TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

**Greete Kelli Aava**

# R-pakett OMOP CDM kujul andmete elukestusanalüüsiks

**Bakalaureusetöö (9 EAP)**

Juhendaja: Markus Haug, MSc

Tartu 2024

## R-pakett OMOP CDM kujul andmete elukestusanalüüsiks

**Lühikokkuvõte:**

Bakalaureusetöö keskendub elukestusanalüüsi tööriista loomisele *Observational Medical Outcomes Partnership* (OMOP) ühtse andmemudeli kujul terviseandmetele. Eesmärgiks on luua R-pakett, mille töövoog sisaldab andmebaasipäringute loomist, teostamist ja saadud vastuste visualiseerimist kasutajaliideses. Töö jaotub teoreetiliseks, kus tutvustatakse elukestusanalüüsi metoodikat ja ühtset andmemudelit, ning praktiliseks osaks, kus kirjeldatakse loodud R-paketti ja selle võimalusi.

**Võtmesõnad:**

Programmeerimiskeel R, R-pakett, elukestusanalüüs, OMOP CDM, Kaplan-Meieri kõver

**CERCS:** P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika

## R-package for survival analysis on OMOP CDM databases

**Abstract:**

The Bachelor's thesis focuses on the creation of a survival analysis tool for health data in the form of the Observational Medical Outcomes Partnership (OMOP) common data model. The goal is to create an R package that includes database query generation, execution of database queries and visualization of results with a graphical user interface. The work is divided into a theoretical part, where the survival analysis methodology and a common data model are introduced, and a practical part, where the created R-package and its capabilities are described.

**Keywords:**

R programming language, R-package, survival analysis, OMOP CDM, Kaplan-Meier curve

**CERCS:** P160 Statistics, operation research, programming, actuarial mathematics

# Sisukord

# Sissejuhatus

Tõenduspõhiste otsuste olulisus meditsiinisüsteemis ja terviseandmete digitaalne hoiustamine on terviseandmetega seotud uuringutes kaasanud informaatikat ja andmeteadust. Hripcsak ja Ryan on kirjutanud, et tõenduspõhiste otsuste tegemine olemasolevate terviseandmete põhjal on keeruline protsess. Autorid on lisanud, et protsess nõuab erineva taustaga spetsialistide koostööd, näiteks statistikud ja tarkvaraarendajad (Ryan & Hripcsak, 2021). Hripcsak on kirjutanud, et 2014. aastal asustatud avatud teaduse kogukond *Observational Health Data Sciences and Informatics* (OHDSI) on võtnud eesmärgiks parendada inimeste tervist ja heaolu informaatika toel. Autor kirjeldas tehtuid töid ja põhimõtteid, kus keskseteks on *Observational Medical Outcomes Partnership Common Data Model* (OMOP CDM) ühtse andmemudeli kasutamine, vabavaraliste tööriistade loomine mudeliga töötamiseks ning reprodutseeritavate metoodikate väljatöötamine (Hripcsak et al., 2015).

Elukestusanalüüs on populaarne andmete analüüsimeetod, Flynn on välja toonud oma artiklis, et metoodikat kasutatakse nii meditsiinivaldkonnas kliinilistes ja vaatluspõhistes uuringutes kui ka väljaspool tervisevaldkonda, näiteks juhtudel, kus uurimisobjektiks on kahe sündmuse vaheline periood (Flynn, 2012). 2023. aastal avaldati López-Güelli juhtimisel loodud R-pakett *cohortSurvival*, mis ühildub OMOP CDM andmekujuga ning teostab elukestusanalüüsi varasemalt defineeritud kohortidel (López-Güell et al. 2023).

Bakalaureusetöö eesmärk on arendada R-pakett, mis võimaldaks elukestusanalüüsi teostamist OMOP CDM andmemudelil ja kuvada graafilisi tulemusi kasutajaliideses. Pakett sisaldab funktsionaalsust andmebaasipäringute koostamiseks ja päringute teostamiseks. Elukestusanalüüsi visualiseerimiseks luuakse liides, kus kuvatakse Kaplan-Meieri ja kumulatiivse riskifunktsiooni kõveraid ning teisi statistiliselt olulisi näitajaid.

Käesolev töö jaguneb neljaks peatükiks. Esimeses peatükis tutvustatakse elukestusanalüüsi ja selle statistilisi meetodeid. Teises peatükis on kokkuvõte ühtsest andmemudelist ja sellega seotud töödest. Kolmandas peatükis antakse ülevaade bakalaureusetöö raames valminud praktilisest väärtusest. Neljandas peatükis viiakse läbi näidisuuring ning analüüsitakse loodud tööriista võimekust.

# 1 Metoodika

Käesolev peatükk katab bakalaureusetöö teoreetilist sisu, mis hõlmab sissejuhatust elukestusanalüüsi ja selle statistilistesse meetoditesse. Alampeatükkide kirjutamisel on lähtutud D. Collett ja A. Kimberi kirjutatud raamatust "Modelling survival data in medical research" (2014), kui ei ole märgitud teisiti.

## 1.1 Elukestusanalüüs

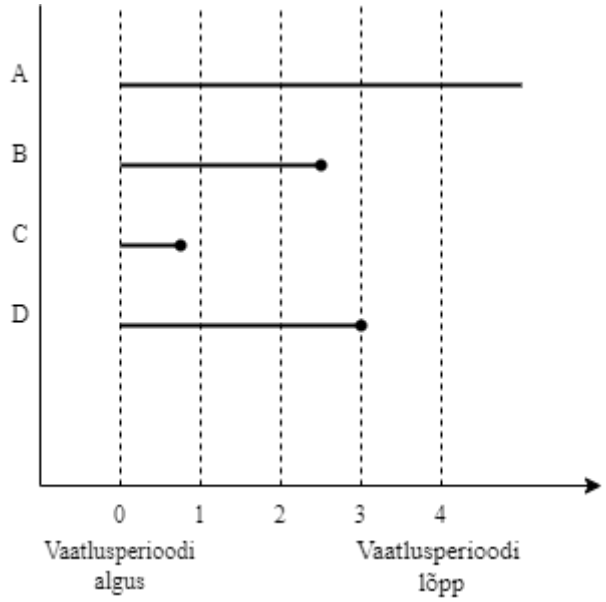
Elukestusanalüüs (ingl *survival analysis*) on Kleinbaumi ja Kleini poolt sõnastatud kui andmeanalüüsi metoodikate kogum, mis on kohandatud uurimaks erinevate juhtude aega mingi sündmuse toimumise või mitte toimumiseni (Kleinbaum & Klein, 2005). Collett ja Kimber on öelnud, et aja analüüsimisel on tähtis kaasata ka ajaperioodi algust märkiv sündmus. Ajaperioodi piiritlemine alustava ja lõpetava sündmusega määrab uuringu sisu. Samuti kirjeldasid Collett ja Kimber elukestusanalüüsi kasutusvõimalusi meditsiinilistes uuringutes ja tehnikaseadmete vastupidavuse uurimisel, analüüsitavaid ajaperioode võivad märkida vastavalt haigestumine kuni suremine ja tehnika kasutuselevõtt kuni pöördumatu vea teke.

Käesolevas bakalaureusetöös alustab ja lõpetab ajaperioodi terviseandmetes registreeritud sündmus, kusjuures mõlemad sündmused on kasutaja poolt defineeritavad. Sellisteks sündmuste paarideks sobivad näiteks mõne haiguse avaldumine ja suremine või diagnoosi saamine ja ravi alustamine. Elukestusanalüüs hõlmab riskifunktsiooni, kumulatiivset riskifunktsiooni ja üleelamisfunktsiooni, lisaks kasutatakse siin töös Kaplan-Meieri mitteparameetrilist kõverat elukestuse visualiseerimiseks ja log-rank testi Kaplan-Meieri kõverate võrdlemiseks. Järgnevalt kirjeldatakse eelmainitud funktsioone ja terviseandmete eripära lähemalt.

## 1.2 Tsenseeritus

Colletti ja Kimberi väitel pole terviseandmed sageli normaaljaotusega, lisaks on vaatlusandmed enamasti ebatäielikud ehk tsenseeritud. Kleinbaumi ja Kleini sõnul on elukestusanalüüsis osaleja andmed tsenseeritud siis, kui vaatlusandmetes ei ole registreeritud uuringu raames defineeritud huvipakkuva sündmuse toimumishetk (Kleinbaum & Klein, 2005).

Collett ja Kimber on teinud ülevaate elukestusanalüüsis esinevatest tsenseerimis võimalustest, mida järgnevalt kokkuvõtvalt kirjeldatakse. Andmed saavad olla vasakult, paremalt ja intervall tsenseeritud. Vasakult tsenseeritud on sellised andmed, kus huvipakkuv sündmus toimub aga sündmuse toimumisaeg on enne vaatlusperioodi lõppu, jättes toimumishetke täpselt määramata. Alloleval joonisel (Joonis 1) kirjeldab juht C vasakult tsenseeritust, kus juhtum toimub enne ajahetke üks. Intervall tüüpi tsenseerimisel, sarnaselt vasakult tsenseerimisega on määramata huvipakkuva sündmuse toimumishetk aga on teada ajavahemik vaatlusperioodist, mil sündmus toimus. Joonisel 1 tähistab intervall tsenseeritust näide B, kus sündmus toimub ajahetkede kaks ja kolm vahel. Enimlevinud, paremalt tsenseerimine kirjeldab olukorda, kus huvipakkuv sündmus ei toimunud uuringu jooksul või sellekohane teave puudub. Joonisel 1 tähistab paremalt tsenseeritust näide A, kus sündmus uuringu vaatlusajajooksul ei toimu. Viimase korral kirjeldab tsenseeritud osalejate elukestust hiliseim ajahetk, mil osaleja uuringus oli. Joonisel 1 tähistab juhtum D olukorda, kus tsenseeritust ei esine, ehk huvipakkuv sündmus toimus ning selle toimumishetk on teada.




