Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kadri Kalamäe

Ühendatud logistiline regressioon elukestusandmete analüüsis TÜ Eesti geenivaramu andmete näitel

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: MSc Anastassia Kolde,
MSc Saskia Kuusk

Tartu 2025

ÜHENDATUD LOGISTILINE REGRESSIOON ELUKESTUSANDMETE ANALÜÜSIS TÜ EESTI GEENIVARAMU ANDMETE NÄITEL

Bakalaureusetöö

Kadri Kalamäe

Lühikokkuvõte

Bakalaureusetöö eesmärk on uurida kahte elukestusanalüüsi meetodit: Coxi võrdeliste riskide mudelit ja ühendatud logistilist regressiooni. Kuigi Coxi võrdeliste riskide mudel on laialt levinud ja palju kasutatud elukestusandmete analüüsimisel, on sellel tülikas võrdeliste riskide eeldus, mis pole praktikas tihtilugu täidetud. Töös analüüsitakse, millal ja miks võib ühendatud logistiline regressioon olla parem alternatiiv. Tartu Ülikooli Eesti geenivaramu andmete põhjal hinnatakse Coxi võrdeliste riskide ja ühendatud logistilise regressiooni mudelid, et uurida seoseid südame-veresoonkonna haiguste esinemisega pärast esimest kõrge kolesteroolitaseme mõõtmist.

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

Märksõnad: Elukestusanalüüs, geenivaramu, kardiovaskulaarsed haigused.

POOLED LOGISTIC REGRESSION IN SURVIVAL DATA ANALYSIS BASED ON DATA FROM THE ESTONIAN BIOBANK AT THE UNIVERSITY OF TARTU

Bachelor thesis

Kadri Kalamäe

Abstract

The aim of this bachelor's thesis is to study two survival analysis methods: the Cox proportional hazards model and the pooled logistic regression. Although the Cox proportional hazards model is widely used and popular in survival analysis, it relies on the proportional hazards assumption, which is often not met in practice. This thesis analyzes when and why pooled logistic regression might be a better alternative. Based on data from the Estonian Biobank at the University of Tartu, the Cox proportional hazards model and pooled logistic regression are applied to assess how predictors influence the risk of cardiovascular disease after the first high cholesterol measurement.

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

Key Words: Survival analysis, biobank, cardiovascular diseases.

Sisukord

Sissejuhatus

Bakalaureusetöö eesmärk on uurida kahte elukestusanalüüsi meetodit: Coxi võrdeliste riskide mudelit ja ühendatud logistilist regressiooni. Kuigi Coxi võrdeliste riskide mudel on laialdaselt levinud, on selle võrdeliste riskide eeldus väga tugev ning praktikas see sageli ei kehti. Alternatiivina on võimalik kasutada diskreetsel ajal põhinevat ühendatud logistilist regressiooni, millel ei ole Coxi mudelile omast võrdeliste riskide eeldust. Töö praktilises osas koostatakse Coxi võrdeliste riskide ja ühendatud logistilise regressiooni mudelid, et analüüsida Tartu Ülikooli Eesti geenivaramu andmete põhjal aega alates esimesest kõrge kolesterooli mõõtmisest kuni südame-veresoonkonna haiguse esinemiseni. Südame-veresoonkonna haigusena käsitletakse käesolevas töös haiguseid, mille RHK-10 koodid on I21 (äge müokardiinfarkt), I22 (korduv müokardiinfarkt), I63 (peaajuinfarkt, v.a I63.6) või surm südame-veresoonkonna haigusesse.

Südame- ja veresoonkonna ehk kardiovaskulaarsed haigused on kõige sagedasem surma põhjus kogu maailmas, põhjustades igal aastal hinnanguliselt 17,9 miljoni inimese surma (World Health Organization, 2025). Eestis suri 2023. aastal vereringeelundite haiguste tõttu ligi 7 500 inimest, mis moodustas umbes 47% kõigist surmajuhtumitest (Statistikaamet, 2024). Maailma Südameföderatsiooni hinnangul on 80% kardiovaskulaarsetest haigustest ennetatavad, kuna nende peamised riskitegurid on seotud elustiiliga: ebatervislik toitumine, vähene füüsiline aktiivsus ning tubaka ja alkoholi tarvitamine. Need tegurid võivad põhjustada kõrgenenud vererõhku, diabeeti, kõrget kolesterooli, samuti ülekaalu ja rasvumist. Haiguste ärahoidmiseks on oluline loobuda tubakast, toituda mitmekülgselt ja tervislikult, liikuda regulaarselt ning vältida alkoholi liigtarbimist. (World Heart Federation, 2025) Kardiovaskulaarhaiguste ennetuses on tähtis ka varajane diagnoosimine ja õigeaegne ravi alustamine. Eestis kasutatakse kõige enam just südame-veresoonkonna ravimeid, sealhulgas vere lipiidisisaldust kontrolli all hoidvaid aineid ning kardiovaskulaarseid tüsistusi ennetavaid ravimeid nagu statiine, mille kasutamine on vii-

mase kümne aasta jooksul kahekordistunud. (Ravimiamet, 2024)

Töö esimene peatükk tutvustab analüüsimetoodikat, kirjeldades elukestusanalüüsi olemust ja kasutatavaid meetodeid. Teises peatükis antakse ülevaade andmetest, viiakse läbi andmete kirjeldav analüüs ning esitatakse geenivaramu andmetel tehtud statistilise analüüsi tulemused.

1 Elukestusanalüüs

Juhul, kui pole märgitud teisiti, põhineb elukestusanalüüsi kirjeldav 1. peatükk David Colletti raamatul „Modelling Survival Data in Medical Research“ (Collett, 2014). Elukestusanalüüsis (*survival analysis*) uuritakse aega alates mingist kindlaks määratud alghetkest kuni huvipakkuva sündmuse toimumiseni või paika pandud lõppmomendini. Vastavat ajavahemikku nimetatakse elukestuseks ning andmeid kestusandmeteks. Meditsiinilistes uuringutes loetakse sageli algmomendiks indiviidi kaasamist katsesse ning huvipakkuvaks sündmuseks võib olla surm või sümptomite ilmnemine. Elukestusanalüüsi meetodid on aga rakendatavad ka teistes valdkondades ja erinevate sündmuste korral, näiteks inseneerias seadme rikkeni kuluva aja hindamiseks.

Kestusandmeid ei ole võimalik üldjuhul standardsete statistiliste meetodite abil analüüsida. Esiteks ei pärine kestusandmed normaaljaotusest, sest ajavahemike pikkused ei saa olla negatiivsed ning nende jaotus ei pruugi olla sümmeetriline, vaid on tihtipeale paremalt pikema sabaga. Teiseks on kestusandmetele iseloomulik tsenseeritus, millele keskendutakse lähemalt järgmises alapeatükis.

1.1 Tsenseeritud andmed

Indiviidi elukestus loetakse tsenseerituks, kui uuringu lõpphetkeks pole tal huvipakkuvat sündmust toimunud või kui indiviid on jälgimise alt millegipärast lahkunud või kadunud. Näiteks võib tsenseerimise põhjuseks olla katsealuse kolimine välismaale, sest seal ei ole teda võimalik enam jälgida. Eristatakse kolme tüüpi tsenseerimist: paremalt tsenseerimine, vasakult tsenseerimine ning intervallis tsenseerimine. Paremalt tsenseerimine on kõige sagedasem ja ka käesolevas bakalaureusetöös kasutatavad kestusandmed on paremalt tsenseeritud.

Sisenegu indiviid uuringusse hetkel t_0 ja toimugu huvipakkuv sündmus hetkel t_0+t, kusjuures t pole teada, sest uuritav on jälgimise alt kadunud või uuringu lõppedes

pole sündmust veel toimunud. Olgu $t_0 + c$ viimane hetk, millal indiviidi staatus on teada. Sellisel juhul nimetatakse aega c tsenseeritud elukestuseks ning tegu on paremalt tsenseeritusega, kus tsenseeritud elukestus on lühem kui tegelik elukestus, mis ei ole teada.

Tsenseeritud kestusandmete analüüsimisel eeldatakse, et indiviidi tegelik elukestus t ja tsenseeritud elukestus c, kus $c < t$, on sõltumatud. See tähendab, et tsenseerimine ei anna teavet selle kohta, millal huvipakkuv sündmus tegelikult toimub. Kirjeldatud omadust nimetatakse mitteinformatiivseks tsenseerimiseks.

