TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Sander Soodla

R-pakett diagnooside vaheliste seoste uurimiseks elukestusanalüüsiga

Bakalaureusetöö (9 EAP)

Juhendajad:
Neeme Ilves, MD
Maria Malk, MSc
Raivo Kolde, PhD

Tartu 2025

R-pakett diagnooside vaheliste seoste uurimiseks elukestusanalüüsiga

Lühikokkuvõte:

Bakalaureusetöö eesmärk oli luua R keele tarkvarapakett diagnooside vaheliste seoste uurimiseks elukestusanalüüsiga ehk statistiliste meetoditega, mis analüüsivad aega mingi uuritava sündmuse toimumiseni. Loodud pakett *Exposure2OutcomeSurv* töötab rahvusvahelise ühtse andmemudeli (OMOP CDM) vormingus terviseandmebaasidega. Pakett võimaldab terviseinformaatika teadlastel uurida, kas ühe diagnoosi (ekspositsiooni) esinemine mõjutab mõne teise diagnoosi (tulemi) esinemist, ja seda korraga paljudel erinevatel seisunditel. Selleks võrreldakse aega tulemi tekkimiseni ekspositsiooniga patsientide ning neile soo ja vanuse poolest sarnase, kuid ekspositsioonita, kontrollgrupi vahel. Pakett visualiseerib tulemi tekkimise tõenäosust ajas, võrdleb gruppe statistiliselt ning hindab, kui palju suurem või väiksem on tulemi risk ühes grupis võrreldes teisega. Kasutajasõbralikkuse suurendamiseks sisaldab pakett Shiny graafilist kasutajaliidest. Paketi tööd demonstreeriti näidisuuringuga.

Võtmesõnad: elukestusanalüüs, R-pakett, OMOP CDM, Kaplan-Meieri kõver, Coxi mudel

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

R package for Investigating Effects Between Diagnoses Using Survival Analysis

Abstract:

The aim of this bachelor's thesis was to create an R software package for investigating effects between medical diagnoses using survival analysis, i.e., statistical methods that analyze the time until the occurrence of a specific event. The developed package, *Exposure2OutcomeSurv*, works with health databases in the OMOP Common Data Model (CDM) format. The package enables health informatics researchers to investigate whether the occurrence of one diagnosis (exposure) affects the occurrence of another diagnosis (outcome), simultaneously for many different diagnoses. For this purpose, it compares the time to outcome occurrence between patients with the exposure and a control group, matched by gender and age but without the exposure. The package visualizes the probability of outcome occurrence over time, statistically compares the groups, and estimates the magnitude of the outcome risk in one group compared to the other. To enhance user-friendliness, the package includes a Shiny graphical user interface. The package's functionality was demonstrated with a pilot study.

Keywords: survival analysis, R package, OMOP CDM, Kaplan-Meier curve, Cox model

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

Sissejuhatus

Terviseandmete analüüsis on suureks väljakutseks sügavamate seoste mõistmine erinevate haiguste vahel ning nende mõju hindamine patsiendi edasisele tervisekäigule. Üha laialdasemalt kasutusele võetavad rahvusvaheliselt standardiseeritud terviseandmebaasid pakuvad enneolematuid võimalusi nende seoste uurimiseks [1]. Samas nõuab selliste aja jooksul ilmnevate seoste analüüs spetsiifilisi statistilisi meetodeid ja kasutajasõbralikke tööriistu, mis võimaldaksid terviseinformaatika teadlastel neid uurimusi tõhusalt läbi viia.

Selliste seoste uurimiseks pakub võimekat raamistikku elukestusanalüüs. Elukestusanalüüs (ingl *survival analysis*) on laialdaselt kasutatav statistiliste meetodite kogum, millega uuritakse aega täpselt defineeritud alguspunktist mingi sündmuse toimumiseni [2], [3]. Kuigi nimi tuleneb surma toimumise uurimisest, saab elukestusanalüüsi kasutada paljude erinevate sündmuste vaatlemiseks nii meditsiinis kui ka teistes valdkondades, sealhulgas meditsiiniliste tulemite (ingl *outcome*) esinemise [2].

Bakalaureusetöö eesmärk on luua R programmeerimiskeele pakett diagnooside vaheliste seoste uurimiseks elukestusanalüüsiga. Loodav pakett *Exposure2OutcomeSurv* töötab ühe suurima rahvusvahelise ühtse andmemudeli, OMOP CDM, vormingus terviseandmebaasidega ning võimaldab terviseinformaatika teadlastel uurida, kas esmase diagnoosi (ekspositsiooni) esinemine mõjutab mõne teise diagnoosi (tulemi) hilisemat esinemist. Selleks võrreldakse aega tulemi tekkimiseni ekspositsiooniga patsientide ning neile soo ja vanuse poolest sobitatud sarnase, kuid ekspositsioonita, kontrollgrupi vahel. Loodav pakett erineb olemasolevatest lahendustest, sest sisaldab graafilist kasutajaliidest ning võimaldab mitme elukestusanalüüsi meetodi rakendamist paljudel diagnoosidel, kasutades sisendina OMOP CDM andmeid.

Töö jaguneb neljaks sisupeatükiks. Esimeses peatükis tutvustatakse tausta ning antakse ülevaade varasematest sarnastest töödest. Teises peatükis kirjeldatakse elukestusanalüüsi ja teisi meetodeid, mida loodud paketis uuringute läbiviimisel kasutatakse. Valminud tarkvarapaketi tehnoloogiast, ülesehitusest ja töövoost kirjutatakse kolmandas peatükis. Neljandas peatükis analüüsitakse saavutatud tulemusi näidisuuringus ning tuuakse välja viisid rakenduse edasi arendamiseks. Lisas I on näidatud paketi kasutajaliidese külgriba, mida kasutatakse kasutajalt sisendi saamiseks. Lisades II ja III näidatakse paketi kasutajaliidese vahelehti. Lisas IV on loodud R-paketi ülesehitust ja töövoogu visualiseeriv joonis. Lisades V ja VI on näidisuuringus kasutatud failid uuritavate seisundite sisestamiseks kasutajaliidesesse. Lisas VII on tööga kaasa pandud arhiivifaili sisu kirjeldus.

1. Taustainfo

Selles peatükis tutvustatakse tänapäeva terviseandmete analüüsis olulist OMOP ühtset andmemudelit, sellest välja kasvanud OHDSI koostööprojekti ja Eesti terviseandmete andmebaasi RITA-MAITT. Lisaks antakse ülevaade sarnastest töödest.

1.1 OMOP ühtne andmemudel

Observational Medical Outcomes Partnership (OMOP) oli USA Toidu- ja Ravimiameti (FDA) algatatud avaliku ja erasektori partnerlus, mille eesmärgiks oli välja töötada ravimiohutuse järelevalve süsteem, kasutades erinevaid olemasolevaid vaatlusandmeid [4]. Sellise süsteemi loomisel osutus suureks takistuseks vaatlusandmebaaside erinev struktuur, mistõttu loodi ühiste analüütiliste meetodite kasutamiseks OMOP ühtne andmemudel (ingl *common data model*, CDM) [5]. See võimaldab ühes kohas koostatud andmebaasi päringuid ja andmeanalüüsi tööriistu kasutada ülemaailmsetes uuringutes [6]. Lisaks loodi ühtses andmemudelis kasutamiseks standardne meditsiiniterminite sõnavara (ingl *standardized vocabularies*) [7]. Need OMOP projektis välja töötatud lahendused on edendanud instituutidevahelist ja rahvusvahelist koostööd ning algselt ravimiohutuse uurimisele keskendunud andmemudeli kasutusvaldkonnad on tänu avatud teaduse põhimõtetele märkimisväärselt laienenud [8]. Ka selle bakalaureusetöö raames valminud rakendus kasutab OMOP CDM kujul andmeid. Peale OMOP projekti lõppu loodi teadustöö jätkamiseks ning ühtse andmemudeli ja standardse sõnavara haldamiseks rahvusvaheline koostööprogramm OHDSI [8], [9].

1.2 OHDSI koostööprojekt

Observational Health Data Sciences and Informatics (OHDSI) [9] on 2014. aastal loodud rahvusvaheline koostööprojekt, mis loob teaduskogukonnale võimaluse avastada suuremahulistest terviseandmetest tõendeid, mis parandavad inimeste terviseotsuseid ja ravi [10], [11]. OHDSI tegevust toetavad mitmed tegurid [6]. Esiteks võimaldab rahvusvaheline koostöö kasutada andmeid ja jagada teadmisi laias ulatuses. Sellise koostöö võimaldamisel on olulisel kohal eelnevalt mainitud OMOP ühtne andmemudel. Teiseks on vastavalt avatud teaduse põhimõttele OHDSI tööriistad, meetodid ja tulemused kõigile avalikult kättesaadavad ning rakendatavad. Mitmed OHDSI arendatud avatud lähtekoodiga rakendused on kasutusel ka selles töös. Edasi tutvustatakse OMOP CDM rakendamisel Eestis loodud terviseandmebaasi RITA-MAITT.

1.3 RITA-MAITT andmestik

RITA ehk „Valdkondliku teadus- ja arendustegevuse tugevdamine“ programmi raames toimus 2019. – 2022. aastal „Masinõppe ja AI toega teenused“ (MAITT) projekt, mis uuris masinõppe ja tehisintellekti kasutamise võimalusi Eesti avalikes teenustes [12]. Solvaki jt [12] koostatud lõpparuande põhjal oli kaasatud ka tervise valdkond, mille jaoks loodi erinevaid terviseväljundite prognoosimise mudeleid. Selleks otsustati kasutada retseptikeskuse, Eesti Haigekassa ja Tervise Infosüsteemi andmekogusid. Andmete täieliku potentsiaali ära kasutamiseks ja olemasolevate rahvusvaheliselt parimate OHDSI tööriistade kasutamise võimaldamiseks teisendati andmed autorite sõnul kokku ühte OMOP CDM kujul andmebaasi. Andmestik koosnes 10% Eesti elanike juhuvalimi (n = 150 824) pseudonüümitud tervise-andmetest perioodist 2012–2019. Valimist õnnestus 149 364 inimese ehk 99.0% andmed teisendada CDM kujule RITA-MAITT andmebaasi [13]. Seda andmebaasi kasutatakse käesoleva töö raames ehitatud rakenduse tulemuste analüüsiks ja uuringu teostamiseks.

1.4 Seotud tööd

Leidub ka teisi tarkvarapakette OMOP CDM andmetel elukestusanalüüsiga terviseseisundite vaheliste seoste uurimiseks. Siin tutvustatakse neist kolme: OHDSI *CohortMethod* [14], DARWIN EU[1] algatuse raames loodud *CohortSurvival* [15] ja Aava bakalaureusetöös valminud *cohortSurvivalAnalysis* [16].

CohortMethod on siinse tööga väga sarnase eesmärgiga R-pakett, mis keskendub laiemalt igasugustele kohortuuringutele [14]. Kohort on Rahvatervishoiu sõnastikus[2] kirjeldatud kui teatud sarnase sündmuse või aja kaudu seotud uuritavate rühm. Siin töös keskendutakse ainult elukestusanalüüsiga kohortuuringule ning kasutatakse teistsugust meetodit võrreldavate patsientide sobitamiseks. Erinevalt *CohortMethod*ist, kus graafiline kasutajaliides on ainult tulemuste visualiseerimiseks, võimaldab selles töös loodud pakett kogu uuringu läbiviimiseks kasutada graafilist kasutajaliidest. See aspekt parandab oluliselt kasutajakogemust, sest kasutaja ei pea ise käitama õiges järjekorras kümneid erinevaid funktsioone, ning teeb paketi kättesaadavamaks ka IT valdkonnast kaugematele teadlastele.

CohortSurvival on pakett, mis võimaldab samuti ühtse andmemudeli andmetel elukestust analüüsida [15]. Sarnaselt *CohortMethod* paketiga koosneb selle kasutusvoog ilma graafilise

[1] https://www.darwin-eu.org/
[2] https://sonaveeb.ee/search/unif/dlall/rtrv/kohort/1/est

kasutajaliideseta järjestikuste funktsioonide välja kutsumisest. Lisaks on siin loodud paketiga võimalik võrrelda gruppide tulemi esinemise riski, mida *CohortSurvival* ei võimalda.