Joonis 1. Näitejoonis, millel on alates ülevalt kujutatud paremalt tsenseerimine, intervall tsenseerimine, vasakult tsenseerimine ja tsenseerimata näide.

Emory Ülikooli epidemioloogide Kleinbaumi ja Kleini sõnul puutuvad kõik meditsiinilised elukestusanalüüsid kokku paremalt tsenseeritud andmetega, põhjusteks on järgnevad uuritava sündmusega seotud asjaolud: huvipakkuv sündmus ei toimunud uuringus osalejaga, osaleja taganeb uuringust enne sündmuse toimumist või elukestusuuringu osaleja sureb enne sündmust uuringu jaoks ebaolulistel põhjustel (Kleinbaum & Klein, 2005).

Tehtud töö praktilises osas arendati R-paketti Eestis kogutud terviseandmetel, mis olid paremalt tsenseeritud, sisaldades juhtusid, kus valitud sündmus ei toimunud patsiendiga. Seega loodud visualiseerimisvõimalus toetab paremalt tsenseeritud andmeid.

## 1.3 Elukestusfunktsioon ja riskifunktsioon

Kleinbaumi ja Kleini seisukohalt on elukestusanalüüsi keskseteks funktsioonideks elukestusfunktsioon ja riskifunktsioon (Kleinbaum & Klein, 2005). Collett ja Kimber on lisanud nende sekka kumulatiivse riskifunktsiooni. Lisaks on nad selgitanud, et kõik kolm funktsiooni põhinevad elukestusanalüüsi vaatlusandmetel. Funktsioonidel on kokkuleppeliselt järgnevad sümbolid:

- $T$, mis tähistab uuringus osaleja kogu elukestust;
- $t$, mis tähistab suvalist ajahetke osaleja elukestuse $T$ jooksul;
- $\Delta t$, mis tähistab ajamuutu.

Elukestusfunktsioon (ingl *survivor function*) on Colletti ja Kimberi sõnastuse kohaselt andmetel leitav tõenäosusfunktsioon, et huvipakkuv sündmus ei toimu enne sisendina saadud ajahetke. Elukestusfunktsiooni esitatakse järgmisel kujul:

$$S(t) = P(T \geq t) \quad (1).$$

Kleinbaum ja Klein (2005) on oma definitsioonile lisanud, et elukestusfunktsiooni kohta kehtivad järgmised omadused:

- joonistuv graafik on monotoonselt mittekasvav;
- ajahetkel t = 0 on S(t) väärtuseks alati 1;
- ajahetkel t = $\infty$ on S(t) väärtuseks alati 0.

Oluline on märkida, et ajahetke, kus $t$ on võrreldav lõpmatusega, praktikas ei teki.

Riskifunktsiooni (ingl *hazard function*) kirjeldavad Collett ja Kimber kui võimalust, et huvipakkuv sündmus toimub suvalisel ajahetkel $t$, funktsioon avaldub järgnevalt

$$h(t) = \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \quad (2).$$

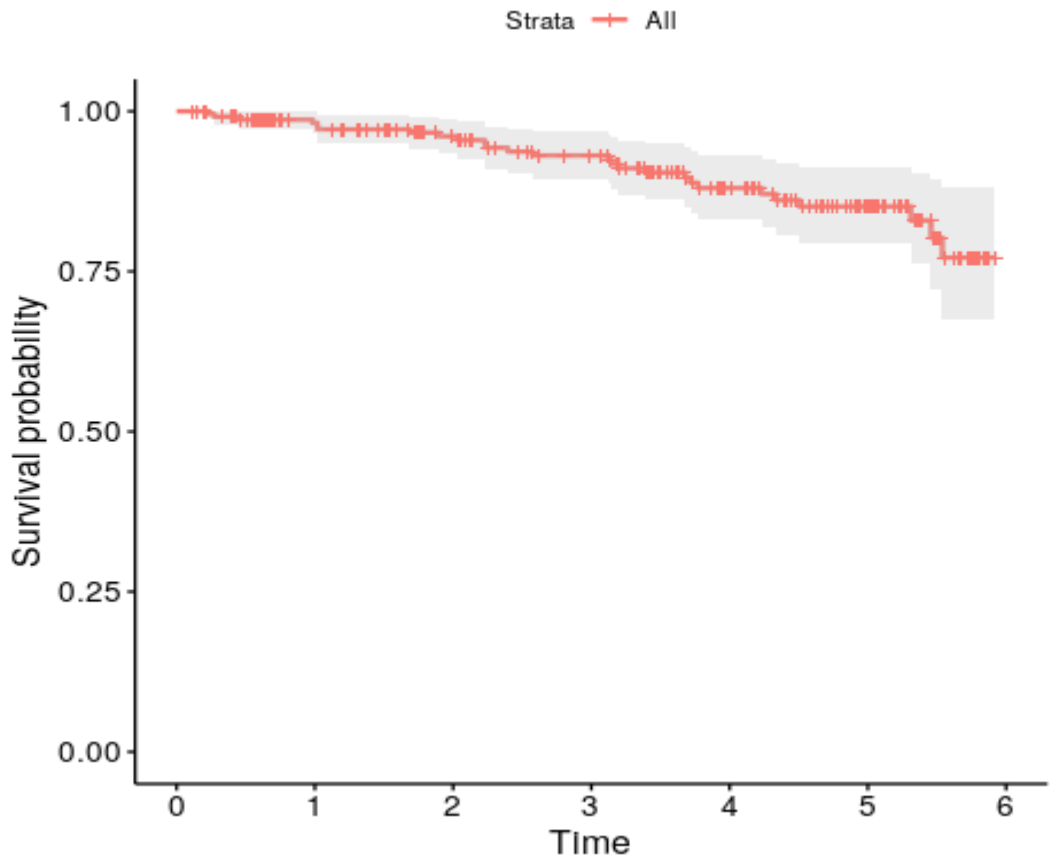
Valemi kohaselt eeldatakse, et elukestus $T$ on vähemalt sama suur kui ajahetk $t$ ehk sündmus ei ole varasemalt toimunud. Seejärel leitakse tõenäosus, et sündmus toimub vähemalt ajahetke $t$ ja t + $\Delta t$ vahel, kus $\Delta t$ ühtlasi läheneb nullile.

Kumulatiivne riskifunktsioon ja elukestusfunktsioon on vahetult seotud ning Collett ja Kimber tõestasid seda oma väljaandes. Nad kirjutasid, et kumulatiivne riskifunktsioon H(t) näitab tõenäosust, et sündmus toimub enne ajahetke *t* ning avaldub kujul

$$H(t) = \quad - \log S(t) \tag{3}.$$

## 1.4 Kaplan-Meieri meetod

Kaplan-Meieri kõverat (vt joonis 2), mis põhineb samanimelisel hinnangul, kasutatakse elukestusanalüüsis kirjeldamaks elukestust tuginedes kogutud andmetele ning arvestades ebatäielike vaatlusandmetega (Kleinbaum & Klein, 2005).




Joonis 2. Kaplan-Meieri kõvera graafik loodud R-pakettiga sooliselt eristamata ja vanusevahemikus 29-42 andmetel.

Järgnev lõik põhineb Kaplani ja Meieri poolt avaldatud artiklil "Nonparametric Estimation from Incomplete Observations" (1958). Kaplan-Meieri kõver (1958) joonistub samanimelise hinnangu põhjal. Meetodi kirjeldamiseks kasutatakse järgnevaid sümboleid:

- *j*, mis tähistab uuringu ajahetke vaatlusperioodist;
- $n_j$, mis tähistab ajahetke j kogusündmuste arvu;
- $n_j'$, mis on võrdne kogu sündmuste ja huvipakkuvate sündmuste vahega ajahetkel *j*;
- *k*, mis tähistab eristatud ajahetkede kogu arvu.

Kaplan-Meieri hinnang on defineeritud järgnevalt

$$\hat{P}(t) = \prod_{j=1}^{k} \frac{n_j'}{n_j} \quad (4).$$

Kaplan-Meieri kõverat on Kleinbaum ja Klein kirjeldanud kui langevat astmelist kõverat. Oma töös seletavad nad, et iga ajahetke kohta leitud tõenäosus lisatakse graafikule ja huvipakkuvate sündmuste mitteesinemine tekitab astmelised platood (vt joonis 2) (Kleinbaum & Klein, 2005).

## 1.5 Log-rank test

Elukestusanalüüsi raames tekkivad graafikud on omavahel analüütiliselt võrreldavad. Mitme Kaplan-Meieri kõvera ekvivalentsuse hindamiseks kasutatakse log-rank testi.