1.2 Tähtsamad definitsioonid

Olgu T elukestust iseloomustav pidev juhuslik suurus, mis on alati mittenegatiivne. Olgu $f(t)$ selle juhusliku suuruse tihedusfunktsioon, siis T jaotusfunktsioon $F(t)$ defineeritakse järgmiselt:

$$F(t) = \mathrm{P}(T \leq t) = \int_0^t f(u)\,\mathrm{d}u.$$

Üleelamisfunktsioon $S(t)$ näitab tõenäosust, et elukestus on pikem kui t ning see avaldatakse eelmise võrrandi kaudu

$$S(t) = \mathrm{P}(T > t) = 1 - F(t).$$

Riskifunktsioon $h(t)$ iseloomustab riski, et sündmus toimub ajahetkel t tingimusel, et sündmust pole enne hetke t toimunud. Riskifunktsioon on defineeritud järgmiselt:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t}. \tag{1}$$

Riskifunktsiooni $h(t)$ võib tõlgendada kui oodatavat sündmuste arvu ajaühiku kohta, eeldades, et sündmus ei ole varem toimunud ja risk on ühes ajaühikus konstantne. Kasutades tingliku tõenäosuse valemit ja tuletise definitsiooni, on võimalik

näidata, et riskifunktsiooni ja üleelamisfunktsiooni vahel kehtib seos

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

1.3 Coxi võrdeliste riskide mudel

Coxi võrdeliste riskide mudel võimaldab uurida seoseid sõltumatute muutujate ja elukestuse vahel. Olgu vaatluse all n indiviidi ja p argumenttunnust $X_1, X_2, \ldots, X_p$, kusjuures nende tunnuste väärtuste vektor i-nda indiviidi jaoks tähistatakse $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$. Eeldatakse, et argumenttunnuste väärtused on teada uuringu algusest ning need ei muutu ajas. Funktsiooni, kus kõigi argumenttunnuste väärtused on võrdsed nulliga, tähistatakse $h_0(t)$ ja nimetataskse baasriskifunktsiooniks.

Riskifunktsioon i-nda indiviidi jaoks kirjutatakse kujul

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

kus $i = 1, 2, \ldots, n$ ja $\psi(\boldsymbol{x}_i)$ on vektorist $\boldsymbol{x}_i$ sõltuv funktsioon, mis näitab suhtelist riski i-nda indiviidi ja indiviidi vahel, kelle kõik argumenttunnused on võrdsed nulliga. Eelduse järgi on tunnuste väärtused fikseeritud terve jälgimisperioodi vältel, järelikult on $\psi(\boldsymbol{x}_i)$ konstantne. Kuna riskide suhe ei saa olla negatiivne, siis võib $\psi(\boldsymbol{x}_i)$ kirjutada eksponentfunktsioonina ning riskifunktsioon avaldub sel juhul järgmiselt:

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

kus β_j on j-ndale argumenttunnusele vastav mudeli parameeter ($j = 1, 2, \ldots, p$) ning $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ on parameetrite vektor. Funktsiooni $\psi(\boldsymbol{x}_i)$ kuju võib olla ka teistsugune, kuid kestusandmete modelleerimiseks on kõige levinum just ülaltoodud kuju (2).

Coxi võrdeliste riskide mudeli koostamiseks R tarkvaras saab kasutada *survival* paketti. Esmalt defineeritakse objekt *Surv(time, status)*, kus *time* tähistab jälgimisaega ja *status* on sündmuse toimumise indikaator, mis on 1, kui sündmus toimus, ja 0, kui ei toimunud. Seejärel sobitatakse mudel käsuga *coxph(Surv(time, status)∼x, data)*, kus ∼ märgist paremale koondatakse plussmärkidega eraldatult argumenttunnused. (Therneau, Lumley, Atkinson *et al.*, 2024)

Coxi võrdeliste riskide mudelit nimetatakse poolparameetriliseks mudeliks, sest baasriskifunktsiooni $h_0(t)$ kuju kohta ei tehta eeldusi, vaid hinnatakse mudeli parameetrid $\boldsymbol{\beta}$. Parameetrite hindamiseks kasutatakse suurima tõepära meetodit, seega on parameetrite hinnangud väärtused, mille korral tõepärafunktsioon saavutab oma maksimumi.

Olgu vaatluse all olevast n indiviidist sündmus toimunud r indiviidil ning ülejäänud $n-r$ indiviidi paremalt tsenseeritud. Eeldatakse, et sündmus ei saa toimuda mitmel indiviidil korraga samal ajal. Toimunud sündmuste järjestatud aegasid tähistatakse $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, kus $t_{(j)}$ on j-ndana toimunud sündmuse aeg. Märkigu $R(t_{(j)})$ indiviidide hulka, kes on hetkel $t_{(j)}$ riskigrupis ehk nendega pole huvipakkuv sündmus veel toimunud ning nad ei ole tsenseeritud. Tõepärafunktsioon, mille David Cox avaldas mudeli (2) jaoks, on kujul

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}, \tag{3}$$

kus $\boldsymbol{x}_{(j)}$ on argumenttunnuste väärtuste vektor indiviidil, kellega toimus sündmus momendil $t_{(j)}$. Kuna funktsioon $L(\boldsymbol{\beta})$ valemis (3) ei võta arvesse täpseid tsenseerimise ja sündmuse toimumise aegasid, vaid nende järjekorda, siis on tegu osalise tõepärafunktsiooniga. Coxi võrdeliste riskide mudeli parameetreid saab hinnata numbriliste meetodite abil.

1.4 Võrdeliste riskide eelduse kontroll

Coxi võrdeliste riskide mudeli rakendamisel eeldatakse, et riskide suhted on võrdelised ning ajast sõltumatud. Eelduse kontrollimiseks on kõige tõhusam meetod kaalutud Schoenfeldi jääkide analüüs.

Esmalt defineeritakse Schoenfeldi jäägid. Olgu vaatluse all n indiviidi, kellel on mõõdetud p tunnust $X_1, X_2, \ldots, X_p$. Lisaks olgu hinnatud Coxi võrdeliste riskide mudeli riskifunktsioon (2) ning saadud mudeli parameetrite vektori $\boldsymbol{\beta}$ hinnang $\hat{\boldsymbol{\beta}}$. Sümboliga δ_i tähistatakse indikaatortunnust, mis näitab, kas i-ndal indiviidil toimus sündmus (sellisel juhul $\delta_i = 1$) või on ta elukestus tsenseeritud (siis $\delta_i = 0$). Mudeli (2) j-nda muutuja X_j i-s Schoenfeldi jääk r_{Sji} avaldub kujul

$$r_{Sji} = \delta_i \left(x_{ji} - \frac{\sum_{l \in R(t_i)} x_{jl} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)} \right),$$

kus x_{ji} on i-nda indiviidi j-nda argumenttunnuse väärtus ($j = 1, 2, \ldots, p$) ning $R(t_i)$ on riskigrupp hetkel t_i. Schoenfeldi jääkide vektor i-nda vaatluse jaoks tähistatakse $\boldsymbol{r}_{Si} = (r_{S1i}, r_{S2i}, \ldots, r_{Spi})'$. On oluline tähele panna, et tsenseeritud vaatluste Schoenfeldi jäägid on nullid, sest tsenseeritud elukestuse korral on $\delta_i = 0$. Et eristada jääke, mis on nullid tsenseerimata vaatluste korral, kasutatakse enamasti tsenseeritud indiviidide jääkide puhul puuduvaid väärtusi.

Kaalutud Schoenfeldi jääkide vektor i-nda vaatluse jaoks koosneb p komponendist $r^*_{S1i}, \ldots, r^*_{Spi}$ ja esitatakse valemiga

$$\boldsymbol{r}^*_{Si} = d \, \mathrm{var}(\hat{\boldsymbol{\beta}}) \boldsymbol{r}_{Si},$$

kus d tähistab n indiviidi hulgas toimunud sündmuste arvu ja $\mathrm{var}(\hat{\boldsymbol{\beta}})$ on Coxi mudeli hinnatud parameetrite kovariatsioonimaatriks. Kaalutud Schoenfeldi jäägi

keskväärtuse jaoks i-nda indiviidi j-nda argumenttunnuse korral kehtib seos

$$E(r^*_{Sji}) \approx \beta_j(t_i) - \hat{\beta}_j,$$

kus $\beta_j(t_i)$ tähistab ajas muutuva kovariaadi X_j parameetrit i-nda indiviidi sündmuse toimumise ajahetkel t_i ja $\hat{\beta}_j$ on j-nda argumenttunnuse parameetri β_j hinnang. Seega seos $r^*_{Sji} + \hat{\beta}_j$ ja elukestuse vahel annab teavet ajas muutuva kovariaadi $\beta_j(t)$ kuju kohta. Horisontaalne joon viitab sellele, et parameeter on konstantne ja võrdeliste riskide eeldus on täidetud. Kaalutud Schoenfeldi jääke kasutab R-i funktsioon *cox.zph() survival* paketis, mis kontrollib, kas sirge tõus on võrdne nulliga (Therneau, Lumley, Atkinson *et al.*, 2024).