Küll aga saab graafilise kasutajaliidesega analüüsi läbi viia Aava *cohortSurvivalAnalysis*[3] paketis [16]. Võrreldes *cohortSurvivalAnalysis* paketiga saab siin töös valminud paketiga uurida erinevusi mingisse haigusesse haigestunud ja neile sobitatud mitte-haigestunud patsientide vahel ning arvutada nende gruppide riskide suhted. Lisaks on *Exposure2OutcomeSurv* pakett ka rohkem avastusliku eesmärgiga, sest võimaldab korraga läbi analüüsida suure hulga erinevaid terviseseisundite kombinatsioone.

On olemas ka teisi R keele ja OHDSI kogukonna väliseid graafilise kasutajaliidesega statistikatööriistu (nt GraphPad Prism[4] ja IBM SPSS Statistics[5]), mis võimaldavad elukestust analüüsida, kuid need tööriistad vajavad sisendiks andmestikku, kus on vajalikud elukestust iseloomustavad näitajad juba olemas. Selle töö raames valminud pakett ja eelnevalt tutvustatud paketid tekitavad need andmed OMOP ühtse andmemudeli andmetest ise. Kokkuvõttes on loodud tarkvarapakett erinev teistest lahendustest, sest see pakub lihtsat graafilist kasutajaliidest mitme elukestusanalüüsi meetodi samaaegseks rakendamiseks paljudel diagnoosidel otse OMOP CDM andmetel.

[3] https://github.com/GreeteKelli/cohortSurvivalAnalysis
[4] https://www.graphpad.com/features
[5] https://www.ibm.com/products/spss-statistics

2. Metoodika

Loodud tööriist võimaldab läbi viia kohort analüüsi korraga paljudele ekspositsiooni-tulemi paaridele. Järgnevalt tutvustatakse analüüsimetoodikat, mida selles protsessis vaja läheb.

2.1 Uuringu struktuur

Selle töö raames valminud rakendus uurib seost valitud diagnooside ehk ekspositsioonide ja neile järgnevate teiste diagnooside ehk tulemite esinemise vahel. Elukestusanalüüsi erinevate meetoditega võrreldakse aega tulemi tekkimiseni kahe grupi vahel. Need grupid on vastavalt eksponeeritud (ingl *exposed*) ja eksponeerimata (ingl *unexposed*) patsientide kohort. Terminit „ekspositsioon“ (ingl *exposure*) kasutatakse siin töös Rahvatervishoiu sõnastiku[6] definitsioonide järgi, tähistades sellega nii terviseseisundit mõjutavat tegurit, mille võimalikku seost tulemiga hinnatakse, kui ka kokkupuudet selle teguriga. „Eksponeeritud“ ja „eksponeerimata“ tähistavad samuti ekspositsiooni olemasolu. Uuringuks rakendatakse sobitamisega kohortuuringu ülesehitust. Nii ekspositsioonidel kui ka tulemitel vaadeldakse analüüsi käigus ainult esimesi toimumisi. Iga uuritava ekspositsioon-tulem paari puhul kaasatakse lähtepopulatsioonist uuringusse patsiendid, kellel ei ole vaatlusperioodi algusest kahe aasta jooksul esinenud tulemit ega ekspositsiooni. See suurendab võimalust, et vasakult tõkestatud andmetes esinev tulem ja ekspositsioon on tõesti esimene esinemine ja mitte patsiendi jaoks krooniline/korduv diagnoos, ingliskeelsed terminid sellise meetodi kohta on *run-in* või *washout period* [17], [18]. Eksponeeritud kohorti arvatakse patsiendid, kui neil esineb uuritav diagnoos ja enne seda ei ole neil esinenud tulemit. Eksponeerimata patsientide kohort moodustatakse sobitamise teel. Järgmisena kirjeldataksegi eksponeerimata patsientide kohordi ehk nn kontrollgrupi sobitamist.

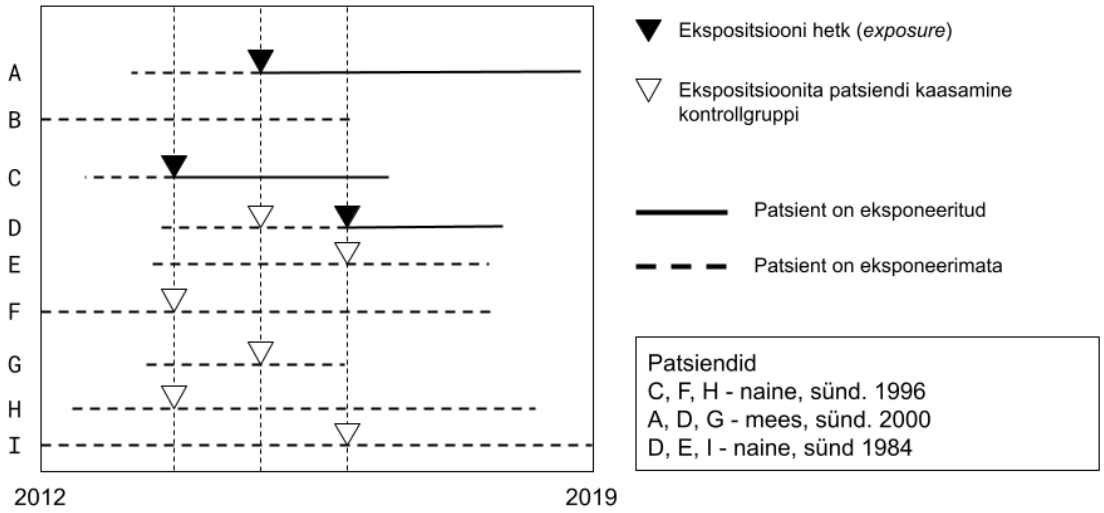
2.2 Kontrollgrupi sobitamine

Sobitamine (ingl *matching*) on meetod, mis seisneb uuringu patsientide valimises viisil, et võrreldavatel gruppidel oleksid teatud tunnuste osas sarnased jaotused, näiteks sugu ja vanus [19]. Võrreldes juhusliku valimiga võimaldab see vähendada segavate tegurite mõju tulemustele. Sellel põhjusel on sobitamine kasutusel ka siin töös.

Valminud rakendus kasutab sobitamist kontrollgrupi loomiseks. Igale eksponeeritud patsiendile valitakse ekspositsiooni hetkel võrdluseks *n* soo ja vanuse poolest sobivat patsienti,

[6] https://sonaveeb.ee/search/unif/dlall/rtrv/ekspositsioon/1/est

kes on selleks hetkeks eksponeerimata ja ka tulemi esinemiseta (vt Joonis 4). Lisaks peavad sobivad patsiendid olema olnud selleks hetkeks vaatluse all vähemalt kaks aastat.



Joonis 4. Sobitamist selgitav joonis, kus valitakse igale eksponeeritud patsiendile kaks sobivat eksponeerimata patsienti. Joonis on tehtud Iwagami ja Shinozaki [19] eeskujul.

Vanuse arvestamisel valitakse *n* lähimat naabrit, soo puhul täpne vaste. Joonise 4 patsiendi D põhjal on näha samuti, et patsiendid saavad olla ühe patsiendi jaoks kontrollgrupis ja hiljem ka eksponeeritud grupis, kus neile sobitatakse omakorda ekspositsioonita kontrollid. Sobitatavate patsientide arvuks *n* on siin töös võetud neli, mis on sageli kasutatav suhe sobitatud uuringutes [19]. Järgmises alapeatükis tutvustatakse analüüsis kasutatavaid elukestusanalüüsi meetodeid.

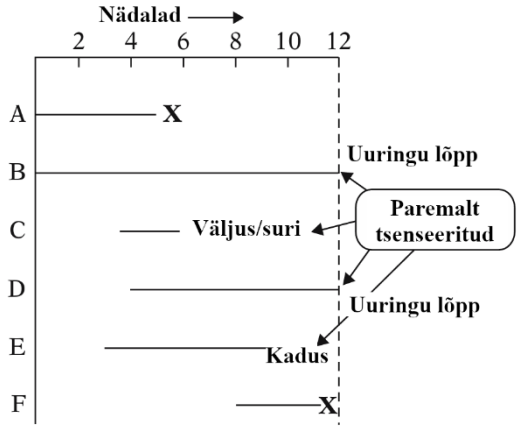
2.3 Elukestusanalüüs

Siin alapeatükis kirjeldatakse paketis rakendatavaid elukestusanalüüsi meetodeid. Kõigepealt tuuakse esile põhjused, miks on elukestus andmete analüüsiks vaja erilisi meetodeid.

2.3.1 Elukestus andmete eripärad

Elukestus andmete analüüsimiseks on loodud eraldi statistilised meetodid, kuna neil on mitmed eripärad võrreldes teiste andmetega [2], [20]. Esiteks ei ole Colletti sõnul [2] sellised andmed normaaljaotuse järgi jaotatud. See ilmneb tema väitel näiteks sarnaste inimeste eluaegades, kus suurem osa surma toimumistest esineb ajavahemiku hilisemas otsas. Teiseks mõjutab elukestus andmeid tsenseeritus, mis Kleinbaumi ja Kleini sõnul [3] tähendab, et kõikide uuritavate elukestused ei ole täpselt teada. Nad toovad välja kolm viisi, kuidas see tavaliselt juhtub (vt Joonis 1). Esiteks, kui patsiendil ei esine uuringu perioodi jooksul uuritavat sündmust. Teiseks siis, kui patsient kaob uuringust mistahes põhjusel. Kolmandaks, kui patsient astub uuringust

välja või sureb ja surm pole uuringus uuritav sündmus. Kõigil nendel kolmel juhul on tegemist paremalt tsenseerimisega, mis on elukestus andmetes kõige rohkem levinud [3].



Joonis 1. Paremalt tsenseerimise põhjuste näited patsientidel A..F. X tähistab uuritava sündmuse toimumist [3].

Töös kasutatud andmebaas vaatleb perioodi alates aastast 2012 kuni 2019. Selle tõttu on andmed esiteks paremalt tsenseeritud ning lisaks ei ole varasemast ajaperioodist mingit infot. Sellistes vasakult piiratud andmetes ei ole teada, kas patsientidel on juba enne andmeperioodi algust esinenud uuritavaid terviseseisundeid. Hilisemas uurimismetoodika peatükis kirjeldatakse meetodit selle piirangu mõju vähendamiseks. Edasi kirjutatakse elukestus-analüüsi peamisest funktsioonist.

2.3.2 Elukestusfunktsioon

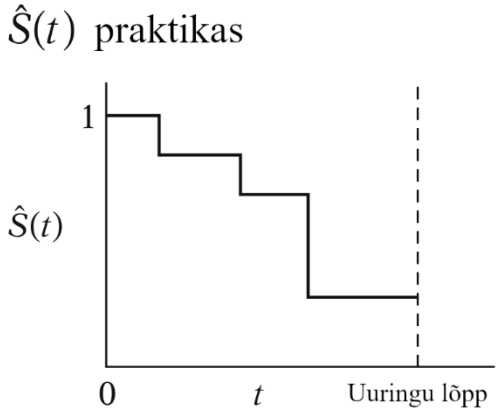
Elukestusfunktsioon või üleelamisfunktsioon (ingl *survival function* või *survivor function*) $S(t)$ on elukestusanalüüsi keskne funktsioon, mis kirjeldab tõenäosust, et subjekti elukestus T on suurem kui antud aeg t [2], [3]. Teisisõnu kirjeldab see tõenäosust, et uuritav sündmus ei toimu enne ajahetke t. Elukestusfunktsioon esitatakse järgnevalt [2]:

$$S(t) = P(T \geq t) = 1 - F(t) \quad (1),$$

kus $F(t)$ on T jaotusfunktsioon. Kleinbaum ja Klein toovad välja elukestusfunktsiooni $S(t)$ kolm omadust [3]:

- $S(t)$ on alati mittekasvav;
- ajahetkel $t = 0$ on $S(t)$ väärtus alati 1;
- ajahetkel $t = \infty$ on $S(t)$ väärtus alati 0.