Järgnev lõik kirjeldab Kleinbaumi ja Kleini raamatu peatükki, milles kirjeldatakse kahe Kaplan-Meieri kõvera sarnasuse hindamist log-rank testi abil. Testi aluseks on nullhüpotees, et kaks kõverat on statistiliselt identsed. Nullhüpotees kummutatakse, kui leitava statistiku väärtus on väiksem kui 0.05 või muu etteantud piir. Statistiku arvutamise valem kahe kõvera *i* (i = 1, 2) jaoks on järgmine:

$$Log - rank\ statistik = \frac{(O_i - E_i)^2}{Var(O_i - E_i)} \quad (5).$$

Eelnevalt kirjeldatud valemis tähistavad $O_i$ ja $E_i$ kõverate *i* (i = 1, 2) korral vastavalt kõigi eristatavate perioodide jooksul toimunud huvipakkuvate sündmuste arvu ja kõigi eristatavate perioodide jooksul toimunud eeldatud juhtude arvu. Lugeja on summa kõigi uuritud ajaperioodide jooksul toimunud ja eeldatud sündmuste vahest ning avaldub kujul

$$O_i - E_i = \sum_{j=1}^{k} (m_{ij} - e_{ij}) \quad (6).$$

Eeldatud sündmuste arvutamine toimub kõvera *i* (i = 1, 2), kus sümbol *v* (v = 1, 2) tähistab kõverat kui i ja sümbol tähistab *j* eristatavat ajahetke, jaoks järgnevalt

$$e_{ij} = \left(\frac{n_{ij}}{n_{ij}+n_{vj}}\right)(m_{ij} + m_{vj}) \quad (7).$$

Nimetaja avaldub kujul

$$Var\left(O_i - E_i\right) = \sum_j \frac{n_{1j}n_{2j}(m_{1j}+m_{2j})(n_{1j}+n_{2j}-m_{1j}-m_{2j})}{(n_{1j}+n_{2j})^2(n_{1j}+n_{2j}-1)} \quad (8),$$

kus $n_1$ ja $n_2$ on vastavalt kõvera 1 ja kõvera 2 isikute arv perioodis *j* ning $m_1$ ja $m_2$ on vastavalt kõvera 1 ja kõvera 2 huvipakkuvate sündmuste toimumiste arv perioodis j (Kleinbaum & Klein, 2005). Käesolevas töös kasutatakse log-rank testi Kaplan-Meieri kõverate võrdlemiseks.

# 2 Ühtne andmemudel

*Observational Medical Outcomes Partnership* (OMOP) *common data model* (CDM) on Ryani ja Hripcsaki poolt kirjeldatud kui standardiseeritud andmemudel, mida kasutatakse rahvusvahelisel tasandil OHDSI nimelise avatud teaduse kogukonna töös (Ryan & Hripcsak, 2021). Bakalaureusetöö praktilises pooles valminud R-pakett loodi tööriistana OMOP CDM kujul andmemudelile ja seda kasutavale kogukonnale.

OMOP ühtse andmemudeli eesmärk on Blacketeri sõnul struktureerida terviseandmed erinevatest andmebaasidest nii, et need oleks rahvusvaheliselt üheselt mõistetavad ja kasutatavad (Blacketer, 2021). Üheselt mõistetaval andmemudelil teostatud uuringud on rahvusvahelisel tasandil reprodutseeritavad. Ryan ja Hripcsak on nimetanud uuringute reprodutseeritavuse, läbipaistvuse ja tõenduspõhisuse ühtlasi OHDSI kogukonna põhieesmärkideks (Ryan & Hripcsak, 2021). Blacketer on lisanud, et andmed on üldandmemudelit kasutades paremini kaitstud, sest laialdaste uuringute läbiviimiseks piisab uuringu sisu, mis koosneb uuringu protokollist, metoodikast ja taasloomiseks vajalikust koodist, jagamisest (Blacketer, 2021).

Tartu Ülikooli terviseinformaatika töögrupp on loonud töövoo, mille väljundiks on Eestis kogutud terviseandmete teisendamine OMOP CDM kujul andmemudelile (Oja et al. 2023). Järgnevalt tutvustatakse OHDSI kogukonda, bakalaureusetöös kasutatud OHDSI hallatud fenotüübi sõnastikku ja vabavaralist tarkvara OMOP andmemudelil töötamiseks.

## 2.1 OHDSI

Ryan ja Hripcsak on kirjutanud OHDSI avatud teatmikus kogukonda tutvustava alalõigu, mida järgnevalt kokkuvõtvalt kirjeldatakse. OHDSI on rahvusvaheline terviseteadusele keskendunud kogukond, mille eesmärgiks on inimeste tervise edendamine ja ravivõtete parendamine andmeteaduse ja informaatika toel. Autorid on lisanud, et OHDSI uurib terviseandmete kasutusvõimalusi ja kasutusmeetmete olulisust. Kogukond töötab avatud teaduse põhimõttel, millest tulenevalt on loodud analüüsimistarkvarad, kasutatud metoodikad ning teaduslikud leiud avalikud (Ryan & Hripcsak, 2021). Alates 2023. aastast on S. Reisbergi juhtimisel Eestis iseseisev OHDSI haru (OHDSI Europe).

## 2.2 OMOP CDM kujule viimine Eesti näitel

Tartu Ülikooli arvutiteaduse instituudi töögrupp avaldas 2023. aastal metoodika Eesti residentide terviseandmete viimise üle OMOP CDM kujule. Metoodika koondab digiretseptid, digilood ja tervisekindlustuse nõuded üheks andmebaasiks. Tuleb märkida, et surmaregister ei ole kaasatud, seega surmapõhjused ei ole ühtses andmemudelis saadaval. Töö osaks oli Eesti elanike terviseandmetel põhineva OMOP CDM loomine. Varasemalt olemasolev MAITT andmestik, 10% juhuvalim (n=150824) Eesti elanikkonnast aastatest 2012 kuni 2019, viidi üle OMOP CDM kujule, kasutades arendatud metoodikat ja demonstreerides valmisolekut viimaks andmed täielikult ühtsele andmekujule (Oja et al. 2023). Lisaks on koostööd alustatud Eesti geenivaramuga, CORIVA projektiga ja Tartu Ülikooli Kliinikumi onkoloogia keskusega (OHDSI Europe).

## 2.3 Fenotüübi sõnastik

Käesoleva bakalaureusetöö praktiline osa kasutab andmebaasipäringute loomeks OHDSI hallatud kohortide definitsioonide kogumikku nimega *Phenotype Library*. Kohort on andmeanalüüsi ja statistika oskussõnastikus[1] kirjeldatud kui inimrühm, kes kogeb ühist elusündmust. Selle töö raames on moodustavad kohordi inimesed, kellel on registreeritud mingi ühene tervisega seotud sündmus. Rao on kirjutanud, et kogumiku *Phenotype Library* eesmärgiks on parendada vaatlusandmete kasutatavust FAIR[2] põhimõtete suunal. Selle eesmärgi täitmiseks on kogumik keskendunud kohordi definitsioonide hoiustamisele, nende asjakohastena hoidmisele ja metaandmete kogumisele. Autor on lisanud, et kogumiku korrashoiuks on tehnilised testid kuid ka OHDSI kogukonna ühine panus pisteliseks kontrolliks on oluline (Rao, 2024). Kohordi definitsioonide kogumik hõlbustab OMOP andmemudelil uuringute teostamist pakkudes eeldefineeritud andmebaasipäringuid, mille edasine täpsustamine on võimalik tööriista ATLAS[3] abil. Kirjeldatud sõnastik hõlbustab OMOP andmemudeliga kokkupuutuvate inimeste tööd.

## 2.4 R-pakett *CohortSurvival*

OHDSI kogukonna eesmärkide täitmiseks on loodud andmete töötlemiseks kui ka analüüsimiseks OMOP CDM-iga ühilduvaid tööriistu. Siinkohal tutvustatakse neist ühte, mis

[1] https://sonaveeb.ee/search/unif/dlall/dsall/kohort/1

[2] https://www.go-fair.org/fair-principles/

[3] https://www.ohdsi.org/software-tools/

avaldati käesolevast lõputööst sõltumatult 2023. aastal DARWIN EU[4] projekti raames ja teostab andmetel elukestusanalüüsi. *CohortSurvival* on R-pakett, mis visualiseerib OMOP CDM andmemudelikujuga andmetel elukestusanalüüsi olemasolevate andmete põhjal (López-Güell et al. 2023). Paketi sisu kattub märkimisväärselt käesoleva bakalaureusetöö praktilise väärtusega, kuid lõputöö kirjutamise hetkel esineb neis erinevusi. Erinevalt mainitud olemasolevast paketist *CohortSurvival* on selle töö raames valminud R-pakett suuteline lisaks elukestusanalüüsile koostama ja teostama andmebaasipäringuid ning analüüside visualiseerimiseks on loodud kasutajaliides. Lisaks kasutab käesolev lõputöö OHDSI fenotüübi sõnastikku. *CohortSurvival* olemasolu kinnitab, et huvi elukestusanalüüsi töövahendite vastu on olemas ning bakalaureusetöö raames loodud paketil on potentsiaal kanda suurt praktilist väärtust.

[4] https://www.darwin-eu.org/

# 3 R-pakett: *cohortSurvivalAnalysis*

Selles peatükis kirjeldatakse loodud R-paketi kasutusvõimalusi ja ülesehitust. Paketi eesmärk on olla töövahend elukestusanalüüsi teostamiseks OMOP CDM kujul andmebaasidel. Loodud R-pakett kombineerib andmebaasipäringute loomise, päringute teostamise ja päringute vastuste visualiseerimise kasutajaliideses.