Coxi võrdeliste riskide mudeli eeldus, et riskide suhted on ajast sõltumatud ja võrdelised, on väga tugev. Praktikas eeldus sageli ei kehti, sest näiteks meditsiinilistes uuringutes võivad riskid varieeruda seetõttu, et inimeste vastuvõtlikkus haigustele on erinev ja ajas muutuv. (Kuitunen, Ponkilainen, Uimonen *et al.*, 2021)

1.5 Kihistatud Coxi võrdeliste riskide mudel

Kui Coxi võrdeliste riskide mudelis ei kehti võrdeliste riskide eeldus kõikide argumenttunnuste jaoks, siis kasutatakse tihti kihistatud võrdeliste riskide mudelit, kus kihistamine toimub eeldusi rikkunud argumenttunnuste järgi. Kihistamiseks kasutatud tunnuseid mudelisse argumenttunnustena ei lisata. Kihistatud mudelis on lubatud riskil muutuda vastavalt kihile, kuid võrdelisi riske eeldatakse kõikide argumenttunnuste korral. Teisisõnu, hinnatavad parameetrid $\boldsymbol{\beta}$ on igas kihis samad, kuid kihtide baasriskifunktsioonid erinevad. (Kleinbaum ja Klein, 2005)

Olgu k kihti, j-ndas kihis n_j indiviidi ning baasriskifunktsioon $h_{0j}(t),\ j = 1, \ldots, k$. Tähistades j-nda kihi i-nda indiviidi riskifunktsiooni $h_{ij}(t)$, kus $i = 1, \ldots, n_j$, siis

kihistatud Coxi võrdeliste riskide mudel p argumenttunnusega avaldub järgmiselt:

$$h_{ij}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})h_{0j}(t),$$

kus $\boldsymbol{x}_{ij}$ on j-nda kihi i-nda indiviidi argumenttunnuste väärtuste vektor. Parameetrite $\boldsymbol{\beta}$ hinnangud leitakse osalise tõepära maksimeerimise abil, maksimeeritava tõepärafunktsiooni leidmiseks peab korrutama omavahel kõikide kihtide tõepärafunktsioonid. (Kleinbaum ja Klein, 2005)

R tarkvaras saab Coxi võrdeliste riskide mudelisse lisada kihistatava tunnuse, kasutades käsku *strata(tunnus)*. See lisatakse mudeli argumenttunnustest plussmärgiga eraldatult ∼ märgist paremale poole. (Therneau, Lumley, Atkinson *et al.*, 2024)

1.6 Ajas muutuvad ja ajas fikseeritud kovariaadid

Elukestusanalüüsi mudelitesse saab kaasata nii ajas muutuvaid kui ajas fikseeritud kovariaate. Ajas fikseeritud muutujad mõõdetakse vaid ühel korral uuringu alguses ning saadud väärtuste põhjal analüüsitakse tunnuste mõju elukestusele. Ajas muutuvateks kovariaatideks loetakse tunnused, mille väärtused võivad uuringu jooksul muutuda ning mida mõõdetakse korduvalt. Eeldatakse, et kõige värskemad andmed annavad elukestuse kohta täpsema prognoosi kui uuringu alguses mõõdetud väärtused. Ajas muutuvaid kovariaate on kahte tüüpi. Esimesed on otseselt seotud uuritava indiviidiga ja neid saab mõõta ainult siis, kui indiviid on elus (*internal variables*), näiteks vererõhk või valgete vereliblede arv. Teise tüübi puhul (*external variables*) ei ole indiviidi elusolek mõõtmise eelduseks, sest nende väärtused on tulevikus kindlalt teada (näiteks indiviidi vanus) või ei sõltu otseselt uuritavast indiviidist (näiteks õhutemperatuur). Käesolevas bakalaureusetöös kasutatakse mudelites vaid ajas fikseeritud muutujaid.

1.7 Diskreetse ajaga meetodid

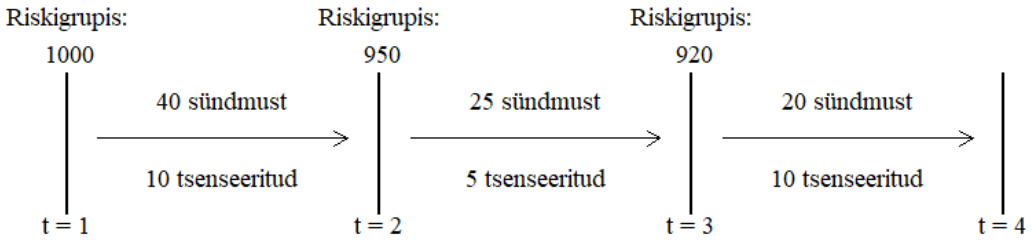
Diskreetse ajaga meetodeid käsitlev alapeatükk 1.7 põhineb Paul D. Allisoni artiklil „Discrete-Time Methods for the Analysis of Event Histories“ (Allison, 1982). Enamik mudeleid käsitleb aega pideva tunnusena, eeldades, et aeg võib võtta ükskõik millise mittenegatiivse väärtuse. Siiski võib mõnel juhul olla sobivam kasutada diskreetse ajaga meetodeid. Esiteks esineb sündmusi, mis toimuvad regulaarselt kindla aja tagant diskreetsetel ajahetkedel. Näiteks ülikoolist väljalangemise uurimisel tuleks kasutada diskreetse ajaga mudeleid, kuna väljalangemine registreeritakse iga semestri lõpus. Teine põhjus diskreetse ajaga mudelite rakendamiseks on olukord, kus sündmused võivad küll toimuda suvalisel ajahetkel, aga andmetes on saadaval vaid ajavahemik, millal sündmus toimus. Näiteks kui uuringus küsitakse vaid aastat, millal indiviid abiellus, kuid mitte täpset kuupäeva, siis ei ole sobilik käsitleda aega pideva tunnusena. Tegelikult on aeg andmetes alati diskreetne, olenemata sellest, kui suure täpsusega see on kirja pandud. Kui täpsus on väga suur võrreldes sündmuse toimumise sagedusega, võib aega käsitleda pideva suurusena. Kui ajaühikuks on aga kuu, aasta või kvartal, siis ei ole enamasti sobilik aega pidevaks tunnuseks lugeda.

Diskreetse ajaga mudelites eeldatakse, et aeg võtab naturaalarvulisi väärtusi ($t = 1, 2, 3, \ldots$). Olgu uuringu alguses ajahetkel $t = 1$ vaatluse all n sõltumatut indiviidi ($i = 1, 2, \ldots, n$). Iga indiviidi jälgitakse kuni ajahetkeni t_i, mil toimub huvipakkuv sündmus või tsenseerimine. Tsenseerimise korral on indiviid jälgimise all kuni ajani t_i, aga mitte enam hetkel $t_i + 1$. Olgu argumenttunnused ajas fikseeritud kovariaadid ning p argumenttunnuse väärtuste vektor i-nda indiviidi jaoks $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$. Diskreetse aja korral on riskifunktsioon analoogne riskifunktsiooniga (1) ning avaldub järgmiselt:

$$P_{it} = \mathrm{P}(T_i = t \mid T_i \geq t, \boldsymbol{x}_i),$$

kus T on sündmuste toimumise hetki kirjeldav diskreetne juhuslik suurus.

Joonis (1) illustreerib, kuidas vaatlusaluste arv diskreetse ajaga mudelites aja jooksul muutub. Olgu algselt (ajahetkel $t = 1$) jälgimise all 1 000 inimest, kellel mõõdetakse riskitegurid. Esimesel vaatlusperioodil toimub huvipakkuv sündmus 40 inimesel ja 10 inimest tsenseeritakse. Riskigrupist eemaldatakse nii tsenseeritud kui ka need inimesed, kellel toimus sündmus, järelikult teise vaatlusperioodi alguses hetkel $t = 2$ on riskigrupis 950 jälgitavat. Sama loogika kehtib ka järgmistel perioodidel.



Joonis 1: Riskigrupi suuruse muutumine ajapunktide kaupa.