Autorid toovad esile ka, et praktikas ei ole uuringu periood kunagi lõpmatu ja elukestusfunktsiooni $S(t)$ hinnangu $\hat{S}(t)$ graafik kahaneb sündmuse toimumise hetkedel sammude võrra (vt Joonis 2), näidates uuringus osalejatest nende hinnangulist proportsiooni, kellel pole selleks hetkeks veel uuritavat sündmust esinenud.



Joonis 2. Elukestusfunktsiooni hinnangu graafik praktikas [3].

Järgmisena kirjeldatakse päriselulistel andmetel elukestusfunktsiooni hindamiseks kasutatavat Kaplan-Meieri meetodit.

2.3.3 Kaplan-Meieri hinnang

Kaplan-Meieri hinnang (ingl *Kaplan-Meier estimate* või *product-limit estimate*) on mitteparameetriline statistiline meetod elukestusfunktsiooni hindamiseks ebatäielike vaatlusandmete korral [3], [20], [21]. Kaplan-Meieri hinnang avaldub järgneva valemi kujul [2]:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \qquad (2).$$

Valemis tähistab [2]:

- k erinevaid sündmuse toimumiste ajahetki;
- j hetkel vaadeldavat ajahetke;
- n_j vaadeldavate arvu, kellel ei ole enne ajahetke j toimunud uuritavat sündmust ega tsenseerimist;
- d_j vaadeldavate arvu, kellel toimub sündmus vaadeldaval ajahetkel j.

Kleinbaumi ja Kleini kirjeldusel [3] nimetatakse Kaplan-Meieri hinnangu graafikut Kaplan-Meieri kõveraks ning see langeb astmeliselt sündmuse toimumise ajahetkedel, nagu on näha hinnangu funktsiooni graafikul Joonisel 2. Kaplan-Meieri kõveraid saab autorite sõnul kasutada erinevate gruppide elukestuse võrdlemiseks, et näiteks uurida kahe erineva ravimeetodi mõju patsientide ellujäämisele. Nende statistiliseks võrdlemiseks kasutatakse sageli *log-rank* testi [3]. Kaplan-Meieri hinnangule saab Greenwoodi valemiga arvutada ka standardviga, mis näitab hinnangu täpsust, ning sellest saab omakorda moodustada usaldusintervalli [2]. Valminud rakenduses kasutatakse Kaplan-Meieri hinnangut ja selle graafikut patsientide elukestuse visualiseerimiseks, kus uuritavaks sündmuseks on mingi tulem-diagnoos.

2.3.4 *Log-rank* test

Log-rank test on statistiline meetod mitme Kaplan-Meieri kõvera võrdlemiseks [3]. Colletti sõnul [2] saab *log-rank* testi tulemusena arvutada p-väärtuse, mis näitab graafikute statistilist sarnasust ning millega saab kontrollida alternatiivse hüpoteesi kehtivust. Testi aluseks võetakse tema väitel nullhüpotees, et võrreldavate gruppide Kaplan-Meieri kõverad ei erine üksteisest statistiliselt. Kleinbaum ja Klein esitavad testi põhjaks oleva *log-rank* statistiku arvutamise valemi kahe võrreldava grupi puhul järgnevalt [3]:

$$log - rank\ statistik = \frac{(O_2 - E_2)^2}{Var(O_2 - E_2)} \quad (3).$$

Valemis 3 tähistab $O_2 - E_2$ vabalt valitud grupi (antud juhul grupi 2; grupi 1 tulemuseks oleks selle vastandarv) igal eristataval sündmuse toimumise hetkel $j \leq k$ toimunud vaadeldavate sündmuste arvu O_2 ja oodatava sündmuste arvu E_2 vahet:

$$O_2 - E_2 = \sum_{j=1}^{k} (m_{2j} - e_{2j}) \quad (4).$$

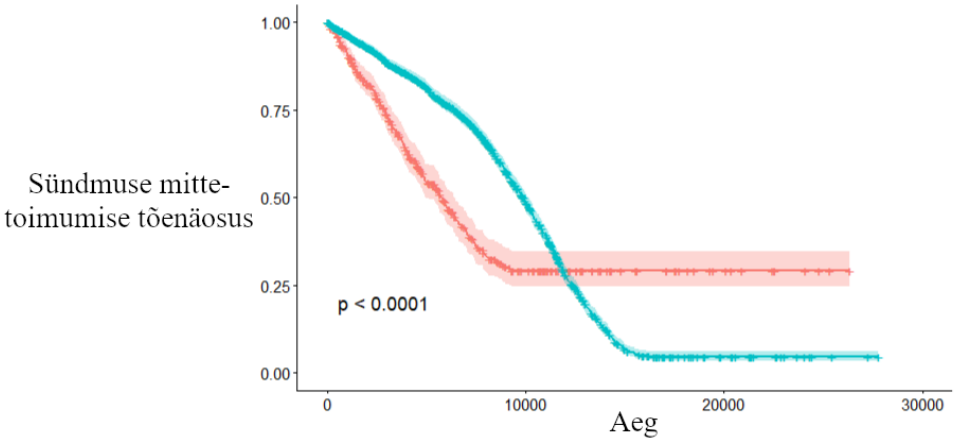
Oodatav sündmuste arv e_{2j} avaldub kujul:

$$e_{2j} = \left(\frac{n_{2j}}{n_{1j} + n_{2j}}\right) \times (m_{1j} + m_{2j}) \quad (5),$$

kus n_{ij} tähistab grupi *i* isikute arvu, kes pole enne hetke *j* veel sündmust kogenud ja m_{ij} grupi *i* huvipakkuvate sündmuste toimumiste arvu hetkel *j*. Seega on grupi oodatav sündmuste arv hetkel j võrdne korrutisega, mis koosneb selle grupi osakaalust kõigist selle hetkeni

„ellujäänud“ isikutest ja mõlemas grupis hetkel j toimunud sündmuste arvust [3]. $Var(O_2 - E_2)$ valemis 3 tähistab $O_2 - E_2$ dispersiooni ehk varieeruvust [3].

Alloleval joonisel (Joonis 3) on kahe Kaplan-Meieri kõvera võrdlemisel *log-rank* testiga tulemuseks p-väärtus < 0.0001, mis tähendab, et kaks uuritavat gruppi erinevad üksteisest väga suurel määral ning tõendid nullhüpoteesi vastu on ülekaalukad [2].



Joonis 3. Kaks Kaplan-Meieri kõverat ja nende võrdlemisel *log-rank* testiga saadud p-väärtus.

Selles töös võrreldakse *log-rank* testiga elukestust tulemi esinemiseni ekspositsiooniga grupi ja neile sobitatud ekspositsioonita patsientide grupi vahel. Järgmises alapeatükis tutvustatakse riskifunktsiooni ja riskide suhet, mis on samuti valminud tarkvarapaketis diagnooside seose väljendamiseks kasutusel.

2.3.5 Riskifunktsioon ja riskide suhe

Lisaks elukestusfunktsioonile on elukestusanalüüsis ka teine tähtis funktsioon – riskifunktsioon. Kui elukestusfunktsioon kirjeldab ellujäämistõenäosust ajas, siis riskifunktsioon kirjeldab vastupidiselt tõenäosust, et uuritav sündmus toimub ajahetkel *t*, eeldusel, et subjektil pole enne seda hetke sündmust toimunud [2], [3]. Riskifunktsioon *h*(*t*) avaldub valemina [2]:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \right\} \quad (6).$$

Valemis tähistab δt nullile lähenevat ajavahemikku ja $P(t \leq T < t + \delta t \mid T \geq t)$ tingimuslikku tõenäosust, et elukestus T jääb ajahetkede t ja $t + \delta t$ vahele, kui T on suurem kui ajahetk t [2]. Elukestusfunktsioon ja riskifunktsioon on omavahel seotud ning teades ühte, on võimalik arvutada teine [3].

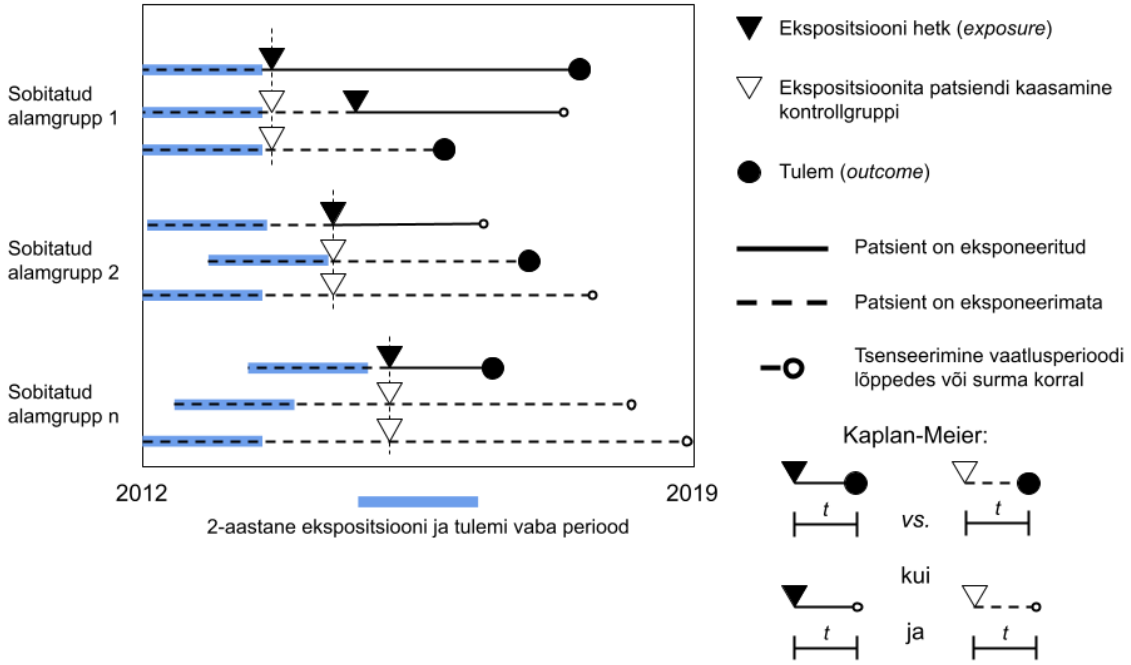
Kui $h_1(t)$ ja $h_2(t)$ on kahe grupi riskifunktsioonid, siis eeldusel, et funktsioonide väärtuse erinevus gruppide vahel ei sõltu ajast, avaldub võrdeliste riskide mudel kujul:

$$h_1(t) = \psi h_2(t) \quad (7),$$

kus ψ on konstant, mida nimetatakse riskide suhteks (ingl *hazard ratio*) [2]. Collett [2] toob ka välja, et kui $\psi < 1$, siis on sündmuse toimumise risk esimesel grupil väiksem kui teisel, ja kui $\psi > 1$, siis on vastupidi. Valemis 7 on allika põhjal esitatud võrdeliste riskide mudeli lihtne kuju ning keerulisemaid väljendusi, mida kasutatakse statistilistes meetodites, siin ei tutvustata. Selle töö raames loodud R-paketis arvutatakse riskide suhet eksponeeritud ja eksponeerimata patsientide gruppide vahel, kasutades Coxi võrdeliste riskide mudelit. Saadud riskide suhtele arvutatakse mudelis ka p-väärtus (Wald testiga), mis väljendab tõenäosust saada vaadeldud tulemus juhul, kui tegelikkuses gruppide vahel erinevust ei esine (st riskide suhe oleks 1) [2]. Edasi võetakse lühidalt kokku elukestusanalüüsi meetodite rakendamine analüüsis.

2.3.6 Elukestusanalüüsi rakendamise metoodika

Iga uuritava ekspositsiooni ja tulemi paari jaoks rakendatakse mitmeid elukestusanalüüsi meetodeid. Analüüsi algushetkeks on iga sobitatud alamgrupi siseselt ekspositsiooni toimumise hetk. Lõpphetkeks on iga patsiendi puhul kas tulemi toimumine või tsenseerimine tema vaatlusperioodi katkemisel. Võrreldavate kohortide hinnangulist tulemi mitte-toimumise tõenäosust aja jooksul visualiseeritakse Kaplan-Meieri kõveratega. Kohortide elukestusjaotuste võrdlemiseks kasutatakse *log-rank* testi. Riskide suhte ja selle usaldusvahemiku arvutamiseks kasutatakse Coxi võrdeliste riskide mudelit. Joonisel 5 on visualiseeritud rakenduses elukestusanalüüsiga uuritavad suurused.



