastad+x3<
          cutoff, synniaastad+x3,cutoff),
             lopp_vanus=lopp_aastad-synniaastad, cens=ifelse(synniaastad+x3<cutoff
```

```
                ,0,1), event=1-cens)
  data=rbind(dliit1, dliit0) #andmed koos
  help1=nrow(dliit1)+nrow(dliit0)
  surv_obj1=Surv(time=data$liit_vanus, time2=data$lopp_vanus, event=data$event)
  mudel1=coxph(surv_obj1~riskitegur_liitudes , data=data)

  ##### MUDEL 2 ######
 #diabeet liitumisel, eemaldame need, kes ei saanud liituda
  dliit1 <- dliit1a %>% mutate(x3=t_tarn+x2) %>% subset(synniaastad+x3 -
      liitumisaastad > 0.01 ) %>%
    mutate(liit_vanus=liitumisaastad-synniaastad, lopp_aastad=ifelse(synniaastad+x3
        <2017, synniaastad+x3,cutoff),
           lopp_vanus=lopp_aastad-synniaastad,  cens=ifelse(synniaastad+x3<cutoff
               ,0,1), event=1-cens, riskigrupp=2)
  #Kui diabeeti liitudes ei ole, siis jagame inimesed kaheks, need kelle tekib
      diabeet uuringu ajala ja need kellele ei teki
  #tekib diabeet uuringu ajal, andmestik enne diabeedi tekkimist
  dliit01 <- dliit0a %>% subset(grupp2==2 & synniaastad + t_tarn - liitumisaastad >
      0.01) %>%
    mutate(liit_vanus=liitumisaastad-synniaastad, lopp_aastad=synniaastad + t_tarn,
        lopp_vanus=lopp_aastad-synniaastad,
           cens=1, event=1-cens, x3=x, riskigrupp=1)
  #teine rida selle kohta, kui diabeet on tekkinud
  dliit02 <- dliit0a %>% subset(grupp2==2 & synniaastad + t_tarn - liitumisaastad >
      0.01) %>%
    mutate(liit_vanus=t_tarn, lopp_aastad=ifelse(synniaastad + t_tarn+x2 < cutoff,
        synniaastad + t_tarn+x2, cutoff),
           lopp_vanus=lopp_aastad-synniaastad,  cens=ifelse(synniaastad + t_tarn+x2
               < cutoff,0, 1), event=1-cens, x3=x2, riskigrupp=2) %>%
    subset(lopp_vanus-liit_vanus > 0.01) #lisakontroll, et uuringus oldnud aeg ei
        oleks väga lühike
  #inimesed kellel diabeeti ei tekigi
  dliit0 <- dliit0a %>% subset(grupp2==0 & synniaastad+x - liitumisaastad > 0.01)
      %>%
    mutate(liit_vanus=liitumisaastad-synniaastad, lopp_aastad=ifelse(synniaastad+x<
        cutoff, synniaastad+x,cutoff),
           lopp_vanus=lopp_aastad-synniaastad, cens=ifelse(synniaastad+x<cutoff,0,1)
               , event=1-cens, x3=x, riskigrupp=1)

  data=rbind(dliit1, dliit01, dliit02, dliit0)
  data <- data %>% arrange(id)
  surv_obj2=Surv(time=data$liit_vanus, time2=data$lopp_vanus, event=data$event)
  mudel2=coxph(surv_obj2~riskigrupp , data=data)
  #salvestame iga simulatsiooni tulemused

  tulemused[i,1]=summary(mudel1)$coefficients[2]#exp(coef)
  tulemused[i,2]=summary(mudel1)$coef[3] #se(soef)
  tulemused[i,3]=summary(mudel2)$coefficients[2]
  tulemused[i,4]=summary(mudel2)$coef[3]
  tulemused[i,5]=summary(mudel1)$coef[5]#p_väärtus
  tulemused[i,6]=summary(mudel2)$coef[5]#p_väärtus
```

```
    tulemused[i,7]=help1 #inimeste arv, kes elasid liitumiseni
    tulemused[i,8]=nrow(dliit1) #inimeste arv, kellel oli riskitegur enne liitumist
    tulemused[i,9]=nrow(dliit02) #inimeste arv, kellel riskitegur tekkis pärast
        liitumist
    tulemused[i,10]=nrow(dliit0) #inimeste arv, kelle ei tekkinud riskitegurit
    tulemused[i,11]=nrow(data[data$event==1,]) #paljud surid

    colnames(tulemused) <- c('m1_c', 'm1_s', 'm2_c', 'm2_s', 'p1', 'p2', 'elus_liit',
        'riskitegur_enne', 'riskitegur_parast', 'riskitegur_ei')
  }
  return (tulemused)
}
tul4=simulatsioonid_aeg(1000, 2)
tul5=simulatsioonid_aeg(1000, 1.2)
tul6=simulatsioonid_aeg(1000, 1.05)
```

Lisa 6 - Näited analüüsis kasutatud tarkvara R koodist

Näide peatükis 3 kasutatud jooniste koodi kohta.

```
m05=survfit(Surv(time=gvdata$liitvanus, time2=gvdata$loppvanus, event=gvdata$surnud)~
    sugu + factor(bmi_grupp), data=gvdata, type='kaplan-meier')

ggsurvplot(m05,
           legend.title = '',
           legend.labs = c("Mees, alakaal", "Mees, normaalkaal", "Mees, ülekaal", "
               Mees, rasvumine", "Naine, alakaal", "Naine, normaalkaal", "Naine, ü
               lekaal", "Naine, rasvumine"),
           legend=c(0.2,0.42),
           xlab="Vanus",
           ylab="Elukestustõenäosus",
           xlim=c(20, 104),
           conf.int = F, #usaldusintervallid
           censor=F, #tsenseerimine
           risk.table = F, #riskitabel
           #tables.height = 0.2,
           #risk.table.title='Inimeste arv riskigrupis',
           #tables.theme = theme_minimal(),
           palette = c("#2166AC", "#4393C3", "#92C5DE", "#D1E5F0", "#B2182B", "#D6604D
               ", "#F4A582", "#FDDBC7" ),
           ggtheme = theme_bw() #joonise stiil
)
```

Näide Coxi võrdeliste riskidega mudeli kasutamise kohta.

```
#vasakult tõkestatust arvestav mudel
m1=coxph(Surv(time=gvdata$liitvanus, time2=gvdata$loppvanus, event=gvdata$surnud)~
    factor(sugu) + Haridus + relevel(suitsstat, ref='Never') + relevel(as.factor(bmi_
    grupp), ref=2), data=gvdata)
summary(m1)

#vasakult tõkestatust mittearvestav mudel
m2=coxph(Surv(time=gvdata$liitvanus, time2=gvdata$loppvanus, event=gvdata$surnud)~
    factor(sugu) + grs_t2d, data=gvdata)
summary(m2)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Marili Zimmermann** (sünnikuupäev: 08.10.1993)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Elukestusanalüüs vasakult tõkestatud andmete ning ajast sõltuva argumenttunnuse korral TÜ Eesti geenivaramu kohordi näitel", mille juhendajad on Krista Fischer ja Nele Taba,

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

   (b) üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguste kehtivuse tähtaja lõppemiseni.

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

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

Tartus, 15.05.2018