1.8 Ühendatud logistiline regressioon

Üheks diskreetse ajaga meetodiks on ühendatud logistiline regressioon. Mudeli rakendamiseks on vajalik jagada aeg diskreetseteks ajaintervallideks (näiteks päev, kuu, kvartal või aasta). Olgu jälgimisperiood $T^* = (0, \tau]$ jagatud K võrdse pikkusega ajaintervalliks $(s_{k-1}, s_k]$, kus $k = 1, \ldots, K$, $s_0 = 0$ ja $s_K = \tau$. Nende intervallide abil saab teisendada andmed kujule, kus üks rida vastab ühele isikuajale (*person-time*). Sellist andmestikku nimetatakse sageli pikas formaadis või ühendatud andmestikuks, sest ühe inimese kohta on andmestikus nii mitu rida, kui mitmel vaatlusperioodil ta jälgimise all on. Selline lähenemine lihtsustab ühendatud logistilise regressiooni mudelisse ajas muutuvate kovariaatide lisamist, kuna igal ajaperioodil võivad argumenttunnuste väärtused muutuda. Lisaks peab andmetes olema indikaatortunnus, mille väärtus on 1, kui vaadeldavas intervallis toimus vaatlusalusel huvipakkuv sündmus, vastasel juhul 0. Näiteks kui indiviidil toimus sündmus 5. intervallis, siis esineb ta andmestikus viiel real, kusjuures esimese nelja

ajahetke korral on uuritava indikaatortunnuse väärtus 0 ja viiendal 1. (Zivich, Cole, Shook-Sa *et al.*, 2025)

Pikas formaadis andmetele saab seejärel rakendada logistilise regressiooni mudelit, mis hindab sündmuse tõenäosust intervallis tingimusel, et indiviidil ei ole enne k-ndat intervalli sündmust toimunud:

$$P_{ik} = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \theta_k)\right)}, \tag{4}$$

kus $P_{ik} = \mathrm{P}(T_i \in (s_{k-1}, s_k] \mid T_i > s_{k-1}, \boldsymbol{x}_i)$, $\beta_0, \beta_1, \ldots, \beta_p$ on hinnatavad parameetrid, $x_{1i}, \ldots, x_{pi}$ on i-nda uuritava argumenttunnuste väärtused ning θ_k on aja s_{k-1} mõju. Mudeli saab kirjutada ka *logit* kujul:

$$\ln \frac{P_{ik}}{1 - P_{ik}} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \theta_k.$$

Logit link on sobivaim juhul, kui sündmused võivad toimuda ainult diskreetsetel ajahetkedel, kuid seda saab rakendada ka olukordades, kus pidev aeg on jagatud intervallideks. Mudeli eksponeeritud koefitsiente saab tõlgendada šansside suhetena. (Zivich, Cole, Shook-Sa *et al.*, 2025; Allison, 1982; Allison, 2010)

Täiend-log-log (*cloglog*) link on sobivam siis, kui sündmused toimuvad tegelikult pidevalt ajas, kuid andmed on grupeeritud diskreetsetesse intervallidesse. Täiend-log-log seosefunktsiooniga logistilise regressiooni mudel kirjutatakse järgmiselt:

$$\ln(-\ln(1 - P_{ik})) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \theta_k$$

ehk alternatiivselt:

$$P_{ik} = 1 - \exp(-\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \theta_k)). \tag{5}$$

Selle mudeli koefitsiente saab eksponeerides tõlgendada kui riskide suhteid. Erinevalt *logit* lingist on *cloglog* funktsioon asümmeetriline. Näiteks on tõenäosuse

muutus vahemikus 0,25 kuni 0,50 *cloglog* skaalal suurem kui sama muutus vahemikus 0,50 kuni 0,75. Oluline on tähele panna, et *logit* funktsiooniga mudelis sõltuva muutuja väärtuste ümberpööramisel (näiteks sündmuse toimumise asemel sündmuse mittetoimumise modelleerimine) muutuvad koefitsiendid lihtsalt vastupidiseks. Täiend-log-log lingi korral saadakse ümberpööramise tulemusena täiesti erinevad parameetrite hinnangud. (Allison, 1982; Allison, 2010)

Aega võib modelleerida mitmel eri kujul. Juhul, kui perioode on palju, siis saab aega modelleerida pideva tunnusena. See ei pruugi aga ajas muutuvat riski hästi kajastada, sest siis eeldatakse, et šansi logaritm sõltub ajast lineaarselt. Teine variant on eeldada, et risk on igas ajaintervallis sama. Kolmas variant on käsitleda aega faktortunnusena ehk iga intervalli jaoks hinnatakse eraldi mudeli parameeter. Kui ajavahemikke on palju, siis peab hindama palju parameetreid ning sellisel juhul on heaks alternatiiviks splainide kasutamine. Splainid võimaldavad ajas muutuvaid riske modelleerida ilma liigse arvu parameetrite hindamiseta. (Zivich, Cole, Shook-Sa *et al.*, 2025)

Mudelite (4) ja (5) parameetrid $\boldsymbol{\beta}$ hinnatakse suurima tõepära meetodi abil seadmata piiranguid aja mõjule. Järgnevalt konstrueeritakse logaritmilise tõepärafunktsiooni kuju ühendatud logistilise regressiooni jaoks, mis kehtib nii *logit* kui *cloglog* seoselingi korral. Mudelite tõepärafunktsiooni saab kirjutada kujul

$$L = \prod_{i=1}^{n} (\mathrm{P}(T_i = t_i))^{\delta_i} (\mathrm{P}(T_i > t_i))^{1-\delta_i}, \tag{6}$$

kus δ_i on sündmuse indikaator (1, kui sündmus toimus, 0, kui andmed on tsenseeritud) ja n on vaatluste arv. Tõenäosused $\mathrm{P}(T_i = t)$ ja $\mathrm{P}(T_i > t)$ võib avaldada järgmiselt:

$$\mathrm{P}(T_i = t) = P_{it} \textstyle\prod_{j=1}^{t-1}(1 - P_{ij}), \qquad \mathrm{P}(T_i > t) = \textstyle\prod_{j=1}^{t}(1 - P_{ij}).$$

Asendades need tõepärafunktsiooni (6) ja võttes logaritmi, saadakse

$$\ln L = \sum_{i=1}^{n} \delta_i \ln \left(\frac{P_{it_i}}{1 - P_{it_i}} \right) + \sum_{i=1}^{n} \sum_{j=1}^{t_i} \ln(1 - P_{ij}).$$

Kui defineerida binaarne tunnus y_{it}, mis on 1, kui indiviidiga i toimub sündmus hetkel t, ja 0 muul juhul, saab eelneva seose esitada kujul

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{t_i} y_{it} \ln \left(\frac{P_{ij}}{1 - P_{ij}} \right) + \sum_{i=1}^{n} \sum_{j=1}^{t_i} \ln(1 - P_{ij}).$$

Siit järeldub, et ühendatud logistilise regressiooni logaritmiline tõepärafunktsioon on sama, mis tavalise logistilise regressiooni logaritmiline tõepärafunktsioon. (Allison, 1982)

Kui ajaintervallid on lühikesed ja risk igas intervallis on väike, siis Coxi võrdeliste riskide mudeli ja ühendatud logistilise regressiooni hinnatud parameetrid on väga sarnased (Cupples, D'Agostino, Anderson *et al.*, 1988). Ühendatud logistilise regressiooni kasutamisel on oluline mõista ka selle piiranguid. Pikka formaati teisendatud andmestiku maht võib olla väga suur, eriti juhul, kui uuritavaid indiviide on palju ning valitud ajaintervalli pikkus on lühike. Intervallide pikendamine (näiteks päevade asemel kuude kaupa) aitab andmestiku mahtu vähendada, kuid toob kaasa teabe kadumise, sest mudel ei kasuta täpset informatsiooni ajahetke kohta, millal huvipakkuv sündmus või tsenseerimine toimus. Näiteks sündmust, mis toimus intervalli alguses, kohtleb mudel samamoodi kui sündmust, mis toimus sama intervalli lõpu poole. (Zivich, Cole, Shook-Sa *et al.*, 2025; Ngwa, Cabral, Cheng *et al.*, 2016)

R-is saab ühendatud logistilise regressiooni koostada funktsiooniga *glm(formula, family = ..., data)*, kus *formula* määrab mudeli kuju (näiteks *sündmus* $\sim$ *tunnus* + ...) ja *family* argumendis määratakse sobiv seoselink. *Logit* link saadakse käsuga *binomial(link = „logit“)* ja täiend-log-log link käsuga *binomial(link = „cloglog“)*. (Mächler, 2025)

2 Andmete analüüs

Käesolevas bakalaureusetöös uuritakse Tartu Ülikooli Eesti geenivaramu (TÜ EGV) andmete põhjal aega alates esmakordsest kõrge tasemega kolesterooli mõõtmisest kuni esimese sündmuse toimumiseni järgmistest sündmustest:

- diagnoos I21 (äge müokardiinfarkt),
- diagnoos I22 (korduv müokardiinfarkt),
- diagnoos I63 (peaajuinfarkt, v.a I63.6) või
- surm südame-veresoonkonna haigusesse[1].