Joonis 5. Elukestusanalüüsi protsessi kirjeldus ühe ekspositsioon-tulemi paari jaoks. Joonis on tehtud Iwagami ja Shinozaki [19] eeskujul.

Järgmisena kirjeldatakse meetodit mudelite rakendamisel saadavate p-väärtuste korrigeerimiseks, mis on ka selles töös vajalik.

2.4 Holm-Bonferroni meetod

Kuna loodud tarkvarapaketis uuritakse korraga suures koguses diagnoose ja seega tehakse suures koguses statistilisi p-väärtuse teste, siis kerkib esile mitmese võrdluse probleem (ingl *multiple comparisons problem*) [22]. Allika sõnul tähendab see seda, et suure testide kogusega suureneb võimalus teha esimest tüüpi vigasid ehk saada valepositiivseid tulemusi statistilise olulisuse kohta. Probleemi vältimiseks kasutatakse paketis *log-rank* testi ja riskide suhte p-väärtuste korrigeerimiseks Holm-Bonferroni meetodit. See meetod järjestatab testide p-väärtused ning võrdleb neid astmeliselt korrigeeritud olulisuslävenditega, lükates null-hüpoteese tagasi ükshaaval alates kõige statistiliselt olulisemast, kuni tingimused enam täidetud pole [23]. Korrektsioon tehakse siin töös kõikide testide üleselt, kus mudeleid rakendati. Kuigi laialdaselt kasutatakse Bonferroni meetodit, siis on selle edasiarendus, Holm-Bonferroni meetod, statistiliselt võimsam ning Aickini ja Gensleri sõnul [24] igati parem. Edasi kirjeldatakse loodud tarkvarapaketti, mis koondab endasse mainitud analüüsimeetodid.

3. R-pakett *Exposure2OutcomeSurv*

Selles peatükis tutvustatakse R-paketi *Exposure2OutcomeSurv* loomiseks kasutatud tehnoloogiat ning kirjeldatakse selle kasutajaliidest ja töövoogu. Lisaks kirjutatakse paketi testimisest. Paketi lähtekood on kättesaadav GitHubis[7] ja dokumentatsioon koos kasutusjuhendiga ka GitHub Pages lehel[8]. Lähtekoodi kirjutamisel kasutati Google Gemini suure keelemudeli erinevate versioonide [25], [26] abi. Abi seisnes peamiselt funktsioonide mustandite genereerimises, vigade otsimises ning dokumentatsiooni koostamises. Kogu kasutatud kood vaadati autori poolt põhjalikult üle ning kohandati vastavalt töö eesmärgile ja nõuetele, tagades selle korrektsuse ja funktsionaalsuse.

3.1 Tehnoloogia

OHDSI avatud teaduse kogukonna töövahenditega ja andmemudeliga sobitumiseks valiti rakenduse arendamiseks programmeerimiskeele R versioon 4.4.1 [27]. OMOP CDM andmebaasi ühenduse loomiseks kasutati OHDSI *DatabaseConnector* paketti ja andmebaasiga suhtluse lihtsustamiseks *CDMConnector* paketti [28], [29]. *CDMConnector* teeb CDM andmebaasi ja selle tabelid koodis objektidena kättesaadavaks. Nendest andmete pärimiseks saab sel juhul kasutada *dplyr* andmetöötluse grammatikat, mille *dbplyr* teisendab automaatselt konkreetsele andmebaasihaldurile sobivateks SQL-päringuteks [30], [31]. Eelnevas peatükis käsitletud elukestusanalüüsi meetodite rakendamiseks kasutati R-is kõige laialdasemalt kasutatavat elukestusanalüüsi paketti *survival* ning Kaplan-Meieri graafikute loomiseks *survminer* paketti [32], [33]. R-paketis sisalduva graafilise kasutajaliidese loomiseks kasutati *shiny* veebirakenduste raamistikku, mis võimaldab luua veebirakenduse samuti R koodis [34]. Tulemuste salvestamiseks ja lugemiseks valiti R keele baasteeki kuuluvad vahendid. Järgnevalt tutvustatakse loodud graafilist kasutajaliidest.

3.2 Kasutajaliides

Kasutajalt sisendandmete saamiseks ja analüüsi tulemuste kuvamiseks loodi *shiny* veebirakendus. Veebirakendus on jaotatud kahte põhikomponenti: *app_ui*, mis paneb paika rakenduse kasutajaliidese vormingu, ja *app_server*, mis juhib kasutajaliidesest saadud sisendi põhjal analüüsiprotsessi ning loob tulemuste kuvamise komponendid. Kasutaja sisend toimub põhiliselt rakenduse külgriba kaudu (vt lisa I). Uuritavate diagnooside valimiseks saab

[7] https://github.com/sandersoodla/Exposure2OutcomeSurv
[8] https://sandersoodla.github.io/Exposure2OutcomeSurv/

kasutada otsingut või laadida üles CSV-failid. Analüüsi tulemused salvestatakse kasutaja määratud nimega faili, kust neid saab kasutajaliidese kaudu ka hiljem uuesti vaatamiseks sisse laadida. Lisaks külgribale on rakendusel kaks vahelehte:

- analüüsi kokkuvõte (ingl *analysis summary*);
- demograafiline ülevaade (ingl *demographic overview*).

Analüüsi kokkuvõtte vahelehel (vt lisa II) on kokkuvõttev tabel vaadeldava tulemuste komplekti kohta. Tabel sisaldab ekspositsioonide ja tulemite paaride:

- riskide suhet (HR, ingl *hazard ratio*) koos 95%-usaldusvahemikuga (CI, ingl *confidence interval*);
- Holm-Bonferroni meetodiga korrigeeritud p-väärtust riskide suhtele ja *log-rank* testi tulemusele;
- uuritud eksponeeritud ja eksponeerimata patsientide arvu;
- mõlemas grupis uuritud patsientide arvu, kellel tekkis tulem.

Tabeli ridade valimisel saab näha ka vastavaid Kaplan-Meieri graafikuid ning tabelit saab salvestada CSV-failina või Exceli failina.

Demograafilise ülevaate vahelehel (vt lisa III) saab näha iga uuritava diagnoosi soolist ja vanuselist jaotust diagnoosi saamise hetkedel, mis on lisainfoks andmete tõlgendamisel. Järgmisena kirjeldatakse nende kuvatavate tulemuste saamise töövoogu.

3.3 Töövoog

Valminud tarkvarapaketi töövoog sisaldab kasutaja sisendi põhjal andmebaasist andmete pärimist, andmete ettevalmistamist elukestusanalüüsiks, elukestusanalüüsi mudelite rakendamist ning tulemuste salvestamist, lugemist ja kuvamist graafilises kasutajaliideses. Töövoo ja ülesehituse visuaalne selgitus on leitav lisas IV. Vajalike CDM lähteandmete saamiseks ja elukestusanalüüsi meetodite sisendandmete loomiseks kasutatakse veebirakenduse serveris funktsiooni *calculateMatchedSurvivalData*. Analüüsi meetodite rakendamiseks on funktsioonid *generateKmPlotObjects* ja *calculateCoxResults*. Selles peatükis kirjeldatakse neid samme lähemalt.

3.3.1 Andmete pärimine

Andmebaasiühendus OMOP CDM andmebaasiga luuakse *DatabaseConnector* paketiga. Ühenduse parameetrid nagu andmebaasi aadress, port, kasutajanimi ja parool saadakse kasutaja

keskkonna failist *.Renviron* keskkonnamuutujatena. *Exposure2OutcomeSurv* toetab kõiki andmebaasihaldureid (DBMS), mida toetab *DatabaseConnector*, ning lisaks ka kohalikke DuckDB[9] faile. DuckDB failide puhul kasutatakse ühendamiseks *DBI* ja *duckdb* R-pakette [35], [36]. Tabelis 1 on välja toodud kõik OMOP CDM tabelid ja väljad, mida loodud paketis kasutatakse.

Tabel 1. Kasutatud OMOP CDM tabelid ja väljad koos selgitusega.

Tabel	**Väli**	**Selgitus**
`person` (Isikud)	`person_id`	Isiku unikaalne kood
	`gender_concept_id`	Sugu
	`year_of_birth`	Sünniaasta
`condition_occurrence` (Terviseseisundite esinemised)	`condition_concept_id`	Esinenud terviseseisundi (diagnoosi) kood
	`person_id`	Seotud isiku kood
	`condition_start_date`	Diagnoosi kuupäev
`observation_period` (Vaatlusperioodid)	`person_id`	Seotud isiku kood
	`…_start_date`	Vaatlusperioodi alguskuupäev
	`…_end_date`	Vaatlusperioodi lõppkuupäev
`concept` (Meditsiinilised kontseptsioonid standardsete koodidena)	`concept_id`	Kontseptsiooni kood
	`concept_name`	Kontseptsiooni nimi
	`domain_id`	Kontseptsiooni liik
`concept_relationship` (Kontseptsioonide seosed)	`concept_id_1`	Kontseptsioon 1
	`concept_id_2`	Kontseptsioon 2
	`relationship_id`	Seos kontseptsiooni 1 ja 2 vahel

Kui kasutaja sisestab uuritavad diagnoosid failidega, siis viiakse enne protsessi jätkamist failis võimalikud mittestandardsed diagnooside koodid üle OHDSI standardse sõnavara koodideks. Seda tehakse funktsiooniga *mapInputToStandardIds*, mis kasutab *concept_relationship* tabelit. Rakenduses saab kasutada näiteks ATHENA[10] või ATLAS[11] tööriistadest saadud terviseseisundite (ingl *condition*) faile, aga töötavad ka kõik teised CSV-failid, kus on üks tulpadest nimega *Id*, *condition_source_concept_id*, *concept_id* või *condition_concept_id*. Otsingu kaudu diagnoose sisestades on diagnoosid juba standardsed. Funktsiooni *calculateMatchedSurvivalData* esimese etapina päritakse andmebaasist vajalik info, kasutades

[9] https://duckdb.org/
[10] https://athena.ohdsi.org/
[11] https://atlas-demo.ohdsi.org/

abifunktsiooni *fetchDataForSurvAnalysis*. Elukestusanalüüsiks vajalikud andmed on siinkohal demograafilised andmed (sugu, sünniaasta), vaatlusperioodid ja uuritavate diagnooside esimeste esinemiste kuupäevad. Järgmises peatükis kirjeldatakse nende andmete töötlemist edasistele analüüsimeetoditele sobivaks.

3.3.2 Andmetöötlus

Andmetöötluse etapp seisneb peatükkides 2.1 ja 2.2 paika pandud uurimismetoodika rakendamisega Kaplan-Meieri ja Coxi meetoditele vajalike lähteandmete tekitamises. Selleks kasutab *calculateMatchedSurvivalData* funktsioon iga ekspositsioon-diagnoosi ja tulem-diagnoosi paari jaoks mitut abifunktsiooni. Funktsioon *filterByWashoutAndGetOutcomeDates* rakendab andmetele uuringusse kaasamise eelduse, et patsientidel ei tohi olla vaatlusperioodi algusest kahe aasta jooksul esinenud tulemit. *defineExposedCohortForPair* defineerib uurimise metoodikat arvesse võttes eksponeeritud patsientide kohordi, mille igale patsiendile sobitatakse *performPairMatching* meetodiga soo ja vanuse põhjal neli eksponeerimata patsienti. Lõpuks luuakse funktsiooniga *calculatePairSurvival* järgmise etapi jaoks vajalikud elukestusandmed. Joonisel 6 on ühe diagnooside paari jaoks saadud tabeli struktuur, kus:

- *set_id* on sobitatud alamgrupi number;
- *exposure_status* näitab, kas patsient oli eksponeeritud või eksponeerimata;
- *time_to_outcome* on aeg päevades ekspositsiooni kuupäevast (*index_date*) uuringust väljumise kuupäevani (*study_exit_date*), kas tsenseerimise või tulemi tekkimise tõttu;
- *outcome_status* näitab, kas tulem tekkis või mitte.

	person_id	set_id	exposure_status	index_date	study_exit_date	time_to_outcome	outcome_status
	<int>	<int>	<dbl>	<date>	<date>	<dbl>	<dbl>
1	191	1	1	1923-02-08	2003-07-22	29384	0
2	1870	1	0	1923-02-08	1961-02-26	13898	0
3	2827	1	0	1923-02-08	1924-02-11	368	1
4	3811	1	0	1923-02-08	2017-01-06	34301	0
5	4017	1	0	1923-02-08	2014-08-15	33426	0
6	154	2	1	1980-09-23	2017-07-20	13449	0
7	1228	2	0	1980-09-23	2016-03-27	12969	1
8	1622	2	0	1980-09-23	2019-04-23	14091	0
9	1769	2	0	1980-09-23	1989-03-10	3090	1
10	3048	2	0	1980-09-23	2018-05-30	13763	0

Joonis 6. Arvutatud elukestusandmete tabeli näide.