Pakett loodi eesmärgiga sobituda rahvusvahelise avatud teaduse kogukonna töövahendite ja andmemudeliga. Sellest tulenevalt kannab R-pakett võõrkeelset nime ning võimalikust sihtgrupist tulenevalt on nii dokumentatsioon kui ka kasutajaliides inglisekeelsed.

## 3.1 Sisu

R-pakett *cohortSurvivalAnalysis*[5] kasutamise eelduseks on ligipääs OMOP CDM kujul andmebaasile. Selles peatükis kirjeldatakse paketi kirjutamisel kasutatud vahendeid ja loodud paketi sisu nii funktsionaalsust kui ka kasutajaliidest.

Pakett arendati vabavaralises programmeerimiskeeles R versioonis 4.2.2. Keel osutus valituks, sest sellele on varasemalt loodud pakette nii OHDSI võrgustiku siseselt kui ka andmetöötluseks ja -visualiseerimiseks ning elukestusanalüüsi teostamiseks. Loodud R-pakett kasutab mitmeid pakette, millest märkimisväärseima panuse annab *Cohort2Trajectory*[6], mida kasutati andmebaasipäringute teostamiseks. Mitmed kasutatud paketid kuuluvad *tidyverse*[7] kogumikku, milles olevad vahendid on loodud andmeteaduse tarbeks (Tidyverse). Andmeid korrastati ja viidi sobivale kujule *dplyr*[8], *tidyr*[9] ja *survival*[10] pakettidega. *Dplyr* ja *tidyr* kuuluvad *tidyverse* kogumikku. *Survival* on aktiivselt hooldatud elukestusanalüüsi matemaatilise analüüsi teostamise abipakett (Therneau et al. 2024). Töös kasutati andmete visualiseerimiseks *ggplot2*[11] ja *survminer* pakette. *Survminer* töötab *ggplot2*, mis on osa andmeteaduse standard pakettide kogumikust, toel ning on loodud elukestuskõverate joonistamiseks andmetel, mille eeltöö on tehtud näiteks *survival* paketi abiga (Kassambara

[5] https://github.com/GreeteKelli/cohortSurvivalAnalysis

[6] https://github.com/HealthInformaticsUT/Cohort2Trajectory

[7] https://www.tidyverse.org/

[8] https://dplyr.tidyverse.org/

[9] https://tidyr.tidyverse.org/

[10] surhttps://cran.r-project.org/package=survivalvival

[11] ggphttps://ggplot2.tidyverse.org/lot2

et al. 2022). Kasutajaliides loodi *Shiny*[12] ja *shinydashboard*[13] toel, mis on loodud mugavalt ühtse disainiga kasutajaliidese loomiseks (Chang et al. 2024; Chang & Ribeiro 2022). Andmebaasipäringute kirjutamisel kasutati *readr*[14], *stringr*[15] ja *rjson*[16] võimalusi, millest esimesed kaks kuuluvad samuti *tidyverse* kogumikku.

R-pakett *cohortSurvivalAnalysis* funktsionaalsus on leitav samanimelisest funktsioonist failis *cohortSurvivalAnalaysis.R*. Mainitud funktsioon sisaldab R-paketi automatiseeritud töövoogu, mille kirjeldus on leitav alapeatükis 3.2 Töövoog. Bakalaureusetöö teoreetilises pooles mainiti elukestusanalüüsi meetodite kasutamist erinevate uuringute jaoks, kus uurimisobjektiks on aeg mingi sündmuseni. Sellest tulenevalt on võimalik huvipakkuvat sündmust määrata töövoogu alustavas väljakutses. Vaikimisi on huvipakkuvaks sündmuse väärtuseks seatud elulõppu tähistav sündmus, mis tuleneb praktilise töö esmasest eesmärgist luua töövahend, millega ennustada tervete inimeste elukestust esmase haigestumise järel. Kirjeldatud funktsioon *cohortSurvivalAnalysis* kasutab vaikimisi *readPhenotypeLibraryToJson* funktsiooni, mille eesmärgiks on lugeda avakoodihoidlast failid ja neid asjakohaselt muuta. Paketi töövoog on automatiseeritud ning mainitud funktsioone on võimalik kasutada ka iseseisvalt.

Elukestusanalüüsi graafikud ja muud andmeanalüüsi tulemused kuvatakse kasutajaliideses. Järgnevalt tutvustatakse põhjalikumalt kasutajaliidese komponente. Loodud liides koosneb peidetavast külgribast ja neljast vahelehest:

- elukestusanalüüs (ingl *survival analysis*),
- andmete ülevaade (ingl *data analytics*),
- andmete võrdlus (ingl *comparison*),
- abi (ingl *help*).

Külgriba võimaldab kasutajal valida andmebaasipäringute tulemusena saadud kohortide vahel. Kohordid on eristatavad nimeliselt, milleks saavad olla haigusnimetused või muud registreeritavad tervisega seotud sündmused. Lisaks sellele on võimalik määrata vanusevahemik, mida analüüsis kuvada soovitakse ning graafikutel võrdluseks eristada sugu.

[12] https://www.rdocumentation.org/packages/shiny/versions/1.8.1.1

[13] shinydashttps://cran.r-project.org/web/packages/shinydashboard/index.htmlhboard

[14] https://readr.tidyverse.org/

[15] https://stringr.tidyverse.org/

[16] https://cran.r-project.org/web/packages/rjson/index.html

Sooliselt eristatakse kõiki sugusid, mis päringutulemuses olemas on. Külgribal on võimalik märkida võrreldavad andmebaasid, sellisel juhul on kolmandal vahelehel, nimega andmete võrdlus (ingl *Comparison*) kuvatud kahe andmebaasi päringutulemused kõrvuti.

Elukestusanalüüsi vahelehel kuvatakse Kaplan-Meieri kõverat ja kumulatiivset riskifunktsiooni. Graafikud joonistatakse vaikimisi kõikide vanusegruppide üleselt ning sooliselt andmeid ei eristata (vt lisa I).

Andmete ülevaate vahelehel kuvatakse soolist jaotumist sektordiagrammiga, kohordi vanuselist jaotust ja üldinfot. Vahelehe andmed ei ole sooliselt eristatavad aga andmeid, mida kirjeldatakse sektordiagrammis ja üldinfos on võimalik vanuseliselt filtreerida (vt lisa II).

Andmete võrdluse vaheleht pakub võimalust võrrelda elukestusanalüüse sama algussündmusega, kuid erinevate lõppsündmuste või teiste andmebaaside vahel. Võrdluseks kasutatavat andmestikku on võimalik valida kaustale, kus andmestik paikneb, viidates vasakul oleval külgribal (vt lisa III & IV).

Abi vahelehel on kirjeldatud kasutajaliidese kasutamist ja otstarvet. Vahelehe eesmärk on aidata lahendada lihtsamaid probleeme ja kirjeldada kasutajale võimalusi (vt lisa V).

## 3.2 Töövoog

Loodud R-paketi töövoog jaguneb kaheks, andmebaasipäringute loomine ja andmebaasi suhtlus. Töövoo alustamiseks tuleb seadistada andmebaasikasutaja ja sessiooni informatsioon ning vajalikud failiteed. Voo käivitamiseks tuleb teha väljakutse *cohortSurvivalAnalysis* funktsioonile.

### 3.2.1 Andmebaasipäringute loome

Andmebaasipäringute loomine toimub R-paketis *readPhenotypeLibraryToJson* nimelise funktsiooniga. Funktsiooniga samas R-failis, *readPhenotypeLibraryToJson*, on mainitud funktsiooni abifunktsioonid *modifyJSON* ja *readInclusionRules*. Päringute algformaadid loetakse sisse fenotüübi sõnastikust *Phenotype Library* kaustast avakoodihoidla Github veebilehelt ning soovitud muudatused teostatakse lokaalselt. Muudatusteks on vaatlusperioodi alustava sündmuse täpsustamiseks kaasamisreeglite lisamine (ingl *inclusion rules*), esmase kriteeriumi seadmine, mis tagab ainult tervisesündmuste esmased juhud, ja vormingu kohendustest, mis on vajalikud päringuna töötamiseks. Loodud päringud salvestatakse kasutaja lokaalsesse kausta. Funktsiooni *readPhenotypeLibrary* on võimalik

kasutada ka osaliselt, sisestades soovitud kohortide identiteedi koodid vastavasse sisendparameetrisse, ilma ülejäänud töövoogu rakendamata või jätta funktsioon töövoos kasutamata. Lisaks on sisendparameetrina võimalik määrata elukestusanalüüsi alustavale sündmusele eelneva perioodi, kus puuduvad registreeritud terviseandmed, pikkust. Kirjeldatud sisendparameeter võimaldab uuringu tingimustes, kus vaatlusperioodi alustavaks sündmuseks on terve inimese haigestumine, määrata, kui palju aega peab viimasest registreeritud terviseandmete sündmusest möödunud olema.