Edaspidi viidatakse loetletud sündmustele kokkuvõtvalt kui huvipakkuv sündmus või kardiovaskulaarsündmus. Vaatluse all on nii doonorid, kellele on määratud ravi, kui ka need, kellele ravimeid välja kirjutatud ei ole. Siinse töö raames tähendab ravi statiinravi ehk lipiidide sisaldust vähendavaid ravimeid, mis kuuluvad statiinide klassi. Statiinid on kliiniliselt kõige laialdasemalt kasutatav ravimirühm lipiidide mõjustamisel, sest need on tugeva LDL-kolesterooli sisaldust alandava toimega (Hedman ja Ristimäe, 2007). Uuring on teostatud väljastuse 6-7/GI/8344 raames ja saanud Eesti bioeetika ja inimuuringute nõukogu kooskõlastuse (24. märts 2020, nr 1.1-12/624).

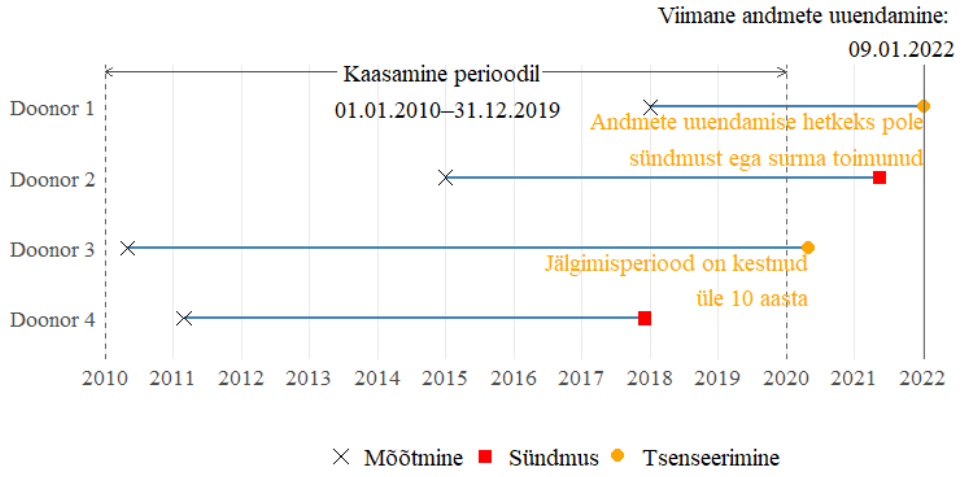
2.1 Andmete ülevaade

Andmete kogumine Tartu Ülikooli Eesti geenivaramusse algas 2002. aastal ning nüüdseks on biopangaga liitunud üle 210 000 eestimaalase ehk umbes 20% Eesti täisealisest elanikkonnast. Kõikidele doonoritele on tehtud pärilikkusaine ehk DNA analüüs, kuid lisaks geeniandmetele sisaldab geenivaramu ka doonorite terviseandmeid, mis võimaldavad uurida geenide ja haiguste vahelisi seoseid. (*Üldinfo. Tartu Ülikooli Eesti geenivaramu* 2025)

[1] RHK-10 (rahvusvahelise haiguste klassifikatsiooni) koodid, mille vasted on leitavad Sotsiaalministeerium, 2024 lehelt.

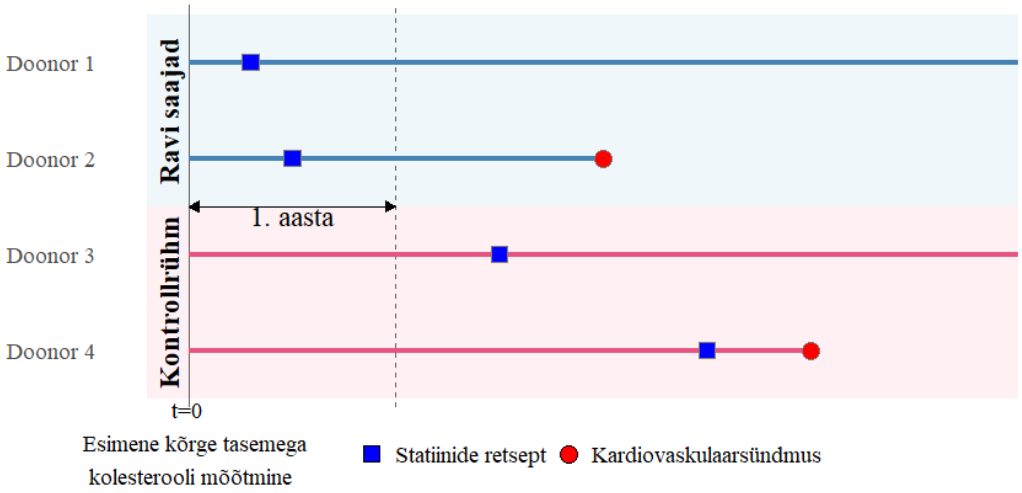
Töös kasutatav valim on koostatud TÜ EGV andmetest, lähtudes Kuusk, 2024 magistritöös „Sihtkatse matkimine: polügeense riskiskoori ja kolesterooli alandava ravi mõju südame-veresoonkonna haiguste riskile“ kirjeldatud eeskirjadest. Valimisse on kaasatud doonorid, kellel on mõõdetud kõrge tasemega kolesterool esmakordselt aastatel 2010–2019. Kolesterooli taset loetakse piisavalt kõrgeks kolesterooli alandava ravi saamiseks, kui LDL-kolesterool $\geq$ 4 mmol/L või kui LDL-kolesterooli mõõtmise puudumisel on üldkolesterool $\geq$ 6,5 mmol/L. Kõrge tasemega kolesterooli mõõtmistest pakub huvi vaid esimene, ülejäänud mõõtmised on eemaldatud.

Jälgimisaeg algab esimese kõrge tasemega kolesterooli mõõtmisest ning lõppeb järgnevatest sündmustest esimese toimumisel: huvipakkuv sündmus, surm või tsenseerimine. Tsenseerimisega on tegu siis, kui viimase andmete uuendamise hetkeks pole esinenud huvipakkuvat sündmust või surma või kui jälgimisperiood on kestnud üle kümne aasta. Doonorite jälgimisaeg kestab maksimaalselt 9. jaanuarini 2022, sest siis toimus viimane andmete uuendamine. Lisaks tuleb siinses töös arvestada vasakult tõkestatusega, sest uuringusse kaasatakse ainult need isikud, kes on uuringu alguses elus. Joonis (2) kujutab doonorite jälgimisperioode alates esmakordsest kõrge tasemega kolesterooli mõõtmisest kuni sündmuse või tsenseerimiseni.



Joonis 2: Doonorite kaasamine ja jälgimine.

Andmestik on puhastatud. Andmetest on eemaldatud juhud, kus statiini retsept on välja kirjutatud enne kõrge tasemega kolesterooli esmakordset mõõtmist või kardiovaskulaarsündmus on toimunud enne esimest kõrge tasemega kolesterooli mõõtmist. Doonor loetakse ravi saajate hulka, kui statiini retsept on välja kirjutatud aasta jooksul pärast esimest kõrge tasemega kolesterooli mõõtmist. Juhul, kui kolesterooli alandavat ravi pole välja kirjutatud või see on välja kirjutatud hiljem, siis arvestatakse geenidoonor ravi mittesaajaks ehk kontrollrühma. Joonisel (3) on näha, kuidas ravi välja kirjutamise aeg määrab, kas doonor kuulub ravi saajate hulka või kontrollgruppi. Näiteks kui doonoril mõõdeti esmakordselt kõrge tasemega kolesterool 2015. aasta septembris ja statiini ravi määrati talle 2016. aasta juulis, siis loetakse ta ravi saajate hulka. Kui aga ravi oleks välja kirjutatud 2016. aasta detsembris, siis kuuluks doonor kontrollgruppi.



Joonis 3: Ravi saajate ja kontrollgrupi määramine.

Kõigi geenidoonorite kohta on teada esimese kõrge tasemega kolesterooli mõõtmise kuupäev, sugu, sünniaasta ning indikaator statiinravi saamisest. Lisaks on iga doonori puhul olemas teave järgmistest tunnustest:

- indikaator I või II tüüpi diabeedi olemasolust,
- suitsetamise staatus,

- kehamassiindeks (KMI),
- haridustase,
- indikaator kroonilise neerupuudulikkuse diagnoosi olemasolust (CKD - *chronic kidney disease*),
- indikaator lipoproteiini ainevahetuse häire olemasolust,
- indikaator kõrgvererõhktõve olemasolust.

Haridustasemeks loetakse viimane teadaolev haridustase ning sellel on kolm taset: kuni põhiharidus, keskharidus ja kõrgharidus. Suitsetamise staatus on samuti jagatud kolmeks: mitte kunagi, endine ja praegune. Kehamassiindeksi väärtus ja suitsetamise tase on leitud kui mõõtmise kuupäevale lähim väärtus. Eespool loetletud tunnused ja lisaks ka geenidoonori sugu, vanus mõõtmisel ning mõõtmise aasta lisatakse mudelitesse segajatena (*confounders*), sest need mõjutavad nii ravi kui väljundit. Segajate mitte kaasamisel võib leida seoseid ravi ja väljundi vahel, mis tegelikult ei kehti. Lisaks kaasatakse mudelisse indikaatortunnus, mis näitab, kas inimene kuulub esimesse või teise Eesti geenivaramu kohorti, sest kohortides on märkimisväärselt erinev südame-veresoonkonnahaiguste risk ja üldsuremus.