Neid andmeid kasutatakse järgnevate elukestusanalüüsi meetodite sisendina.

3.3.3 Elukestusanalüüsi rakendamine

Töödeldud andmetele saab järgmisena rakendada metoodikas välja toodud elukestusanalüüsi meetodeid. Kaplan-Meieri graafikute loomiseks ja *log-rank* testi p-väärtuste saamiseks on

funktsioon *generateKmPlotObjects*. Mudeli valem koosneb paketi *survival* funktsiooniga *Surv* loodud elukestus objektist ning *group* muutujast, mis ütleb mudelile, et tulemused arvutataks kahe võrreldava patsiendigrupi jaoks eraldi (põhineb *exposure_status* tulbal, vt Joonis 6):

```
kmFit <- survival::survfit(survival::Surv(time_to_outcome,
outcome_status) ~ group, data = survData)
```

Kaplan-Meieri graafikud luuakse hiljem *kmFit* mudeli põhjal, kasutades *survminer* paketi funktsiooni *ggsurvplot*. Coxi võrdeliste riskide mudelit rakendatakse funktsioonis *calculateCoxResults* paketi *survival* käsuga *coxph*. Mudeli valemis võetakse arvesse sobitatud alamhulkasid (*cluster(set_id)*) ning jällegi seda, kas patsient oli eksponeeritud või mitte (*exposure_status*):

```
coxFit <- survival::coxph(survival::Surv(time_to_outcome,
outcome_status) ~ exposure_status + cluster(set_id),
data = currentSurvData)
```

Eelnevas Kaplan-Meieri mudelis alamhulkasid parameetrina arvesse ei võetud, sest tegu on mitteparameetrilise mudeliga. Kui vähemalt ühes grupis ei esinenud kellelgi tulemit, siis Coxi mudelit ei rakendata ning riskide suhte andmed jäetakse tühjaks, sest see tuleks lõpmatu või 0. Enne tulemuste kuvamist ja salvestamist korrigeeritakse mittetühjad *log-rank* testi ja Coxi mudelist saadud riskide suhte p-väärtused Holm-Bonferroni meetodiga, kasutades R-i sisseehitatud *p.adjust* funktsiooni. Tulemused salvestatakse kasutaja seadistatud kausta .rds laiendiga faili ja neid näidatakse kasutajaliidese peatükis mainitud analüüsi kokkuvõtte vahelehel. Kasutatav RDS-formaat on mõeldud R keele objektide faili kirjutamiseks ja lugemiseks, mis vaikimisi ka tihendab sisu [27]. Tulemuste faili saab R-i sisse lugeda ka väljaspool paketi kasutajaliidest. Edasi kirjeldatakse loodud paketi koodi testimist.

3.4 Testimine

R-paketi testimiseks kasutati ühiktestimise paketti *testthat* [37]. Testimisel keskenduti elukestusandmeid ettevalmistavale funktsioonile *calculateMatchedSurvivalData* ja selle erinevatele abifunktsioonidele, et veenduda Kaplan-Meieri ja Coxi mudelite sisendandmete korrektsuses. Kokku kirjutati 14 erinevat ühiktesti, millest kümme testivad *calculateMatchedSurvivalData* funktsiooni ja neli tükki erinevaid abifunktsioone. Abifunktsioonide testides on igasse testi plokki koondatud mitu juhtumit, seega on paketis testjuhtumeid kokku 30 ning erinevaid individuaalseid väärtuste kontrolle 158. Testimise täiustamiseks saaks testida ka rakenduse teisi osasid ja kasutajaliidest.

4. Tulemused

Selle peatüki esimeses osas kirjeldatakse paketi funktsionaalsuse demonstreerimiseks läbi viidud näidisuuringu diagnooside valikut ja tähtsamaid tulemusi. Peatüki teises osas käsitletakse rakenduse võimalikke edasiarendusi.

4.1 Näidisuuring

Töö raames loodud tarkvarapaketi võimekuse katsetamiseks analüüsiti sellega valitud täpsustamata terviseseisundite mõju erinevate kasvajate esinemisele. Edasi tutvustatakse kõigepealt täpsemalt uuringusse valitud diagnoose ning seejärel saadud tulemusi.

4.1.1 Diagnooside valik

RHK ehk rahvusvaheline haiguste klassifikatsioon (ingl ICD, *international classification of diseases*) on rahvusvaheline standard, mis määrab igale haigusele, vigastusele või seisundile unikaalse koodi [38]. RHK-10 XVIII peatükk (koodid R00–R99) hõlmab täpsustamata terviseseisundeid, mis on defineeritud kui „sümptomid, tunnused ja ebanormaalsused ning ebamäärased haigusseisundid, millele ei saa panna mujal klassifitseeritud diagnoosi“ [39]. Rakenduse töö demonstreerimiseks valiti juhendaja eelvaliku alusel RITA-MAITT andmebaasist 30 enim esinenud täpsustamata RHK-10 terviseseisundit. Nende hulka võeti eelmainitud täpsustamata terviseseisundeid ja ka teisi diagnoose, mis sisaldavad sõna „*unspecified*“ ehk „täpsustamata“. Kogu uuritud täpsustamata diagnooside sisendfaili sisu on nähtav töö lisas V. Kuna OMOP ühtse andmemudeli terviseseisundite tabel RHK diagnoose otse ei kasuta, kaardistati need paketis standardseteks meditsiiniterminite kontseptsioonideks (ingl *standard concept*). Selle kaardistamise tulemusena vastasid näiteks RHK koodid „R07.3“ ja „R07.4“ mõlemad samale standardsele diagnoosile „Chest pain“ (kood 77670), mis tähendas, et uuritavate unikaalsete standardsete ekspositsioonide arv oli 29. Tulemiteks võeti RHK-10 koodide jaotisest C30–C39 ehk hingamiselundite ja rindkeresiseste elundite pahaloomulistest kasvajatest 20 diagnoosi, mis paketi kasutajaliidesesse sisestades kaardistati 18-ks erinevaks andmestikus esinenud standardseks diagnoosiks. Tulemiteks valitud kasvajate sisendfail on leitav lisas VI. Seejärel teostati valitud diagnooside analüüs.

4.1.2 Uuringu tulemused

Uuringu tulemuseks saadi statistika 522 diagnoosikombinatsiooni kohta (29 ekspositsiooni × 18 tulemit). Kõigi tulemuste tabel on lõputööga kaasa pandud arhiivis, mida tutvustatakse lisas

VII. Statistiliselt kaks kõige olulisemat tulemust riskide suhte p-väärtuse järgi tulid täpsustamata hüpotüreoosi (kilpnäärme alatalitlus) ja täpsustamata stenokardia (rinnaangiin) ning sagarate, bronhi või kopsu pahaloomuliste kasvajate vahel (vt Tabel 2). Need täpsustamata diagnoosid kaardistati rakenduses standardseteks diagnoosideks „hüpotüreoos“ ja „stenokardia“, mis on standardses sõnavaras neile lähimad vastavad kontseptsioonid. Diagnooside analüüsi kaasatud ja tulemi esinemisega patsientide arvud on esitatud tabelis 3.

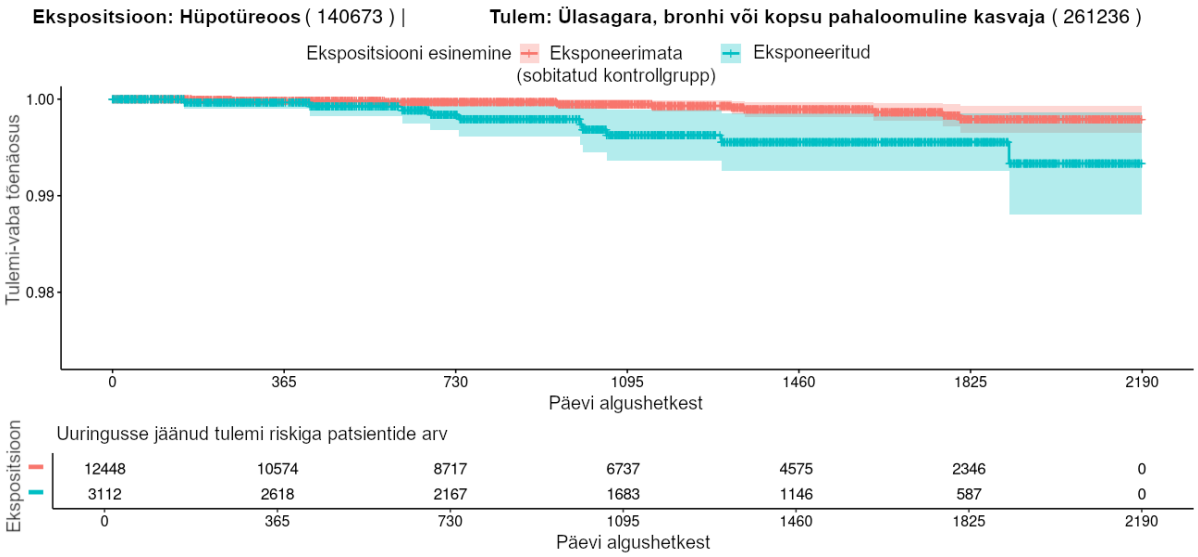
Tabel 2. Kahe statistiliselt kõige olulisema tulemuse näitajad.

Ekspositsioon	**Tulem**	**Riskide suhe (*HR*)**	**Riskide suhte 95% usaldus-vahemik**	***HR* p-väärtus**	***Log-rank* testi p-väärtus**
Täpsustamata hüpotüreoos RHK-10: E03.9 Standardne diagnoos: Hüpotüreoos (kood 140673)	Ülasagara, bronhi või kopsu pahaloomuline kasvaja (kood 261236) vastav RHK-10: C34.1	3.65	1.612 – 8.243	**0.00189** Holm-Bonferroni: 0.168	**0.00152** Holm-Bonferroni: 0.249
Täpsustamata stenokardia RHK-10: I20.9 Standardne diagnoos: Stenokardia (kood 321318)	Kesksagara, bronhi või kopsu pahaloomuline kasvaja (kood 256646) vastav RHK-10: C34.2	12.1	2.441 – 60.103	**0.00227** Holm-Bonferroni: 0.200	**0.00009** Holm-Bonferroni: 0.0161

Tabel 3. Analüüsi kaasatud ja tulemi esinemisega patsientide arvud.