### 3.2.2 Andmebaasi suhtlus

Andmebaasi päringute tegemine toimub R-paketi *Cohort2Trajectory* vahendusel, mis on Tartu Ülikoolis Haugi magistritöö raames valminud pakett patsientide ravitrajektooride loomiseks meditsiiniandmete põhjal. Haug on kirjutanud loodud paketi kohta, et trajektooride loomisel valitav sihtkohort (ingl *target cohort*) eraldab vastavalt defineeritud kohordile inimrühma koos nende vaatlusaegadega ning neelduv seisund lõpetab vaatlusperioodid, kui need peaks esinema (Haug, 2022). Käesolevas bakalaureusetöös kasutatakse sellest osa, mis tagastab sisendina saadud JSON-failile patsientide ravitrajektoorid defineeritud seisunditega. Sihtkohortiks valitakse töövoole ette antud kohortide definitsioonid ning seisundi kohordiks või neelduvaks seisundiks vaikimisi surm. Elukestusanalüüsi tegemiseks vaadeldakse igat patsienti, kes esinesid sihtkohordis kuni nad sattusid neelduvasse seisundisse või nende vaatlusperiood lõppes. Andmebaasi suhtlus on jagatud kaheks etapiks, kus esimeses määratakse soovitud sihtkohordid ja neelduv seisund ning teises osas tehakse andmebaasipäringud soovitud trajektooride koostamiseks.

## 3.3 Edasiarendusvõimalused

Bakalaureusetöö praktilise osa koostamise käigus ilmnesid mitmed nõrkused, mille parandamine tõstaks paketi olulisust ja kasutusmugavust. Käesolevas töös ilmnesid puudujäägid töö hilisemas etapis.

Kasutajaliidese ja andmeanalüüsi edendamiseks soovitab käesoleva töö autor lisada edasise analüüsi valmidus uurimaks neid, kellega huvipakkuv sündmus toimus. Lisaks kuvatavatele väärtustele on mõistlik lisada andmete filtreerimisvõimalusi kasutajaliideses. Kirjeldatud lisandväärtus võimaldaks saada ülevaadet tervisega seotud sündmuste järgnevustest inimrühmade seas.

Kasutajaliidese ülesehitus võiks olla intuitiivsem ning teavitada kasutajat võimalikest vigadest ja viivistest, näiteks kui analüüsi ootamiseks kulub märkimisväärselt rohkem aega. Käesoleva töö autor lisaks kasutajaliidesele raporti koostamise võimaluse, mis koosneks kasutaja valitud joonistest ja informatsioonist.

Tarkvara terviklikuks edendamiseks soovitab autor paketi testimist ja töökeerukuse optimeerimist. Funktsionaalsuse testimine tõstaks paketi töökindlust ning testimise käigus võivad ilmneda võimalused, kus kasutajakogemust on võimalik parendada. Keerukuse optimeerimine on oluline samm kindlustamaks paketi laialdast kasutamist.

# 4 Näidisuuring

Käesolevas peatükis esitletakse varasemalt kirjeldatud R-paketti *CohortSurvivalAnalysis*, selle tarbeks viiakse läbi näidisuuring. Näidisuuringu tarbeks on valitud sihtkohordiks, ehk vaatlusperioodi alustavaks sündmuseks, süvaveeni tromboos (ingl *Deep Vein Thrombosis*) ja neelduvaks seisundiks, ehk vaatlusperioodi huvipakkuv sündmus, kõrgvererõhutõve diagnoos (ingl *Essential Hypertension*). Fenotüübi sõnastikus on sündmused koodidega vastavalt 1152 ja 770. Kõrgvererõhutõbi osutus valituks alustavaks sündmuseks, sest see on CORIVA andmebaasis COVID-19 järel enim esinev tervisesündmus ning sihtkohordi valikuni jõuti juhuslikult.

## 4.1 Andmed

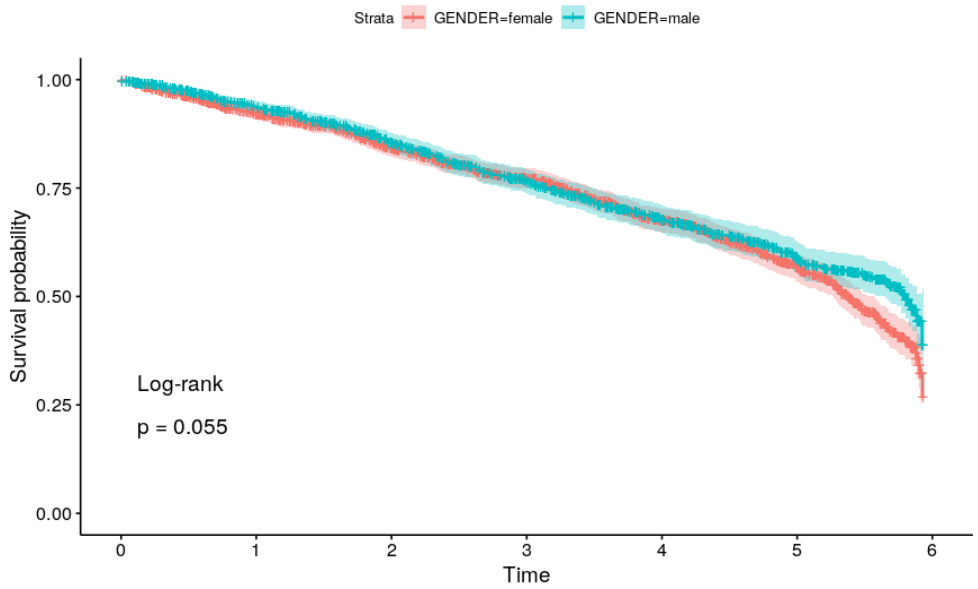
Näidisuuring teostati kahel OMOP CDM kujul andmebaasil, CORIVA ja RITA-MAITT. RITA-MAITT on varasemalt, peatükis OMOP CDM kujule viimine Eesti näitel, tutvustatud andmebaas. CORIVA andmebaas loodi samanimelise projekti raames, mille eesmärk oli SARS-CoV-2 viirusinfektsiooni süsteemne analüüs, keskendudes nii viiruse levikule kui ka COVID-19 põdemise tagajärgedele. Andmebaas koosneb ligikaudu 438 000 isiku andmetest (Uusküla et al., 2022).

Andmestiku kuju, mida visualiseerimisel kasutatakse on väljund *Cohort2Trajectory* paketi tööst, mis loob vastavalt päringule andmetabeli (vt lisa VI). Patsiendid on eristatavad andmetabelis *SUBJECT_ID* toel, mis on tabelis omistatud identiteedi kood. Esitletud näidises on iga patsiendi kohta kolm või neli andmerida, olenevalt huvipakkuva sündmuse mittetoimumisest või toimumisest. Ühte patsienti kirjeldavad read on eristatavad veeru *STATE_LABEL*, ehk seisundi nimetus, väärtustega, milles *targetCohort* tähistab tervise sündmuse avaldumist, *absorbingState* neelduva sündmuse toimumist, *start* tähistab vaatlusandmete perioodi algust ning *end* tähistab vaatlusandmete perioodi lõppu. Eristatavaid seisundeid iseloomustavad *STATE_START_DATE* ehk seisundi algus kuupäev, *STATE_END_DATE* ehk seisundi lõppkuupäev, *STATE_ID* ehk seisundi identiteedi kood ja *SEQ_ORDINAL*, mis näitab mitmendat korda seisund antud patsiendiga toimub. Lisaks on igal andmereal välja toodud *GENDER_CONCEPT_ID* ehk patsiendi sootunnus ja *AGE* ehk vanus.

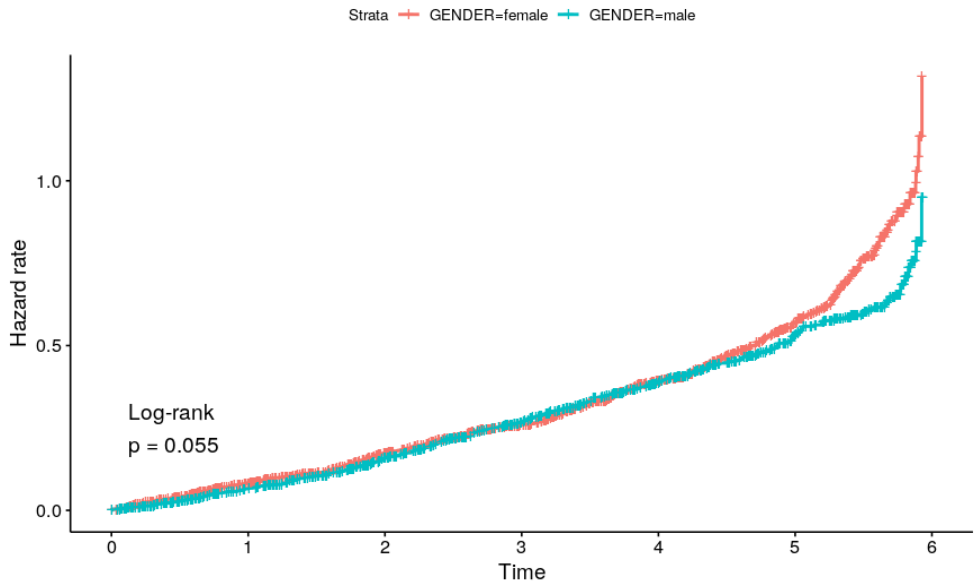
## 4.2 Tulemused

Selles peatükis tutvustatakse bakalaureusetöö raames loodud paketi *cohortVisualAnalysis* näidisuuringu tulemusi, nii graafilise liidese väljundit kui ka töövoo kiirust. Näidisuuringu eesmärk on esitleda paketi võimekust, seega statistilisi järeldusi järgnevas peatükis ei tehta.