2.2 Kirjeldav analüüs

Valim koosneb 14 233 geenidoonorist, kellest 4 719 (33,2%) on mehed ja 9 514 (66,8%) naised. Statiinravi saajaid on 1 236 ehk 8,7%. Keskmine vanus kõrge tasemega kolesterooli mõõtmise kuupäeval on ligikaudu 50,8 aastat, ravi saajate hulgas aga märgatavalt kõrgem – 57,5 eluaastat. Huvipakkuv sündmus esineb jälgimisperioodil 395 vaatlusalusel ehk ligikaudu 2,78% geenidoonoritest, kusjuures keskmiselt vanuses 67,1 eluaastat. Mediaanaeg esimesest kõrge kolesterooli taseme mõõtmisest kuni kardiovaskulaarsündmuse toimumiseni on 3,3 aastat. Jälgimisaeg (aeg mõõtmisest kuni huvipakkuva sündmuseni, surmani või tsenseerimiseni) on keskmiselt 5,6 aastat, ravi saanud doonorite seas aga 6,3 aastat. Tabel (1) võrdleb erinevaid

omadusi (nt vanus, erinevate haiguste olemasolu, haridustase jm) ravi saajate ja ravi mittesaajate hulgas. Lisas 1 esitatud tabel (5) toob välja samade näitajate väärtused pikas formaadis andmestikus, kus aeg on diskreetses formaadis.

Tabel 1: Algse andmestiku karakteristikud.

Tunnus		**Kontrollrühm** $n = 12\ 997$	**Ravi saajad** $n = 1\ 236$
Sugu	Naised	8 709 (67,0%)	805 (65,1%)
Vanus mõõtmisel		50,1 (12,6)	57,5 (10,8)
Üldkolesterool		6,8 (0,7)	7,0 (0,8)
LDL-kolesterool		4,5 (0,5)	4,8 (0,7)
Kardiovaskulaarne surm	Jah	95 (32,3%)	26 (42,6%)
Jälgimisaja pikkus		5,5 (2,3)	6,3 (2,8)
Sündmus	Jah	321 (2,5%)	74 (6,0%)
Diabeet	Jah	678 (5,2%)	202 (16,3%)
Suitsetamise staatus	Praegune	3 129 (24,1%)	301 (24,4%)
	Endine	3 041 (23,4%)	309 (25,0%)
	Mitte kunagi	6 827 (52,5%)	626 (50,6%)
KMI grupp	< 25	4 297 (33,1%)	281 (22,7%)
	$25 - 30$	4 954 (38,1%)	466 (37,7%)
	$30 - 35$	2 571 (19,8%)	314 (25,4%)
	35+	1 175 (9,0%)	175 (14,2%)
Haridustase	Kuni põhiharidus	550 (4,2%)	89 (7,2%)
	Keskharidus	6 077 (46,8%)	697 (56,4%)
	Kõrgharidus	6 370 (49,0%)	450 (36,4%)
CKD	Jah	46 (0,4%)	11 (0,9%)
Kõrgvererõhktõvi	Jah	4 268 (32,8%)	752 (60,8%)
Lipoproteiini ainevahetuse häire	Jah	565 (4,3%)	60 (4, 9%)
Kohort	2	8 828 (67,9%)	671 (54,3%)

Märkus: Arvuliste tunnuste puhul on tabelis välja toodud keskmine ja standardhälve, mittearvuliste tunnuste puhul sagedus ja osakaal.

Ühendatud logistilise regressiooni rakendamiseks moodustatakse algsete andmete põhjal pikas formaadis andmestik, kus iga doonor esineb nii mitmel real, kui mitu aastat ta jälgimise all on. Seega andmestiku iga rida vastab ühele isiku-aastale. Tabel (2) illustreerib, kuidas üksikvaatlused teisendatakse diskreetse ajaga pikas

formaadis andmestikuks. Doonor, kelle ID on 1, esineb pikas formaadis andmestikus kahel real. Kuna esimese aasta jooksul (01.01.2012 – 31.12.2012) ei toimu huvipakkuvat sündmust, siis sellele aastale vastavas reas on sündmuse indikaatortunnuse väärtus 0. Kuna sündmus toimub teisel aastal (01.01.2013 – 31.12.2013, täpsemalt 01.03.2013), siis on teise aasta reale märgitud sündmuse väärtus 1. Doonoril ID-ga 2, keda vaadeldakse kolmes ajaintervallis ja kellel huvipakkuvat sündmust ei toimunud, on pikas formaadis andmestikus igale aastale vastavas reas sündmuse väärtus 0.

Tabel 2: Algne andmestik.

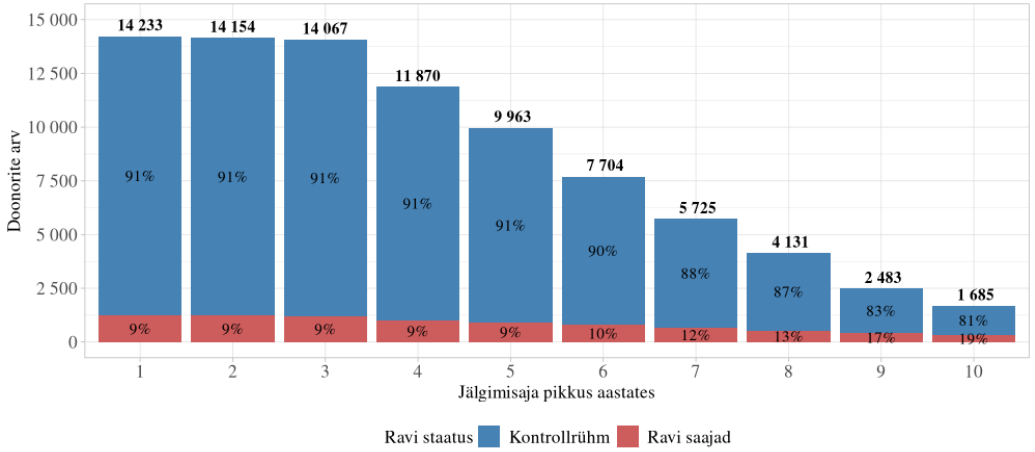
ID	**Ravi**	**Sugu**	**Mõõtmise kuupäev**	**Jälgimise lõpp**	**Sündmus**
1	1	Mees	01.01.2012	01.03.2013	1
2	0	Naine	30.06.2019	09.01.2022	0

Teisendatud pikk andmestik.

ID	**Ravi**	**Sugu**	**Aeg**	**Perioodi algus**	**Perioodi lõpp**	**Sündmus**
1	1	Mees	1	01.01.2012	31.12.2012	0
1	1	Mees	2	01.01.2013	01.03.2013	1
2	0	Naine	1	30.06.2019	29.06.2020	0
2	0	Naine	2	30.06.2020	29.06.2021	0
2	0	Naine	3	30.06.2021	09.01.2022	0

Kokku sisaldab pikas formaadis andmestik 86 015 vaatlust, millest 32,7% (28 159) moodustavad mehed ja 67,3% (57 856) naised. Statiinravi retsept on välja kirjutatud 8 314 doonorile ehk 9,67% uuritavatest. Keskmine vanus kõrge kolesteroolitaseme mõõtmisel on 50,5 aastat, ravi saajate seas aga kõrgem – 57,4 aastat. Joonisel (4) on kujutatud doonorite jaotust ning ravi saajate osakaalu vastavalt jälgimisaja pikkusele (aastates). Esimesel aastal panustab uuringusse 14 233 doonorit, kuid iga järgneva aastaga doonorite arv väheneb, kuna osa neist langeb vaatluse alt välja sündmuse toimumise, tsenseerimise või surma tõttu. Ravi saajate osakaal püsib esimese viie jälgimisaasta jooksul stabiilselt 9% juures, kuid hakkab seejärel

järk-järgult suurenema. 10 aastat jälgimise all olnud doonoritest 19% on uuringusse kaasatud ravi saajatena.



Joonis 4: Doonorite arv ja ravi saajate osakaal jälgimisaja jooksul.