Ekspositsioon	**Tulem**	**Tulemi saanud / ekspositsiooniga patsientide koguarv**	**Tulemi saanud / ekspositsioonita patsientide koguarv**
Täpsustamata hüpotüreoos RHK-10: E03.9 Standardne diagnoos: Hüpotüreoos (kood 140673)	Ülasagara, bronhi või kopsu pahaloomuline kasvaja (kood 261236) vastav RHK-10: C34.1	10 / 3112	11 / 12448
Täpsustamata stenokardia RHK-10: I20.9 Standardne diagnoos: Stenokardia (kood 321318)	Kesksagara, bronhi või kopsu pahaloomuline kasvaja (kood 256646) vastav RHK-10: C34.2	6 / 4677	2 / 18708

Nende diagnoosipaaride riskide suhte 95%-usaldusvahemik (vt Tabel 2) oli märkimisväärselt üle ühe, mis tähendab, et siin analüüsis oli hüpotüreoosi ja stenokardia esinemisel suurem risk saada välja toodud pahaloomulisi kasvajaid, võrreldes patsientidega kellel neid ekspositsioone ei esinenud. Ka riskide suhte statistilist olulisust näitav p-väärtus (vt Tabel 2 *HR* p-väärtus) jääb tulemustel märgatavalt alla olulisuse piiri p = 0,05 (95%-kindluse piir), mistõttu selle järgi saab väita, et erinevus võrreldud gruppide vahel eksisteerib. Vastavad Holm-Bonferroni meetodiga korrigeeritud p-väärtused jäävad liiga suureks, et sama nivooga statistilist olulisust väita. Joonisel 7 on hüpotüreoosi ja ülasagara, bronhi või kopsu pahaloomulise kasvaja tulemuseks saadud Kaplan-Meieri graafik. Ka sellelt on näha väikest erinevust kasvaja esinemises hüpotüreoosiga ja hüpotüreoosita patsientide vahel. Päevade arv algushetkest näitab aega hüpotüreoosi diagnoosi saamise hetkest.



Joonis 7. Hüpotüreoosi ja ülasagara, bronhi või kopsu pahaloomulise kasvaja Kaplan-Meieri graafik. Tõlgitud eesti keelde.

Küll aga oli tulemi saanud patsientide arv mõlemas võrreldavas grupis väga väike (vt Tabel 3), seega võib vaadeldud seos diagnooside vahel kehtida, kuid selleks, et seda kindlamalt väita, oleks vaja edasist analüüsi, näiteks suuremal andmebaasil või üldisemate diagnoosidega. Saadud tulemuste põhjal saab võtta need diagnoosid edasise uurimise alla ja veenduda seose tegelikus olemasolus teiste uurimismeetoditega. Kuigi näidisuuringus ilmnenud seosed vajavad edasist kinnitust, tõestas uuring loodud *Exposure2OutcomeSurv* paketi toimivust ja võimekust avastada haiguste seoseid, mille uurimisse suunata rohkem tähelepanu.

Analüüsitud 522 diagnoosipaari tulemuste saamiseks kulus 6 tundi 49 minutit, ehk ühe paari kohta umbes 47 sekundit, kuid ühe diagnoosipaari jaoks kuluv aeg sõltub ekspositsiooniga patsientide arvust, sest neile sobitatakse võrdlusgrupp. Kuigi jõudlust saaks kindlasti

parandada, siis suure diagnooside kogusega kulubki aega ja seega pole loodud rakendus mõeldud reaalajas tulemuste saamiseks, vaid korraga paljude diagnooside uurimiseks, mille tulemusi saab kasutaja hiljem vaadelda. Kiiresti tulemuste saamiseks saab valida uurimiseks ka väikse koguse diagnoose. Protsessi kõige aeganõudvam osa on ülekaalukalt patsientide võrdlusgruppide sobitamine, mis tehakse igal diagnoosipaaril eraldi, et arvestada ekspositsiooni ja tulemi varasema esinemise ning vaatlusaja kriteeriumitega. Töö käigus tehti mitmeid parandusi sobitamise efektiivsuse tõstmiseks, mille tulemusel saavutati eelmainitud ajakulu. Ilmselt oleks võimalik sobitamise protsessi kiirust veelgi edasi arendada, näiteks paralleelarvutust kasutades. Teistest, funktsionaalsetest, edasiarendusvõimalustest kirjutatakse järgnevas alampeatükis.

4.2 Võimalikud edasiarendused

Kuigi tulemusena õnnestus luua tarkvarapakett, mis vastab püstitatud eesmärgile, siis on valminud lahendusel mitmeid arendamisvõimalusi. Selleks, et valminud pakett saaks olla sobiv alternatiiv *CohortMethod* ja *CohortSurvival* pakettidele, siis peaks saama kasutajaliideses seada uuringu parameetreid (*washout* periood, sobitamise suhe, ajavahemik jne). Hetkel on loodud paketi kasutamine palju mugavam, kuid uuringustruktuuri kohandamise poolest vähem võimas. Lisaks oleks kasulik, kui uuritavaks tulemiks saaks võtta surma. Kasutaja soovil võiks saada rakendusse sisestatud üldisematesse diagnoosidesse kaasata kõik selle alamdiagnoosid, et vähendada diagnooside killustamist liiga spetsiifilisteks osadeks, mida võib olla raske uurida. Korraga mitme andmebaasiga töötavatele teadlastele oleks kasulik võimalus seadistada mitu andmebaasiühendust, mille vahel saaks kasutajaliideses valida. Demograafilise ülevaate vahelehe populatsioonipüramiide ei salvestata praeguses versioonis faili, seega neid saab ainult vaadata samas sessioonis, kus analüüs käivitati, või kui kasutaja salvestab need eraldi piltidena. Ka seda saaks parandada. Nende ja teiste potentsiaalsete täienduste elluviimine muudaks valminud paketi veelgi mitmekülgsemaks ja võimsamaks tööriistaks terviseandmete uurijatele.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli luua R programmeerimiskeele tarkvarapakett diagnooside vahelise mõju uurimiseks elukestusanalüüsiga. Elukestusanalüüs uurib aega mingi sündmuse toimumiseni ning töös uuriti erinevusi tulemi esinemiseni inimestel, kellel esines mingi varasem diagnoos, ja nendel, kellel seda ei esinenud. Valminud R-pakett *Exposure2OutcomeSurv* pärib kasutaja sisestatud diagnooside põhjal andmed OMOP CDM andmebaasist, valmistab andmed ette elukestusanalüüsiks, rakendab erinevaid elukestusanalüüsi meetodeid ning võimaldab lisaks tulemuste kuvamisele graafilises kasutajaliideses ka nende salvestamist ja lugemist. Võrreldavad patsientide grupid saadakse soo ja vanuse põhjal sobitamise teel. Pakett arvutab Coxi võrdeliste riskide mudeli abil võrreldavate gruppide riskide suhted ning visualiseerib elukestuse erinevusi ka Kaplan-Meieri graafikutega. Statistilist olulisust näitavate p-väärtuste korrigeerimiseks kasutatakse Holm-Bonferroni meetodit. Graafiline kasutajaliides eristab valminud tarkvarapaketti teistest lahendustest nagu OHDSI kohordiuuringu tööriist *CohortMethod* ning võimaldab laialdasemat ja mugavamat kasutust. Teiste graafilise kasutajaliidesega tööriistadega võrreldes on paketil eriline võimekus analüüsida nii Kaplan-Meieri kui ka Coxi mudeliga paljusid erinevaid diagnoose ning seda otse OMOP CDM andmetel. Loodud lahenduse tööd katsetati näidisuuringus RITA-MAITT andmebaasil. Näidisuuringus leiti potentsiaalne seos kahel diagnoosipaaril, kuid uuritud kasvajate vähese esinemise arvu tõttu ei saa sellest kindlaid järeldusi teha. Sellele vaatamata tõestas näidisuuring loodud paketi tehnilist toimivust ja võimekust avastada seoseid, mida edasi uurida. Töö lõpus toodi välja võimalused rakenduse edasi arendamiseks, millest olulisemad olid uuringu parameetrite seadmine ning tulemina surma uurimine.

Töö viidi läbi vastavalt TÜ eetikakomitee ja Eesti bioeetika ja inimuuringute nõukogu lubadele (load nr 300/T-23 ja 1.1-12/3088) ning projektide TEM-TA72 ja PRG1844 raames. Projekt TEM-TA72 on rahastatud Euroopa Liidu ja kaasrahastatud Haridus- ja Teadusministeeriumi poolt. Projekt PRG1844 on rahastatud Eesti Teadusagentuuri poolt.

Viidatud kirjandus

[1] I. Reinecke, M. Zoch, C. Reich, M. Sedlmayr, and F. Bathelt, 'The Usage of OHDSI OMOP – A Scoping Review', in *Studies in Health Technology and Informatics*, R. Röhrig, T. Beißbarth, J. König, C. Ose, G. Rauch, U. Sax, B. Schreiweis, and M. Sedlmayr, Eds., IOS Press, 2021. doi: 10.3233/SHTI210546.

[2] D. Collett, *Modelling survival data in medical research*, Fourth edition. in Texts in statistical science. Boca Raton: CRC Press, Taylor and Francis Group, 2023.

[3] D. G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text*. 2005. doi: 10.1007/0-387-29150-4.

[4] P. E. Stang *et al.*, 'Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership', *Ann. Intern. Med.*, vol. 153, no. 9, pp. 600–606, Nov. 2010, doi: 10.7326/0003-4819-153-9-201011020-00010.

[5] J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, and P. E. Stang, 'Validation of a common data model for active safety surveillance research', *J. Am. Med. Inform. Assoc.*, vol. 19, no. 1, pp. 54–60, Jan. 2012, doi: 10.1136/amiajnl-2011-000376.

[6] G. Hripcsak *et al.*, 'Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers.', *Stud. Health Technol. Inform.*, vol. 216, pp. 574–578, 2015.

[7] C. Reich and A. Ostropolets, 'Standardized Vocabularies', in *The Book of OHDSI*, Observational Health Data Sciences and Informatics, 2021. Accessed: Dec. 03, 2024. [Online]. Available: https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html

[8] P. Ryan and G. Hripcsak, 'The OHDSI Community', in *The Book of OHDSI*, Observational Health Data Sciences and Informatics, 2021. Accessed: Dec. 03, 2024. [Online]. Available: https://ohdsi.github.io/TheBookOfOhdsi/OhdsiCommunity.html

[9] 'OHDSI – Observational Health Data Sciences and Informatics'. Accessed: Dec. 03, 2024. [Online]. Available: https://www.ohdsi.org/

[10] Observational Health Data Sciences and Informatics, 'Who We Are – OHDSI'. Accessed: Dec. 04, 2024. [Online]. Available: https://ohdsi.org/who-we-are/

[11] Observational Health Data Sciences and Informatics, 'Mission, Vision & Values – OHDSI'. Accessed: Dec. 04, 2024. [Online]. Available: https://ohdsi.org/who-we-are/mission-vision-values/

[12] M. Solvak *jt.*, 'Programmi RITA tegevuse 1 projekti „Masinõppe ja AI toega teenused" lõpparuanne', 2022.

[13] M. Oja *et al.*, 'Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned', *JAMIA Open*, vol. 6, no. 4, p. ooad100, Dec. 2023, doi: 10.1093/jamiaopen/ooad100.

[14] M. Schuemie, M. Suchard, and P. Ryan, *CohortMethod: New-User Cohort Method with Large Scale Propensity and Outcome Models*. (2024). [Online]. Available: https://ohdsi.github.io/CohortMethod

[15] K. López-Güell, E. Burn, M. Catala, X. Li, D. Newby, and N. Mercade-Besora, *CohortSurvival: Estimate Survival from Common Data Model Cohorts*. (2025). [Online]. Available: https://darwin-eu-dev.github.io/CohortSurvival/

[16] G. K. Aava, 'R-pakett OMOP CDM kujul andmete elukestusanalüüsiks', Bakalaureusetöö, TÜ arvutiteaduse instituut, 2024.

