TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Karin Rosenberg

R-pakett meditsiiniliste kohortide kirjeldamiseks

Bakalaureusetöö (9 EAP)

Juhendaja:
Kerli Mooses, PhD

Tartu 2025

R-pakett meditsiiniliste kohortide kirjeldamiseks

Lühikokkuvõte:

Standardiseeritud terviseandmete kasutamine võimaldab senisest paremini mõista patsientide ravi-, diagnoosi- ja teenusekasutuse mustreid ning toetada tervishoius andmepõhist otsustamist. Lõputöö eesmärk oli arendada töövoog ja interaktiivne rakendus, mis võimaldab kirjeldada kohorte OMOP CDM formaadis esitatud terviseandmetel rahvusvaheliste haiguste klassifikatsiooni koodide kaudu. Usaldusväärne kohortanalüüs eeldab täpselt määratletud ja selgelt kirjeldatud kohorti, mille saavutamisele aitab kaasa põhjalik ülevaade kohordi koosseisust ja omadustest — seda toetab arendatud tööriista kasutamine. Töös loodud paketi CohortExplorerICD abil saab määratleda esmaseid ja kaasuvaid diagnoose rahvusvahelise haiguste klassifikatsiooni alusel, rakendada filtreid soo, vanuse ja ajavahemike lõikes ning analüüsida valitud kohorti kuuluvate patsientide statistilisi näitajaid. Rakenduses on arvestatud andmekaitse- ja ligipääsetavuse põhimõtetega: tulemusi ei kuvata, kui andmehulgas on viis või vähem inimest, ning visualiseerimisel on kasutatud värviskeeme, mis sobivad ka värvipimedatele kasutajatele. Valminud tööriist võimaldab kasutajatel uurida valitud patsientide rühma omadusi interaktiivsete jooniste kaudu, pakkudes visuaale diagnooside, vanusejaotuse, suremuse ja meditsiiniteenuste kasutuse kohta.

Võtmesõnad: Programmeerimiskeel R, shiny, kohordi kirjeldamine, OMOP CDM, visualiseerimine

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

R package for describing medical cohorts

Abstract:

The use of standardized health data enables a better understanding of patient treatment, diagnosis, and service utilization patterns and supports data-driven decision-making in healthcare. The aim of this thesis was to develop a workflow and an interactive application for describing cohorts based on health data presented in the OMOP CDM format, with a focus on characterizing cohorts through ICD-10 codes. Reliable cohort analysis requires a precisely defined and clearly described cohort, which is facilitated by providing a comprehensive overview of the cohort's characteristics — supported by the use of the developed tool CohortExplorerICD. Using the application, it is possible to define primary and comorbid diagnoses based on the International Classification of Diseases, apply filters by gender, age, and time intervals, and analyze statistical indicators of patients belonging to the selected cohort. The application adheres to data protection and accessibility principles: results are not displayed when the data set includes five or fewer individuals and visualization uses color schemes suitable for color-blind users. The developed tool enables users to explore the characteristics of the selected patient group through interactive charts, providing visualizations related to diagnoses, age distribution, mortality, and service utilization.

Keywords: R programming language, shiny, cohort description, OMOP CDM, visualization

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

Sisukord

Sissejuhatus

Raghupathi jt [1] on rõhutanud, et terviseandmete töötlemise tõhusus ja kvaliteet on muutumas üheks olulisemaks teguriks andmepõhise tervishoiu arendamisel. Autorite sõnul loob suurenev andmemaht — sealhulgas elektroonilised terviselood, retseptiinfo, kindlustusandmed ja teenuseosutajate kirjed — mitmeid uusi võimalusi nii teaduslikuks analüüsiks kui ka praktiliseks rakendamiseks tervishoius. Terviseandmete kasutamine aitab tuvastada ravimustreid, hinnata kliinilisi tulemusi ja toetada otsuste tegemist tervishoiusüsteemi eri tasanditel. Selleks, et analüüsi oleks võimalik usaldusväärselt ja korratavalt läbi viia, tuleb tagada andmete hea struktuur, standardiseeritus ning turvaline käitlemine [2].

Kohortanalüüsi kasutatakse laialdaselt terviseandmete uurimisel, et mõista erinevate inimrühmade omadusi ja käitumist ajas. Seda meetodit rakendatakse nii epidemioloogilistes uuringutes, tervishoiuteenuste planeerimisel kui ka andmepõhiste otsuste tegemisel [3]. Käesolev lõputöö keskendub Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) andmemudelis esitatud terviseandmete kohortanalüüsile. Töö teema valik tulenes praktilisest vajadusest arendada tööriist, mis võimaldaks kohordi kirjeldamist OMOP CDM andmetel rahvusvahelise haiguste klassifikatsiooni kümnenda versiooni ehk RHK-10 koodide kaudu. Vastava vajaduse mugava ja visuaalselt arusaadava tööriista järgi tõi esile juhendaja, kellel on varasem kogemus Eesti terviseandmete analüüsiga.