2.3 Statistiline analüüs

Elukestusanalüüsi mudelitesse kaasatakse kolesterooli alandava ravi saamise indikaatortunnus, kohordi määrav binaarne tunnus ning eelnevalt välja toodud segajad, mille valik põhineb erialastel teadmistel. Kuna koostatud ühendatud logistilise regressiooni mudeleid kasutatakse hiljem sekkumise ehk ravi määramise tõenäosuse hindamiseks, siis eesmärk on tagada ravi saajate ja mittesaajate gruppide võrreldavus ning mudelitest ei eemaldata statistiliselt mitteolulisi tunnuseid.

Esmalt koostatakse Coxi võrdeliste riskide mudel ning kontrollitakse selle eelduse kehtivust. Kuigi praktikas see eeldus sageli ei kehti, on antud juhul see siiski täidetud. Edasi koostatakse kihistatud Coxi mudel, kus kihistamine toimub diskreetse aja tunnuse alusel, et mudel sarnaneks struktuurilt ühendatud logistilise regressiooniga – iga diskreetse ajahetke jaoks on eraldi baasriskitase. Ka kihistatud mudeli puhul on võrdeliste riskide eeldus täidetud ning saadud parameetrid on väga sarnased tavalise Coxi mudeli hinnangutega. Ühendatud logistilise regressiooni (ÜLR)

mudelites on ajavahemike modelleerimisel kasutatud kolme vabadusastmega splaine. Tabelisse (3) on koondatud mudelite hinnatud parameetrid. Kõigi mudelite puhul osutuvad samad argumenttunnused statistiliselt oluliseks.

Tabel 3: Mudelite hinnatud parameetrid.

	Tavaline Cox $\hat{\boldsymbol{\beta}}$	Kihistatud Cox $\hat{\boldsymbol{\beta}}$	ÜLR *logit* $\hat{\boldsymbol{\beta}}$	ÜLR *cloglog* $\hat{\boldsymbol{\beta}}$
Vabaliige			**255,062**	**252,825**
Aeg (splaini komponent 1)			−0,058	−0,051
Aeg (splaini komponent 2)			0,294	0,289
Aeg (splaini komponent 3)			−0,229	−0,224
Ravi: „Jah“	−0,085	−0,085	−0,083	−0,082
Sugu: „Naised“	**−0,673**	**−0,673**	**−0,681**	**−0,672**
Vanus mõõtmisel	**0,086**	**0,086**	**0,087**	**0,086**
Mõõtmisaasta	**−0,105**	**−0,105**	**−0,131**	**−0,130**
Diabeet: „Jah“	**0,525**	**0,525**	**0,530**	**0,519**
Suitsetamine: „Endine“	**−0,875**	**−0,875**	**−0,882**	**−0,875**
Suitsetamine: „Mitte kunagi“	**−0,814**	**−0,814**	**−0,820**	**−0,813**
KMI grupp: „25 − 30“	−0,025	−0,025	−0,023	−0,025
KMI grupp: „30 − 35“	0,161	0,161	0,163	0,160
KMI grupp: „35+“	0,217	0,217	0,206	0,204
Haridustase: „Keskharidus“	−0,265	−0,265	−0,270	−0,265
Haridustase: „Kõrgharidus“	**−0,385**	**−0,385**	**−0,388**	**−0,384**
CKD: „Jah“	**0,944**	**0,944**	**0,967**	**0,924**
Kõrgvererõhktõvi: „Jah“	**0,338**	**0,338**	**0,342**	**0,340**
Lipoproteiini ainevahetuse häire: „Jah“	0,093	0,093	0,093	0,092
Kohort: 2	**−0,285**	**−0,285**	**−0,284**	**−0,282**

Märkus: Rasvases kirjas on tähistatud parameetrid, mille $p \leq 0{,}05$.

Tavalise või kihistatud Coxi mudeli või täiend-log-log seosefunktsiooniga ühendatud logistilise regressiooni eksponeeritud parameetreid saab tõlgendada kui suhtelist riski. Kõigi kolme mudeli põhjal on kardiovaskulaarse sündmuse riski suurenemisega seotud järgmised tegurid: ravi puudumine, suitsetamine (võrreldes sellega, et oli endine suitsetaja või pole kunagi suitsetanud), madalam haridustase, kõrgem vanus esimesel kõrge tasemega kolesteroolitaseme mõõtmisel, diabeedi esinemine,

kehamassiindeksi kuulumine rühmadesse $30-35$ või $35+$ (võrreldes < 25 rühmaga), kroonilise neerupuudulikkuse olemasolu, kõrgvererõhktõvi ja lipoproteiini ainevahetuse häire. Meestel on risk kõrgem kui naistel. Lisaks selgub, et mida hilisemal aastal on esmakordselt diagnoositud kõrge kolesteroolitase, seda väiksem on risk kardiovaskulaarseks sündmuseks. Samuti on hilisemasse kohorti kuulumine seotud madalama riskiga. *Logit* seoselingiga koostatud ühendatud logistilise regressiooni eksponeeritud parameetreid tõlgendatakse šanssidena. Järeldub, et eespool loetletud tegurite korral on kardiovaskulaarse sündmuse šanss suurem, samas hilisemasse kohorti kuulumine ja hilisem mõõtmisaasta vähendavad šansse.

Tabel (4) võrdleb omavahel kihistatud Coxi võrdeliste riskide mudeli ja *logit* seosefunktsiooniga saadud ÜLR-i hinnatud parameetreid ning kihistatud Coxi võrdeliste riskide mudeli ja *cloglog* seosefunktsiooniga saadud ÜLR-i hinnatud parameetreid. Eesmärk on võrrelda protsentuaalset erinevust hinnatud kordajate vahel, mis arvutatakse järgmise valemi järgi:

$$R = \frac{\hat{\gamma} - \hat{\beta}}{\hat{\gamma}} \cdot 100,$$

kus $\hat{\gamma}$ ja $\hat{\beta}$ on vastavalt ühendatud logistilise regressiooni ja kihistatud Coxi võrdeliste riskide mudeli hinnatud parameetrid (D'Agostino, Lee, Belanger *et al.*, 1990).

Selgub, et enamik parameetreid on üksteisele väga sarnased, protsentuaalne erinevus jääb absoluutväärtuselt enamasti alla 5%. Kui ühendatud logistilises regressioonis on valitud ajaintervallid piisavalt lühikesed ja sündmuse risk on igas intervallis väike, siis ÜLR-i ja Coxi mudeli parameetrid tulevadki väga sarnased. Kõige rohkem erinevad antud juhul mudelite parameetrid mõõtmisaasta tunnuse jaoks, kui erinevus on peaaegu 20%.

Tabel 4: Mudelite hinnatud parameetrite võrdlus.

	Kihistatud Cox vs ÜLR *logit* (R, %)	Kihistatud Cox vs ÜLR *cloglog* (R, %)
Ravi: „Jah“	−2,4	−3,7
Sugu: „Naised“	1,2	−0,1
Vanus mõõtmisel	1,1	0
Mõõtmisaasta	19,8	19,2
Diabeet: „Jah“	0,9	−1,2
Suitsetamine: „Endine“	0,8	0
Suitsetamine: „Mitte kunagi“	0,7	−0,1
KMI grupp: „25 − 30“	−8,7	0
KMI grupp: „30 − 35“	1,2	−0,6
KMI grupp: „35+“	−5,3	−6,4
Haridustase: „Keskharidus“	1,9	0
Haridustase: „Kõrgharidus“	0,8	−0,3
CKD: „Jah“	2,4	−2,2
Kõrgvererõhktõvi: „Jah“	1,2	0,6
Lipoproteiini ainevahetuse häire: „Jah“	0	−1,1
Kohort: 2	−0,4	−1,1

Kokkuvõte

Bakalaureusetöö eesmärk oli uurida kahte elukestusanalüüsi meetodit: Coxi võrdeliste riskide mudelit ning ühendatud logistilist regressiooni. Esmalt kirjeldati elukestusanalüüsi olemust ning toodi välja, et üldjuhul ei ole võimalik kestusandmeid standardsete statistiliste meetodite abil analüüsida, sest need ei pärine normaaljaotusest ja neile on iseloomulik tsenseeritus.

Defineeriti Coxi võrdeliste riskide mudel, mis on kestusandmete analüüsimisel laialt levinud. Näidati, et tegu on poolparameetrilise mudeliga, kuna baasriskifunktsiooni kuju kohta ei tehta eeldusi, vaid osalise tõepärafunktsiooni abil hinnatakse mudeli parameetrid. Coxi võrdeliste riskide mudeli rakendamisel eeldatakse, et riskide suhted on võrdelised ning ajast sõltumatud. Eeldus on aga väga tugev ja praktikas see sageli ei kehti.