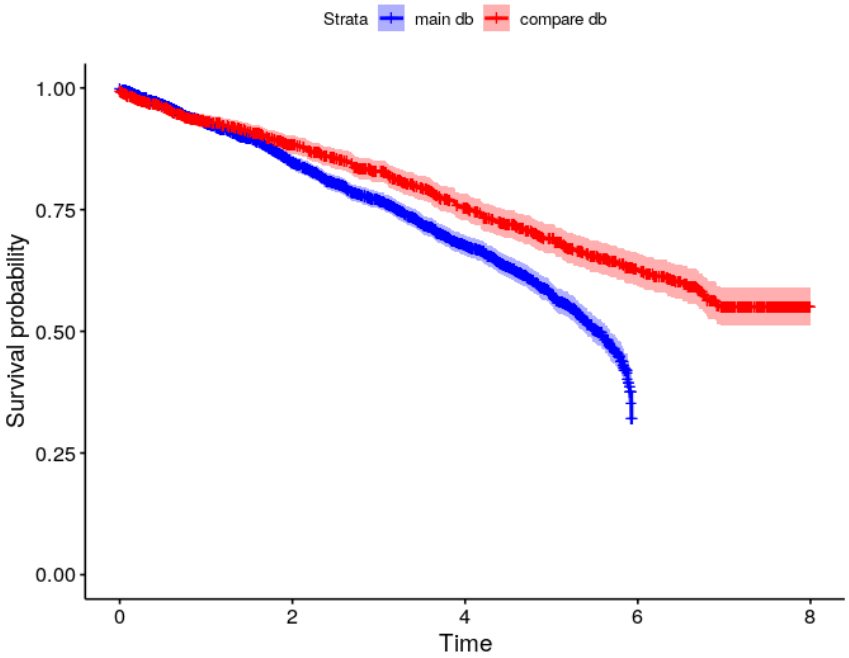
Kasutajaliidese vahelehtede tulemusi näidisuuringu tingimustel on kujutatud joonistel 3, 4, 5.




Joonis 3. Kaplan-Meieri kõver näidisuuringu tingimustel vanusegruppide üleselt ja sooliselt eristatud.




Joonis 4. Kumulatiivne riskifunktsioon näidisuuringu tingimustel vanusegruppide üleselt ja sooliselt eristatult.




Joonis 5. Kaplan-Meieri kõverad näidisuuringu tingimustel vanusegruppide üleselt ja sooliselt eristamata andmebaasidel RITA-MAITT ja CORIVA.

Joonistel on legendis väljatoodud andmestike kihitamine[17] (ingl *strata*), mis tähistab joonisel eristatud andmekihte. Joonisel 3 on kujutatud kaks Kaplan-Meieri kõverat, kus on ilmne, et ajahetkede viis ja kuus vahel on elulemuses selge erinevus, siiski arvutatud p-väärtus log-rank testmeetodiga statistilist erinevust ei tuvasta. Joonisel 3 on eristatavateks kihtideks sootunnus. Sarnaselt eelnevalt kirjeldatud joonisele on joonisel 4 kumulatiivse riskifunktsiooni graafikul alates neljanda ja viienda ajahetke vahelt kuni vaatlusaja lõpuni näha erineva intensiivsusega riskifunktsioone, kuid log-rank meetodiga arvutatud p-väärtus erinevust ei kinnita. Joonisel 5 kuvatud kaks Kaplan-Meieri kõverat, mis põhinevad kahel erineval andmebaasil, on visuaalselt eristatava elulemusega. Võrdluseks valitud andmestik, kaheksa eristatava ajahetkega, oli pikema vaatlusperioodiga kui esmaselt valitud andmestik, kuue eristatava ajahetkega (vt joonis 5). Lisaks on tuvastatav joonisel 5 tugev platoo vaatlusperioodi lõpus võrdluseks valitud andmestikul. Arvestades graafiku eelnevat monotoonset langust on platoo loomulik teke ebatõenäoline ning see võib viidata puudulikele andmetele või töövoo veale.

Järgnevalt kirjeldatakse paketi kasutusele kuluvat aega. Töövoo mõõtmisel eelistati ajahetkede vahe leidmist, kus algus ja lõpphetk on määratud R programmeerimiskeele

[17] https://sonaveeb.ee/search/unif/dlall/aso/kihitamine/1

funktsiooniga *Sys.time*[18]. Töövoo esmase etapi, ehk andmebaaside päringute loomine, läbimiseks kulus kogu fenotüübisõnastiku sisu lugemisel ligikaudu 6.5 minutit. Andmebaasipäringute etappi eraldiseisvalt ei hinnatud, kuna etapi põhiosa moodustab kasutatud pakett *Cohort2Trajectory*. Kasutajaliidese võimekuse hindamiseks mõõdeti graafikute tekkimise kiirust juhuslikult valitud andmestikega. Mõõtmises kasutati elukestusanalüüsi vahelehe Kaplan-Meieri kõvera graafikut, andmete võrdluse vahelehe graafikut kahe Kaplan-Meieri kõveraga ja andmete võrdluse vahelehe kõigi nelja graafikute teket (vt tabel 1).

| | Aritmeetiline keskmine (s) | Mediaan väärtus (s) | Standardhälve (s) |
|---|---|---|---|
| elukestusanalüüsi vahelehe Kaplan-Meieri kõvera graafiku kuvamiseks kulunud aeg | 1.04 | 0.35 | 1.52 |
| andmete võrdluse vahelehe Kaplan-Meieri kõverate graafiku kuvamiseks kulunud aeg | 1.66 | 0.73 | 2.05 |
| andmete võrdluse vahelehe täielikuks kuvamiseks kulunud aeg | 3.72 | 1.66 | 4.63 |

Tabel 1. Kasutajaliidese graafikute kuvamiseks kulunud aeg sekundites.

Rakenduse esmasele käivitamisele kulunud aega mõõtmistel ei arvestatud. Mõõtmised viidi läbi Tartu Ülikooli arvutivõrgu töökeskkonnas, mis eraldati autorile töö teostamiseks. Sellest tulenevalt võib kasutamiskiirus erineda tabelil 1 väljatoodust.

[18] https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/Sys.time

# Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli luua R-pakett elukestusanalüüsi teostamiseks OMOP CDM kujul andmebaasidel. Pakett loob ja teostab andmebaasi päringud. Kasutajaliideses visualiseeritakse asjakohaseid graafikuid, mis andmeid kirjeldavaid. Tehtud töö kasutab varasemalt loodud OHDSI tööriistu.

Elukestusanalüüs on metoodika uurimaks aega mingil kindlal ajaperioodil mitmel uurimisobjektil. Analüüsi keskseteks meetoditeks on elukestusfunktsioon, mis kirjeldab tõenäosust, et uurimisobjektiga ei toimu kokkuleppelist sündmust, ja riskifunktsioon, kirjeldab näitu, et mingil hetkel vaatlusperioodi jooksul huvipakkuv sündmus toimub. Elukestusanalüüsi levinud võtteks on Kaplan-Meieri kõver, astmeline mittekasvav joonis visualiseerib elukestusfunktsiooni.

Tervishoiu vaatlusandmete digitaalne hoiustamine annab võimaluse kasutada olemasolevaid andmeid uuringute läbiviimiseks. Uuringute reprodutseerimiseks peavad olema taasloodavad uuringu keskkonna tingimused. Tervishoiu andmete tundlikkusest tulenevalt ei ole jätkusuutlik nende jagamine. OMOP CDM on ühtne andmemudel, mille abil on võimalik uuringute metoodikaid ja loodud tööriistu taaskasutada erinevatel OMOP CDM andmestikel. Ühtset andmemudelit ja sellele loodud tööriistu ning metoodikaid haldab OHDSI nimeline avatud teaduse põhimõttel töötav kogukond.

Bakalaureusetöö raames valmis R-pakett *cohortSurvivalAnalysis*, millega saab teostada elukestusuuringut OMOP CDM andmemudelitel. Pakett sisaldab andmebaasi päringute loomist ja teostamist, etappe saab kasutada eraldi kui ka ühe töövoona. Elukestusanalüüsi graafikute ja teiste statistiliselt oluliste näitajate kuvamine toimub kasutajaliidesega. Kasutajaliideses on võimalik filtreerida andmeid vanuse järgi ning kõrvutada andmeid märgitud soo põhjal. Loodud paketi sobivust OMOP CDM andmemudelil katsetati Tartu Ülikooli teadusgrupi poolt loodud CORIVA ja RITA-MAITT andmestikega. Töö edasiarendusteks pakkus autor välja andmeanalüüsi täpsustamise, paketi põhjalik testimise ning kasutajaliidese kohandamise.

# Kasutatud kirjandus


Ryan, P. Hripcsak, G. (2021). The OHDSI Community. The Book of OHDSI. 2021-01-11. https://ohdsi.github.io/TheBookOfOhdsi/OhdsiCommunity.html (10.05.2024)

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M.A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., van der Lei, J., Pratt, N., Norén, G. N., Li, Y.-C., Stang, P. E., Madigan, D., & Ryan, P. B. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in health technology and informatics*, vol *216*, 574–578.