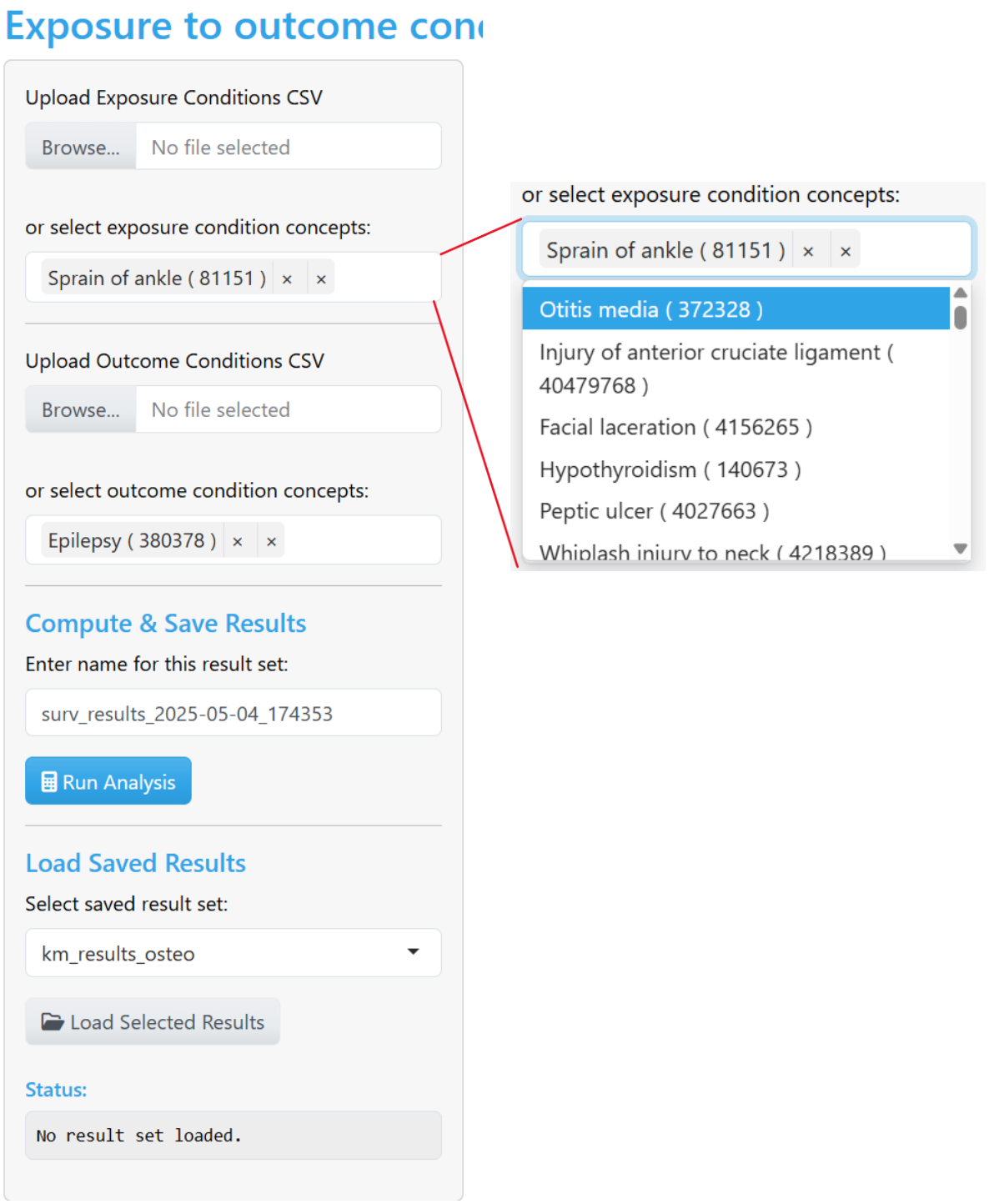
[17] D. R. T. Laursen, A. S. Paludan-Müller, and A. Hróbjartsson, 'Randomized clinical trials with run-in periods: frequency, characteristics and reporting', *Clin. Epidemiol.*, vol. Volume 11, pp. 169–184, Feb. 2019, doi: 10.2147/CLEP.S188752.

[18] O. Plana-Ripoll *et al.*, ‘Impact of washout duration to account for left truncation in register-based epidemiological studies: a case study on estimating the risk of mental disorders’, Feb. 27, 2025, *Epidemiology*. doi: 10.1101/2025.02.26.25322987.
[19] M. Iwagami and T. Shinozaki, ‘Introduction to Matching in Case-Control and Cohort Studies’, *Ann. Clin. Epidemiol.*, vol. 4, pp. 33–40, Apr. 2022, doi: 10.37737/ace.22005.
[20] J. P. Klein and M. L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data*, Second edition. New York: Springer, 2003.
[21] E. L. Kaplan and P. Meier, ‘Nonparametric Estimation from Incomplete Observations’, *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, Jun. 1958, doi: 10.1080/01621459.1958.10501452.
[22] T. A. Ryan, ‘Multiple comparison in psychological research.’, *Psychol. Bull.*, vol. 56, no. 1, pp. 26–47, 1959, doi: 10.1037/h0042478.
[23] S. Holm, ‘A Simple Sequentially Rejective Multiple Test Procedure’, *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979, Accessed: Apr. 24, 2025. [Online]. Available: http://www.jstor.org/stable/4615733
[24] M. Aickin and H. Gensler, ‘Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods.’, *Am. J. Public Health*, vol. 86, no. 5, pp. 726–728, May 1996, doi: 10.2105/ajph.86.5.726.
[25] Google, *Gemini 2.5 Pro (experimental)*. (2025). [Online]. Available: https://gemini.google.com
[26] Google, *Gemini 2.0 Flash*. (2025). [Online]. Available: https://gemini.google.com
[27] R Core Team, *R: A Language and Environment for Statistical Computing*. (2024). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/
[28] M. Schuemie and M. Suchard, *DatabaseConnector: Connecting to Various Database Platforms*. (2025). [Online]. Available: https://ohdsi.github.io/DatabaseConnector/
[29] A. Black, A. Gorbachev, E. Burn, and M. C. Sabate, *CDMConnector: Connect to an OMOP Common Data Model*. (2025). [Online]. Available: https://darwin-eu.github.io/CDMConnector/
[30] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan, *dplyr: A Grammar of Data Manipulation*. (2023). [Online]. Available: https://dplyr.tidyverse.org
[31] H. Wickham, M. Girlich, and E. Ruiz, *dbplyr: A ‘dplyr’ Back End for Databases*. (2024). [Online]. Available: https://dbplyr.tidyverse.org/
[32] T. M. Therneau, *A Package for Survival Analysis in R*. (2024). [Online]. Available: https://CRAN.R-project.org/package=survival
[33] A. Kassambara, M. Kosinski, and P. Biecek, *survminer: Drawing Survival Curves using ‘ggplot2’*. (2024). [Online]. Available: https://rpkgs.datanovia.com/survminer/index.html
[34] W. Chang *et al.*, *shiny: Web Application Framework for R*. (2024). [Online]. Available: https://shiny.posit.co/
[35] R Special Interest Group on Databases (R-SIG-DB), H. Wickham, and K. Müller, *DBI: R Database Interface*. (2024). [Online]. Available: https://dbi.r-dbi.org
[36] H. Mühleisen and M. Raasveldt, *duckdb: DBI Package for the DuckDB Database Management System*. (2025). [Online]. Available: https://r.duckdb.org/
[37] H. Wickham, ‘testthat: Get Started with Testing’, *R J.*, vol. 3, pp. 5–10, 2011, [Online]. Available: https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf
[38] Tervise Arengu Instituut, ‘RHK ehk rahvusvaheline haiguste klassifikatsioon | Tervise Arengu Instituut’. Vaadatud: 15.04.2025. Saadaval: https://tai.ee/et/instituudist/meditsiiniterminoloogia-kompetentsikeskus/who-klassifikaatorid-rfk-ja-rhk/rhk-ehk

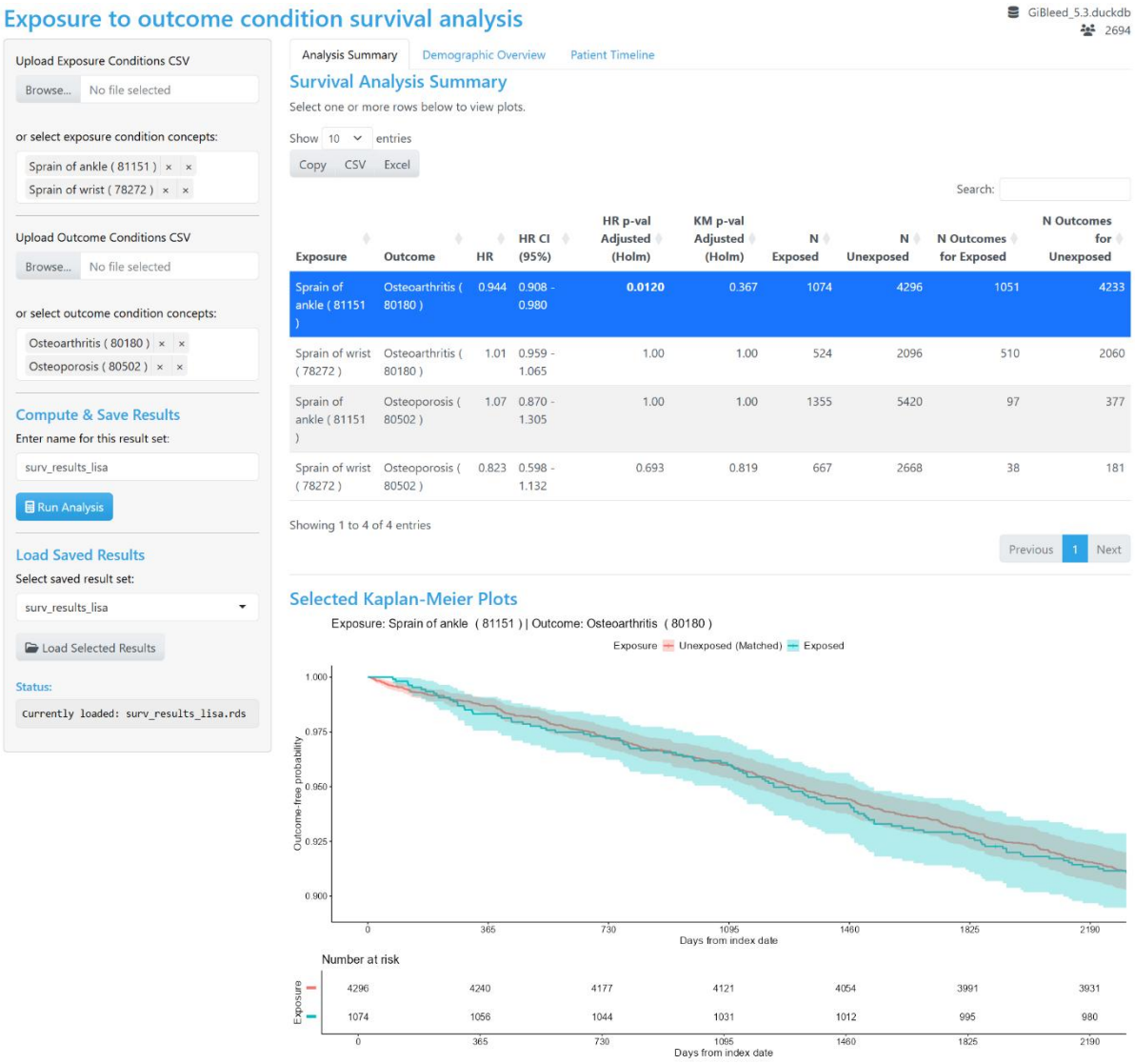
[39] Sotsiaalministeerium, ‘RHK: päringute sooritamine’. Vaadatud: 15.04.2025. Saadaval: https://rhk.sm.ee/