Seejärel käsitleti diskreetse ajaga elukestusanalüüsi meetodeid ning selgitati, millistes olukordades võiks neid eelistada mudelitele, mis käsitlevad aega pideva suurusena. Üheks diskreetse ajaga meetodiks on ühendatud logistiline regressioon. Mudeli rakendamiseks on vajalik jagada aeg diskreetseteks ajaintervallideks ja koostada pikas formaadis andmestik, kus ühe inimese kohta on andmestikus nii mitu rida, kui mitmel vaatlusperioodil ta jälgimise all on. Selline lähenemine lihtsustab ühendatud logistilise regressiooni mudelisse ajas muutuvate kovariaatide lisamist. Andmetele saab seejärel rakendada logistilise regressiooni mudelit, mis hindab sündmuse tõenäosust intervallis tingimusel, et indiviidil ei ole enne vaadeldavat intervalli sündmust toimunud. Mudeli parameetreid hinnatakse suurima tõepära meetodi abil, kusjuures näidati, et ühendatud logistilise regressiooni logaritmiline tõepärafunktsioon on sama, mis tavalise logistilise regressiooni logaritmiline tõepärafunktsioon. Kui ajaintervallid on lühikesed ja risk igas intervallis on väike, siis Coxi võrdeliste riskide mudeli ja ühendatud logistilise regressiooni hinnatud parameetrid on väga sarnased.

Töö praktilises osas analüüsiti Tartu Ülikooli Eesti geenivaramu andmete põhjal ae-

ga alates esimesest kõrge kolesterooli mõõtmisest kuni südame-veresoonkonna haiguse esinemiseni. Südame-veresoonkonna haigusena käsitleti haiguseid, mille RHK-10 koodid olid I21 (äge müokardiinfarkt), I22 (korduv müokardiinfarkt), I63 (peaajuinfarkt, v.a I63.6) või surm südame-veresoonkonna haigusesse. Koostati tavaline ja kihistatud Coxi võrdeliste riskide mudel ning *logit* ja täiend-log-log seosefunktsiooniga ühendatud logistilise regressiooni mudelid. Mudelitesse kaasati kolesterooli alandava ravi saamise indikaatortunnus ning erinevad segajad. Kuna Coxi mudeli võrdeliste riskide eeldus oli antud näites täidetud, ühendatud logistilise regressiooni ajaintervallid valiti piisavalt lühikesed ja risk oli igas intervallis väike, siis tulid Coxi mudeli ja ühendatud logistilise regressiooni hinnatud parameetrid üksteisele sarnased. Sellises olukorras võib kasutada mõlemat lähenemist. Kui aga võrdeliste riskide eeldus ei ole täidetud, ei ole Coxi mudeli tulemused enam usaldusväärsed. Sellisel juhul võib ühendatud logistiline regressioon olla hea alternatiiv, kuna see ei eelda võrdelisi riske ja on tehniliselt lihtne rakendada. Mudelites kasutati vaid ajas fikseeritud muutujaid, kuid töö edasiarendusena saaks argumenttunnustena kaasata ajas muutuvaid kovariaate.

Kasutatud allikad

Allison, P. D. (1982). "Discrete-Time Methods for the Analysis of Event Histories". *Sociological Methodology* 13, lk. 61–98. URL: https://doi.org/10.2307/270718.

– (2010). *Survival Analysis Using SAS: A Practical Guide*. 2. väljaanne. Cary, NC: SAS Institute Inc.

Collett, D. (2014). *Modelling Survival Data in Medical Research*. 3. väljaanne. New York: Chapman ja Hall/CRC.

Cupples, L. A., R. B. D'Agostino, K. Anderson *et al.* (1988). "Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study". *Statistics in medicine* 7.1-2, lk. 205–218. URL: https://doi.org/10.1002/sim.4780070122.

D'Agostino, R. B., M.-L. Lee, A. J. Belanger *et al.* (1990). "Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study". *Statistics in Medicine* 9.12, lk. 1501–1515. DOI: https://doi.org/10.1002/sim.4780091214.

Hedman, A. ja T. Ristimäe (2007). "Statiinid ja ateroskleroos". *Eesti Arst* 86.3, lk. 157–160. URL: https://doi.org/10.15157/ea.v0i0.10190.

Kleinbaum, D. G. ja M. Klein (2005). "The Stratified Cox Procedure". Teoses: *Survival Analysis: A Self-Learning Text*. Springer New York, lk. 173–210. URL: https://doi.org/10.1007/0-387-29150-4_5.

Kuitunen, I., V. T. Ponkilainen, M. M. Uimonen *et al.* (2021). "Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review". *BMC Musculoskelet Disord* 22.489. URL: https://doi.org/10.1186/s12891-021-04379-2.

Kuusk, S. (2024). "Sihtkatse matkimine: polügeense riskiskoori ja kolesterooli alandava ravi mõju südame-veresoonkonna haiguste riskile". Magistritöö. Tartu Ülikool.

Mächler, M. (2025). *Fitting Generalized Linear Models*. URL: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html.

Ngwa, J. S., H. J. Cabral, D. M. Cheng *et al.* (2016). "A comparison of time dependent Cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study". *BMC Medical Research Methodology* 16.148. URL: https://doi.org/10.1186/s12874-016-0248-6.

Ravimiamet (2024). *Südame-veresoonkonna preparaadid on Eestis enimkasutatavad ravimid*. URL: https://www.ravimiamet.ee/uudised/sudame-veresoonkonna-preparaadid-eestis-enimkasutatavad-ravimid (vaadatud 16.04.2025).

Sotsiaalministeerium (2024). URL: https://rhk.sm.ee/ (vaadatud 13.04.2025).

Statistikaamet (2024). *Surmad*. URL: https://www.stat.ee/et/avasta-statistikat/valdkonnad/rahvastik/surmad (vaadatud 22.04.2025).

Therneau, T. M., T. Lumley, E. Atkinson *et al.* (2024). *Package 'survival'*. URL: https://cran.r-project.org/web/packages/survival/survival.pdf.

World Health Organization (2025). *Cardiovascular diseases*. URL: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (vaadatud 22.04.2025).

World Heart Federation (2025). *Prevention*. URL: https://world-heart-federation.org/what-we-do/prevention/ (vaadatud 22.04.2025).

Zivich, P. N., S. R. Cole, B. E. Shook-Sa *et al.* (2025). "Estimating equations for survival analysis with pooled logistic regression". URL: https://doi.org/10.48550/arXiv.2504.13291.

Üldinfo. Tartu Ülikooli Eesti geenivaramu (2025). URL: https://geenidoonor.ee/geenivaramu (vaadatud 09.04.2025).

Lisa 1. Andmestiku karakteristikud

Tabel 5: Pikas formaadis diskreetse ajaga andmestiku karakteristikud.

Tunnus		**Kontrollrühm** $n = 77\ 701$	**Ravi saajad** $n = 8\ 314$
Sugu	Naised	52 412 (67,5%)	5 444 (65,5%)
Vanus mõõtmisel		49,8 (12,5)	57,4 (10,7)
Üldkolesterool		6,8 (0,7)	7,0 (0,8)
LDL-kolesterool		4,5 (0,5)	4,8 (0,7)
Kardiovaskulaarne surm	Jah	397 (30,1%)	140 (42,2%)
Jälgimisaja pikkus		6,4 (2,3)	7,4 (2,5)
Sündmus	Jah	321 (0,4%)	74 (0,9%)
Diabeet	Jah	4 113 (5,3%)	1 348 (16,2%)
Suitsetamise staatus	Praegune	18 906 (24,3%)	1 980 (23,8%)
	Endine	17 813 (22,9%)	2 037 (24,5%)
	Mitte kunagi	40 982 (52,7%)	4 297 (51,7%)
KMI grupp	< 25	25 499 (32,8%)	1 858 (22,3%)
	$25 - 30$	29 540 (38,0%)	3 134 (37,7%)
	$30 - 35$	15 575 (20,0%)	2 186 (26,3%)
	$35+$	7 087 (9,1%)	1 136 (13,7%)
Haridustase	Kuni põhiharidus	3 259 (4,2%)	607 (7,3%)
	Keskharidus	36 287 (46,7%)	4 606 (55,4%)
	Kõrgharidus	38 155 (49,1%)	3 101 (37,3%)
CKD	Jah	288 (0,4%)	77 (0,9%)
Kõrgvererõhktõvi	Jah	25 686 (33,1%)	5 020 (60,4%)
Lipoproteiini ainevahetuse häire	Jah	3 418 (4,4%)	413 (5, 0%)
Kohort	2	52 602 (67,7%)	4 441 (53,4%)

Märkus: Arvuliste tunnuste puhul on tabelis välja toodud keskmine ja standardhälve, mittearvuliste tunnuste puhul sagedus ja osakaal.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kadri Kalamäe,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Ühendatud logistiline regressioon elukestusandmete analüüsis TÜ Eesti geenivaramu andmete näitel“, mille juhendajad on Anastassia Kolde ja Saskia Kuusk, 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.

Kadri Kalamäe
15.05.2025