Flynn, R. (2012). Survival analysis. Journal of Clinical Nursing, vol 21, 2789–2797.

K. López-Güell, D. Newby, I. Koblbauer, X. Li, B. Raventós, M. Català, M. de Ridder, T. Duarte Salles, D. Prieto-Alhambra, & E. Burn. (2023) *CohortSurvival: An R package for survival analysis using the OMOP CDM*. https://www.ohdsi.org/wp-content/uploads/2023/10/Lopez-Kim_CohortSurvival_2023symposium-Kim-Lopez-Guell.pdf

Collett, D., & Kimber, A. (2014). Modelling Survival Data in Medical Research. CRC Press LLC.

Kleinbaum, D. G., & Klein, M. (2005). Survival Analysis: A Self-Learning Text, Second Edition. Springer.

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, vol *53*, 457–481.

Blacketer C. The Common Data Model. *The Book of OHDSI*. 2021-01-11. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html (10.05.2024)

Oja, M., Tamm, S., Mooses, K., Pajusalu, M., Talvik, H.-A., Ott, A., Laht, M., Malk, M., Lõo, M.,Holm, J., Haug, M., Šuvalov, H., Särg, D., Vilo, J., Laur, S., Kolde, R., & Reisberg, S. (2023). Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: Lessons learned. *JAMIA Open*, vol *6*.

OHDSI Europe. https://www.ohdsi-europe.org/index.php/national-nodes/estonia (10.05.2024)

Rao, G. The OHDSI phenotype library. 2024. https://ohdsi.github.io/PhenotypeLibrary/ (10.05.2024)

Tidyverse. https://www.tidyverse.org/ (10.05.2024)

Therneau, T. M. Package 'survival'. April 24, 2024. https://cran.r-project.org/web/packages/survival/survival.pdf (10.05.2024)

Kassambara, A., Kosinski, M., Biecek, P. Package 'survminer'. October 14, 2022. https://cran.r-project.org/web/packages/survminer/survminer.pdf (10.05.2024)

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B. Package 'shiny'. April 2, 2024. https://cran.r-project.org/web/packages/shiny/shiny.pdf (10.05.2024)

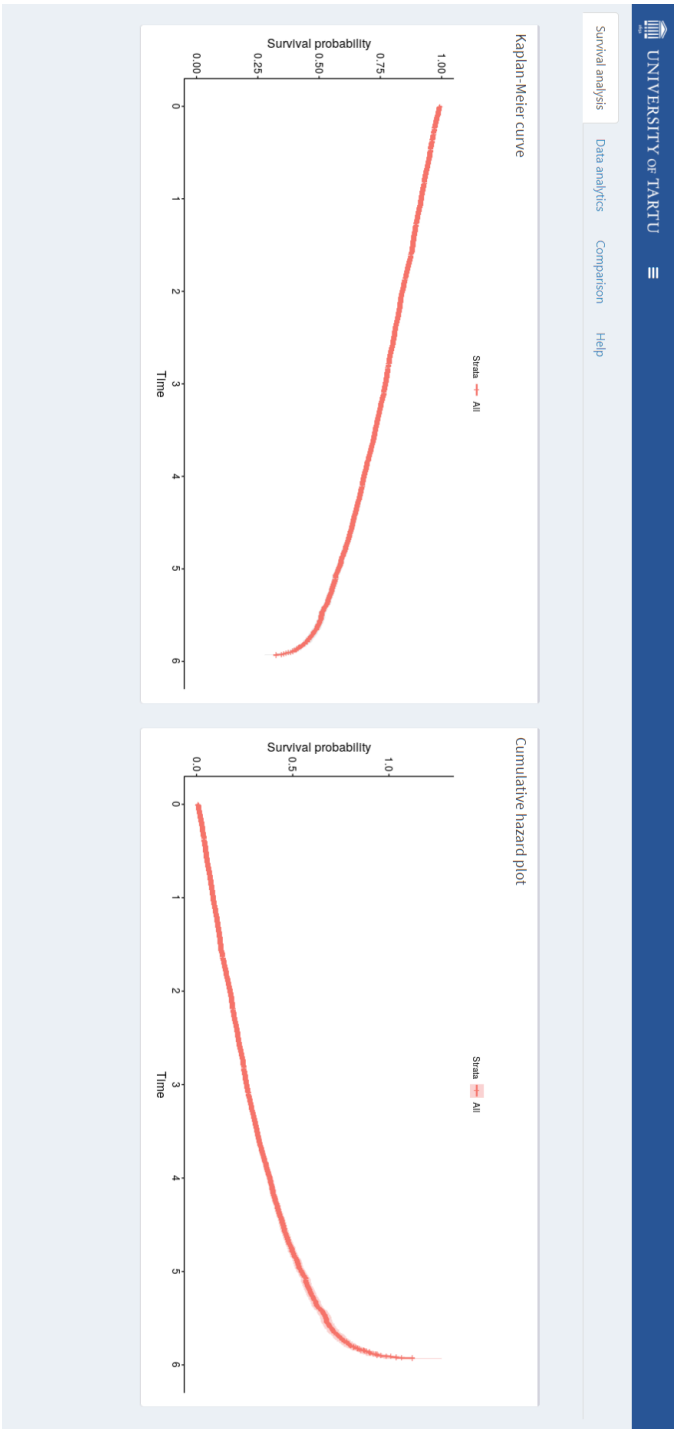
Chang, W., Ribeiro, B. B. Package 'shinydashboard'. October 14, 2022. https://cran.r-project.org/web/packages/shinydashboard/shinydashboard.pdf (10.05.2024)

Haug, Markus. (2022). „Patsientide ravitrajektooride modelleerimine Markovi ahelatega", TÜ arvutiteaduse instituudi magistritöö. https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=74982 (10.05.2024).

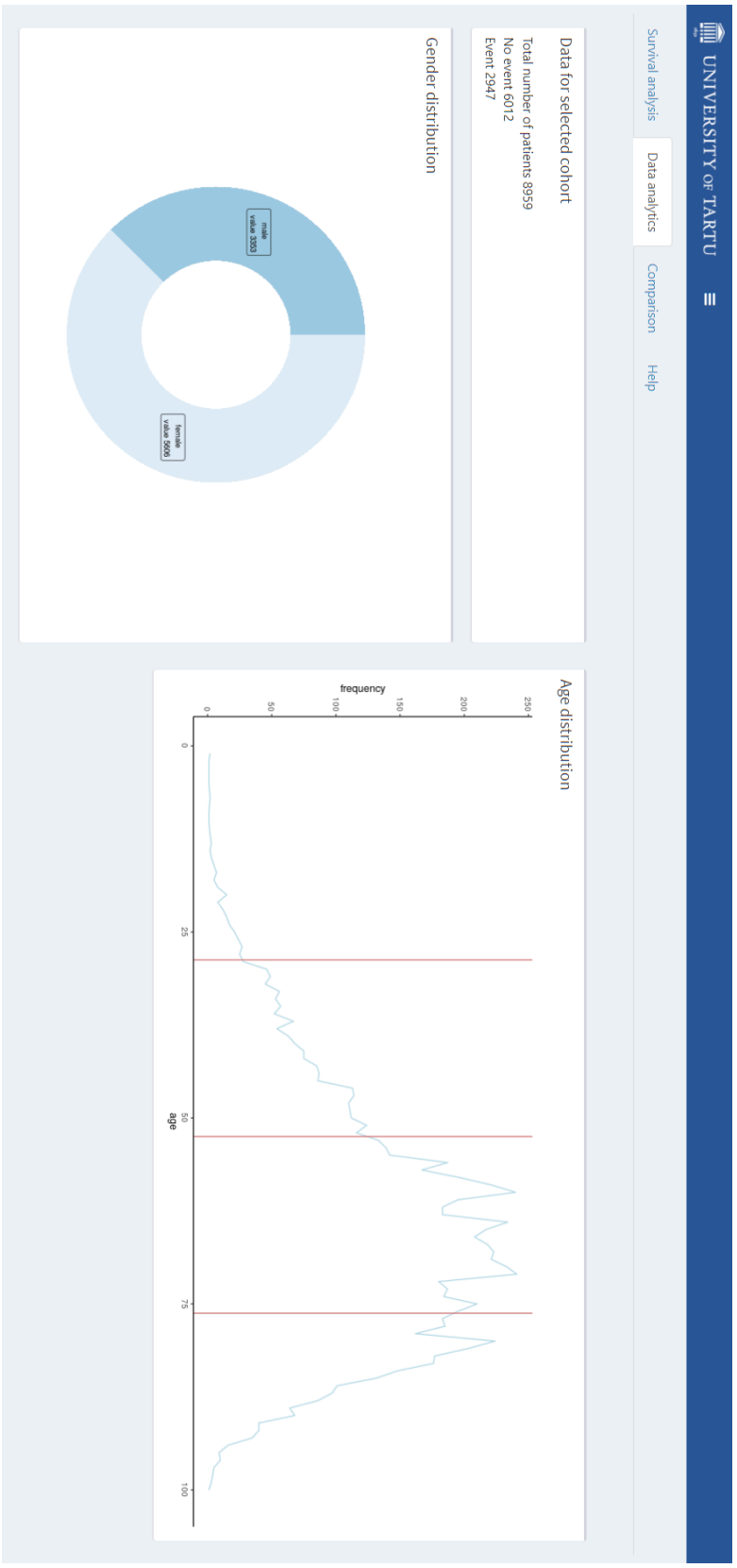
Uusküla, A., Meister, T., Kalda, R., Suija, K., Piirsoo, M., Kolde, R., Milani, L., & Karo-Astover, L. (2022). COVID-19 haigusjuhtumite analüüs ja riskirühmade väljaselgitamine Eestis: Lõpparuanne. https://www.etag.ee/wp-content/uploads/2022/05/RITA1.02-120-lo%CC%83pparuanne-02.05.22-LOPLIK.pdf (13.05.2024)