Lisad

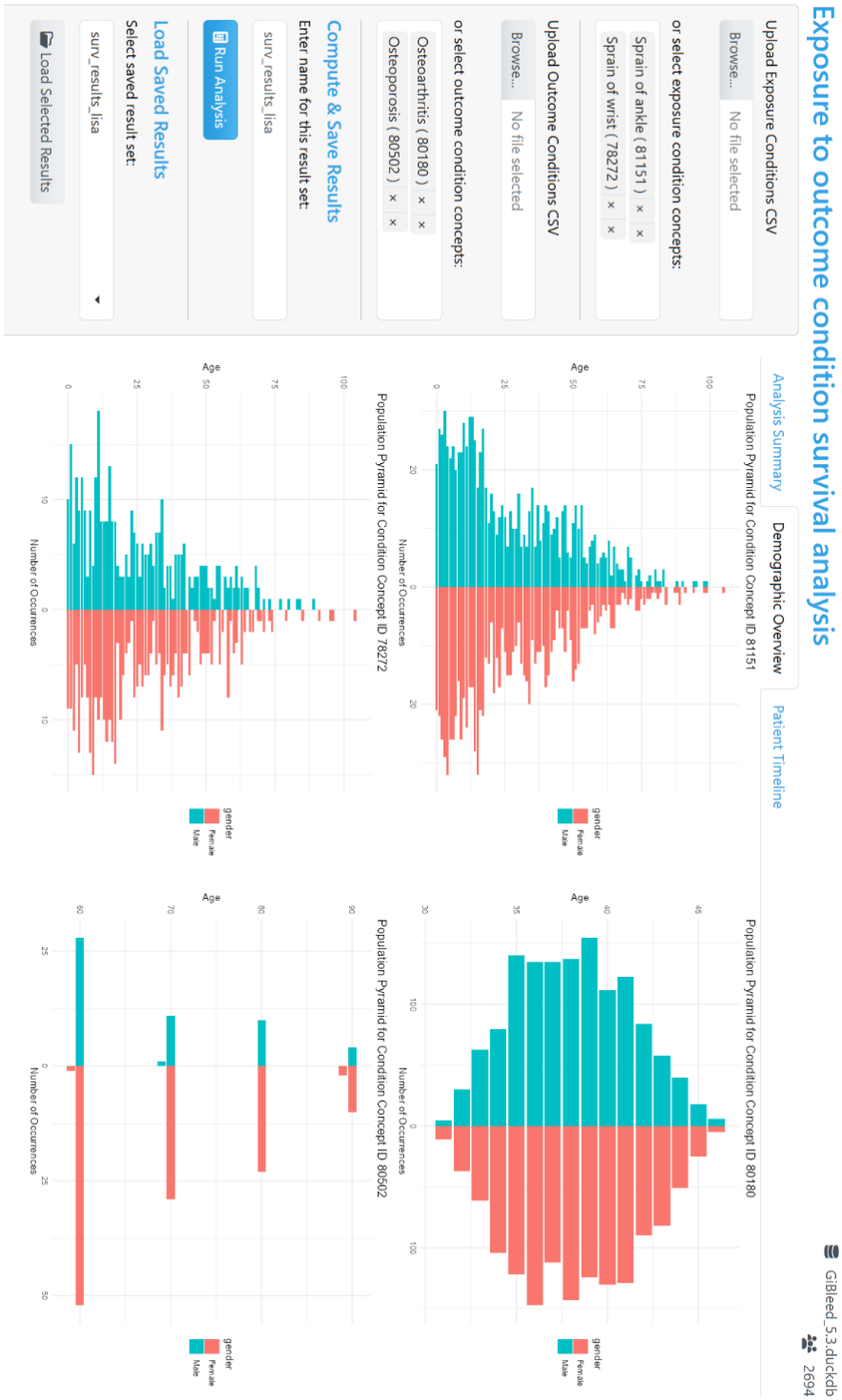
I. Paketi *Exposure2OutcomeSurv* kasutajaliidese külgriba



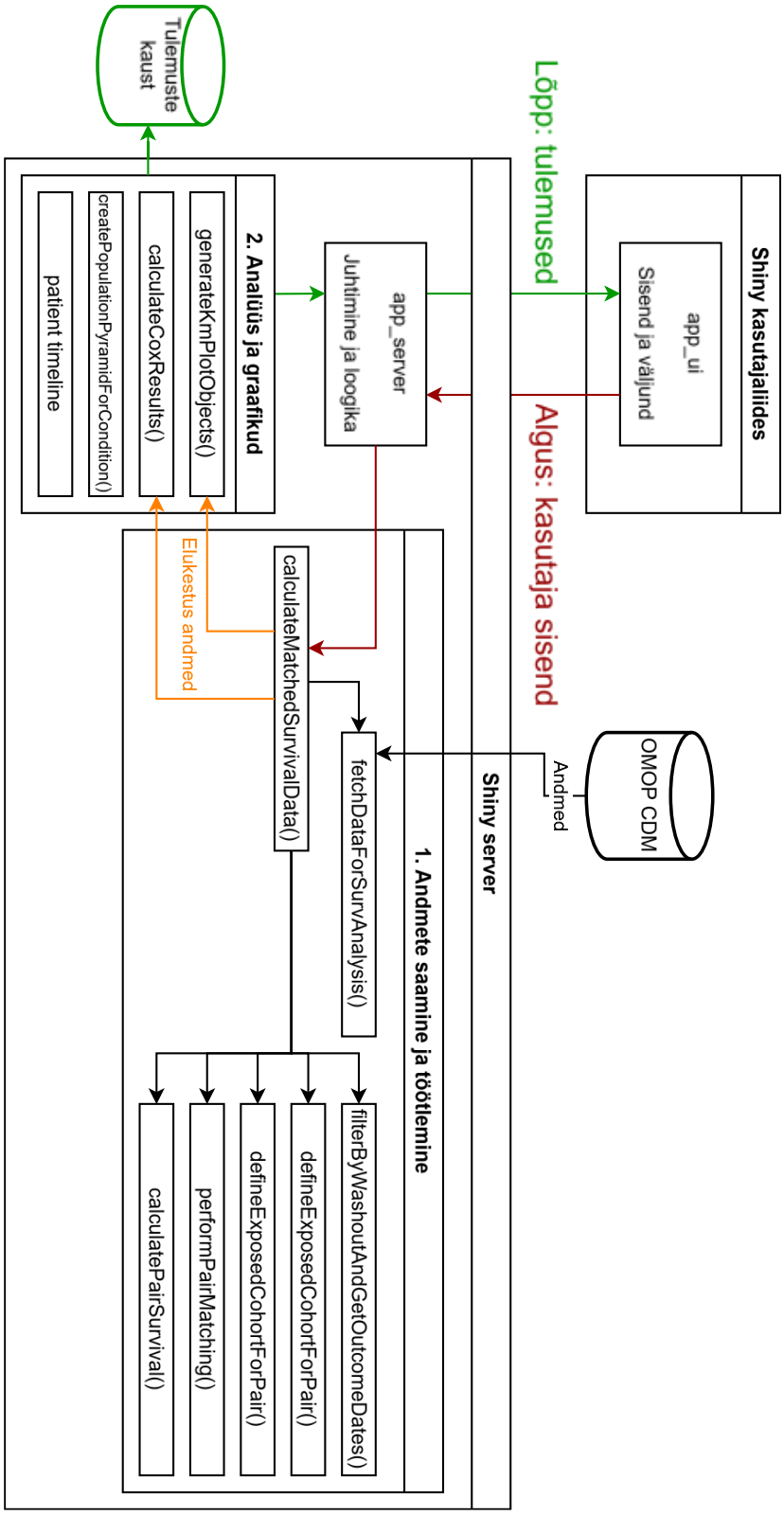
II. Paketi *Exposure2OutcomeSurv* kasutajaliidese analüüsi kokkuvõtte vaheleht



III. Paketi *Exposure2OutcomeSurv* kasutajaliidese demograafilise ülevaate vaheleht



IV. Paketi *Exposure2OutcomeSurv* töövoo ja ülesehituse joonis



V. Näidisuuringu täpsustamata terviseseisundite sisendfaili sisu

```
condition_source_concept_id,condition_source_value,concept_name,count
45568113,R10.4,Other and unspecified abdominal pain,38636
45558474,R51,Headache,35494
45537945,E03.9,"Hypothyroidism, unspecified",33960
45568132,R42,Dizziness and giddiness,25442
45543180,I49.4,Other and unspecified premature depolarization,24802
45568139,R52.2,Other chronic pain,19999
45548502,M79.2,"Neuralgia and neuritis, unspecified",17208
45587505,R29.8,Other and unspecified symptoms and signs involving the
nervous and musculoskeletal systems,16510
45562339,I20.9,"Angina pectoris, unspecified",14378
45567145,H81.9,"Disorder of vestibular function, unspecified",12885
45553741,R53,Malaise and fatigue,10799
45552888,J39.3,"Upper respiratory tract hypersensitivity reaction, site
unspecified",10435
45572718,N50.9,"Disorder of male genital organs, unspecified",9734
45558450,R07.4,"Chest pain, unspecified",9290
45568141,R55,Syncope and collapse,8967
45755625,R03,"Abnormal blood-pressure reading, without diagnosis",8771
45577526,N39.9,"Disorder of urinary system, unspecified",6112
45573006,R06.5,Mouth breathing,6081
45597166,R07.3,Other chest pain,5829
45591043,E28.9,"Ovarian dysfunction, unspecified",5108
45548974,R52,"Pain, not elsewhere classified",4433
45553726,R31,Unspecified haematuria,4207
45606790,R00.0,"Tachycardia, unspecified",4031
45566772,F03,Unspecified dementia,4030
45547996,H91.9,"Hearing loss, unspecified",3606
45586131,E05.9,"Thyrotoxicosis, unspecified",3447
45592416,R30.0,Dysuria,3229
45591632,L25.9,"Unspecified contact dermatitis, unspecified cause",2836
45558446,R00.2,Palpitations,1998
45534459,R59.0,Localized enlarged lymph nodes,1966
```

VI. Näidisuuringus uuritud tulemite sisendfaili sisu

```
Id,Code,Name,Class
45600487,C32.0,Malignant neoplasm: Glottis,ICD10 code
45556972,C32.3,Malignant neoplasm: Laryngeal cartilage,ICD10 code
45595644,C32.9,"Malignant neoplasm: Larynx, unspecified",ICD10 code
45595642,C32,Malignant neoplasm of larynx,ICD10 Hierarchy
45595643,C32.8,Malignant neoplasm: Overlapping lesion of larynx,ICD10 code
45566581,C32.2,Malignant neoplasm: Subglottis,ICD10 code
45561793,C32.1,Malignant neoplasm: Supraglottis,ICD10 code
45585989,C33,Malignant neoplasm of trachea,ICD10 Hierarchy
45542592,C34.9,"Malignant neoplasm: Bronchus or lung, unspecified",ICD10 code
45605263,C34.3,"Malignant neoplasm: Lower lobe, bronchus or lung",ICD10 code
45552250,C34.0,Malignant neoplasm: Main bronchus,ICD10 code
45595647,C34.2,"Malignant neoplasm: Middle lobe, bronchus or lung",ICD10 code
45595645,C34,Malignant neoplasm of bronchus and lung,ICD10 Hierarchy
45542590,C34.8,Malignant neoplasm: Overlapping lesion of bronchus and
lung,ICD10 code
45595646,C34.1,"Malignant neoplasm: Upper lobe, bronchus or lung",ICD10 code
45755317,C38,"Malignant neoplasm of heart, mediastinum and pleura",ICD10
Hierarchy
45595648,C38.8,"Malignant neoplasm: Overlapping lesion of heart, mediastinum
and pleura",ICD10 code
45605265,C38.4,Malignant neoplasm: Pleura,ICD10 code
45605266,C39,Malignant neoplasm of other and ill-defined sites in the
respiratory system and intrathoracic organs,ICD10 Hierarchy
45556974,C39.8,Malignant neoplasm: Overlapping lesion of respiratory and
intrathoracic organs,ICD10 code
```

VII. Kaasapandud arhiivifaili sisu kirjeldus

Tööga kaasapandud arhiivi failis „tulemused.csv“ on kogu töös läbi viidud näidisuuringu tulemuste tabel. Failis on elukestusanalüüsi tulemused 29 sisestatud ekspositsiooni (*exposure*) ja 18 tulemi (*outccome*) kohta ehk kokku 522 diagnoosikombinatsiooni kohta. NA ehk tühjad väljad on tingitud sellest, kui ühes võrreldavas grupis ei esinenud kellelgi tulemit ning Coxi võrdeliste riskide mudelit ei saanud rakendada. Kui mõlemas grupis ei esinenud kellelgi tulemit, siis on tühjad ka Kaplan-Meieri (*km*) väljad.

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Sander Soodla ,
(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   R-pakett diagnooside vaheliste seoste uurimiseks elukestusanalüüsiga ,
   (lõputöö pealkiri)

   mille juhendajad on Neeme Ilves, Maria Malk, Raivo Kolde ,
   (juhendaja nimi)

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi 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.

Sander Soodla
15.05.2025