# Lisad

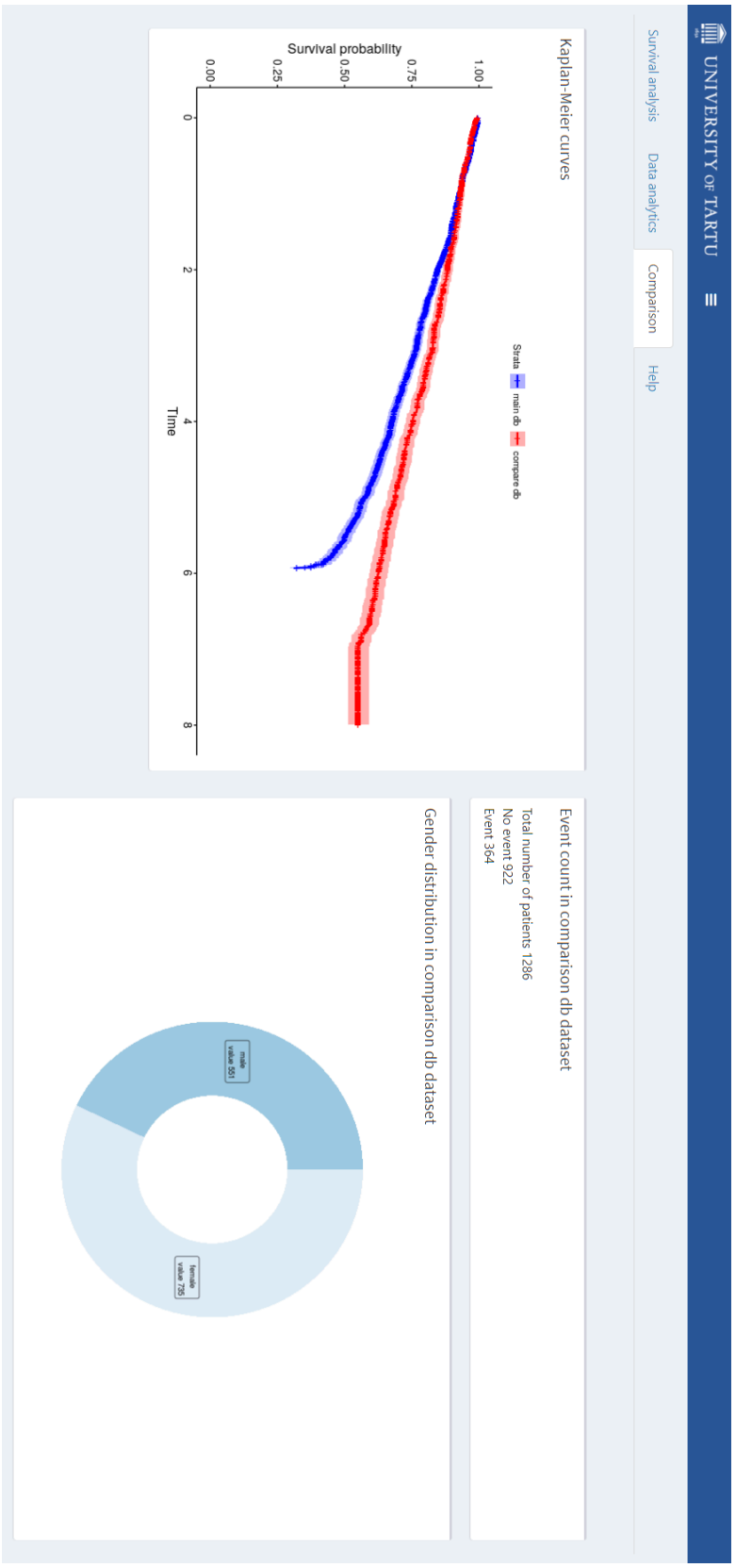
## I. Paketi *cohortSurvivalAnalysis* kasutajaliidese elukestusanalüüsi vaheleht



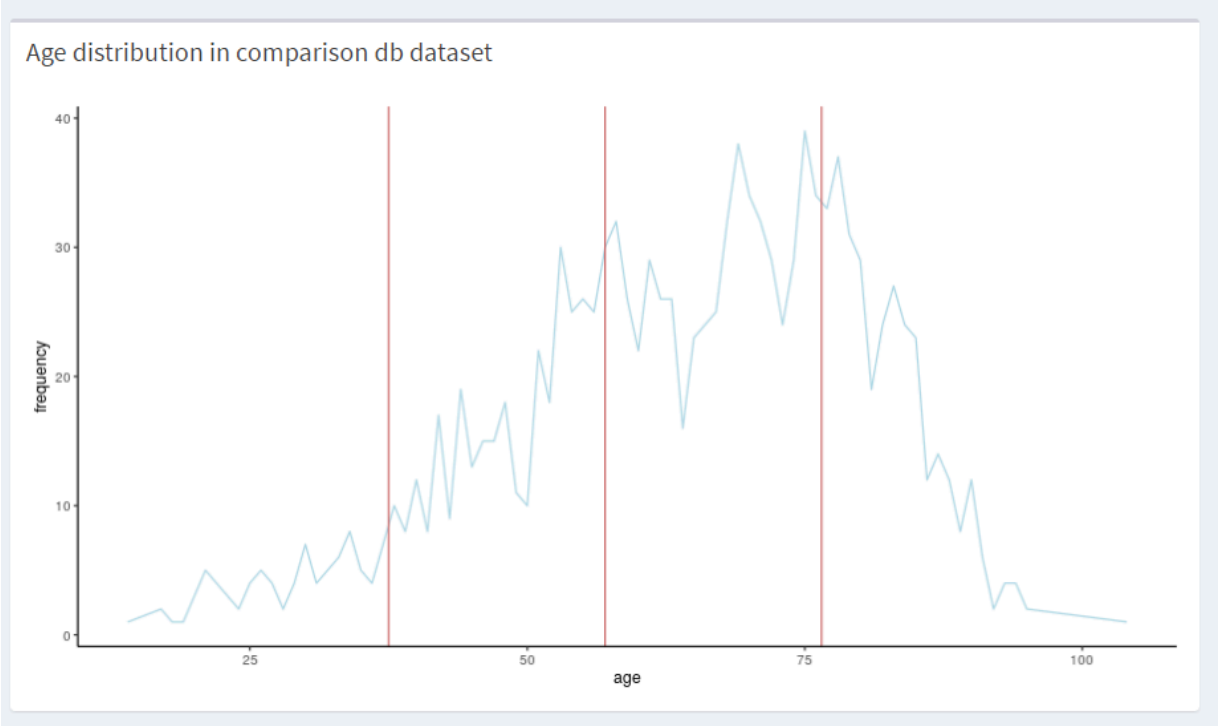
## II. Paketi *cohortSurvivalAnalysis* kasutajaliidese andmeanalüüsi vaheleht



## III. Paketi *cohortSurvivalAnalysis* kasutajaliidese andmete võrdluse vaheleht (1)



## IV. Paketi *cohortSurvivalAnalysis* kasutajaliidese andmete võrdluse vaheleht (2)



## v. Paketi *cohortSurvivalAnalysis* kasutajaliidese külgriba koos abi vahelehega

# VI. Paketi *cohortSurvivalAnalysis* kasutajaliideses kasutatava andmestiku näidis

| SUBJECT_ID | STATE_LABEL | STATE_START_DATE | STATE_END_DATE | STATE_ID | TIME_IN_COHORT | SEQ_ORDINAL | GENDER_CONCEPT_ID | AGE |
|---|---|---|---|---|---|---|---|---|
| 1 | START | 2012-09-17 | 2012-09-17 | 1 | 0 | 1 | 8532 | 86.713 |
| 1 | targetCohort | 2012-09-17 | 2012-09-17 | 2 | 0.002738 | 1 | 8532 | 86.713 |
| 1 | EXIT | 2020-01-01 | 2020-01-01 | 4 | 7.288 | 1 | 8532 | 94.001 |
| 2 | START | 2014-08-13 | 2014-08-13 | 1 | 0 | 1 | 8532 | 59.617 |
| 2 | targetCohort | 2014-08-13 | 2014-08-13 | 2 | 0.002738 | 1 | 8532 | 59.617 |
| 2 | absorbingState | 2019-12-31 | 2019-12-31 | 3 | 5.382 | 1 | 8532 | 64.999 |
| 2 | EXIT | 2020-01-01 | 2020-01-01 | 4 | 5.385 | 1 | 8532 | 65.002 |
| … | … | … | … | … | … | … | … | … |

# VII. Litsents

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Greete Kelli Aava,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „R-pakett OMOP CDM kujul andmete elukestusanalüüsiks", mille juhendaja on Markus Haug, 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 4.0, 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.

Greete Kelli Aava

14.05.2024