Lõputöö eesmärgiks on arendada töövoog ja interaktiivne rakendus, mis võimaldab detailselt loodud kohorte kirjeldada ja visualiseerida. Rakendus võimaldab määrata esmaseid ja kaasuvaid diagnoose, rakendada filtreid soo, vanuse ja ajavahemike lõikes ning vaadelda valitud kohordi statistilisi näitajaid. Rakenduse loomisel on arvestatud oluliste andmekaitse ja ligipääsetavuse põhimõtetega: tulemusi ei kuvata, kui valitud andmehulk sisaldab vähem kui viite inimest, ning graafikutes kasutatakse värviskeeme, mis sobivad ka värvipimedatele kasutajatele. Tööriista abil saadud tulemusi saab kasutada edasiseks uurimiseks või täiendavaks andmeanalüüsiks.

Lõputöö koosneb kolmest peatükist. Esimeses peatükis tutvustatakse OMOP CDM raamistikku ja seda arendanud organisatsiooni, RHK-10 diagnoosikoode, kohortanalüüsi metoodilist tausta, juba olemasolevaid lahendusi ning kasutatud andmestikku. Teises peatükis kirjeldatakse loodud töövoo tehnilist poolt, sealhulgas andmete laadimist, töötlemist ja visualiseerimist. Kolmandas peatükis vaadeldakse kasutusstsenaariumi näitel rakenduse praktilist toimimist, analüüsitakse rakenduse eellaadimise jõudlust ning pakutakse edasisi arendussuundi. Lisadest on leitav litsents ning GitHubi link loodud rakendusele.

1. Põhimõisted ja taustinformatsioon

Selles peatükis antakse ülevaade olulisematest mõistetest, lühenditest ja taustinfost, mis on vajalikud ülejäänud töö paremaks mõistmiseks. Peatükk jaguneb viieks alapeatükiks. Esmalt tutvustatakse OHDSI organisatsiooni ja selgitatakse OMOP andmemudeli olemust. Seejärel tutvustatakse rahvusvahelisest haiguste klassifikatsiooni RHK-10 ning selle rolli terviseandmete tähenduse mõistmisel. Kolmandas osas käsitletakse kohortanalüüsi põhimõtteid ning põhjendatakse selle kasutamist terviseandmete analüüsimisel. Neljandas peatükis tutvustatakse juba olemasolevaid sarnaseid lahendusi ja tuuakse välja erinevused siinse töö raames loodud rakendusega. Viimases alapeatükis tutvustatakse kasutatud andmestikku ja täpsustatakse, milliseid tabeleid kasutatakse antud töö raames loodud tööriista juures.

1.1 OHDSI organisatsioon ja OMOP CDM andmemudel

Observational Health Data Sciences and Informatics (OHDSI) on rahvusvaheline teadus- ja arendusvõrgustik, mille eesmärk on arendada avatud teaduspõhiseid meetodeid ja tööriistu terviseandmete standardiseeritud analüüsiks. Kodulehe järgi on organisatsioon loodud 2014. aastal ning ühendab üle maailma tuhandeid teadlasi, arendajaid ja kliinikuid. Organisatsioon toetub kogukonnapõhisele arendusmudelile ja avatud lähtekoodiga tarkvarale. Tegevuse keskmes on andmete võrreldavus ja reprodutseeritavus — terviseandmed, mis pärinevad eri riikidest ja süsteemidest, viiakse ühtsesse struktuuri, võimaldades ühesuguseid analüüse erinevates kontekstides. See toetab tõenduspõhise meditsiini arendamist, suurandmete kasutamist teadusuuringutes ning teadustulemuste usaldusväärsust. [4]

Euroopa tasandil koordineerib tegevusi OHDSI Europe, mis toimib platvormina akadeemiliste ja kliiniliste asutuste, rahvatervise organisatsioonide ja tööstuspartnerite vahel [5]. Eesti Arstis avaldatud artiklis [6] kirjeldatakse Eesti OHDSI kogukonna kujunemist ja tegevust, mille keskmes on Tartu Ülikooli arvutiteaduse instituudi terviseinformaatika töörühm. Töögrupi tegevuse käigus on loodud ja testitud metoodika, mille abil viidi valik Eesti terviseandmeid OMOP CDM ehk Observational Medical Outcomes Partnership Common Data Model formaati. Esimesed katsetused tehti geenivaramu küsimustiku vastuste põhjal 2017. aastal. Artikli autorite sõnul loob see võimaluse kasutada Eesti andmeid rahvusvahelistes võrreldavates uuringutes ja suurandmete analüüsides, tugevdades seeläbi Eesti osalust OHDSI võrgustikus. Eesti OHDSI kogukond on esindatud ka sümpoosionitel ja rahvusvahelistes aruteludes [7]. Käesolev töö põhineb samuti avatusel ja reprodutseeritavusel ning on osa OHDSI kommuuni vabavarast, seetõttu on nii rakenduse kood kui ka kasutajaliides inglise keeles.

Järgnev tekst tugineb OMOP CDM ametlikule dokumentatsioonile [8]. OMOP CDM on tabelipõhine andmemudel, mille eesmärk on võimaldada erinevatest allikatest, näiteks haiglate infosüsteemidest, apteegi- ja retseptiregistritest, ravikindlustuse andmetest või muudest terviseandmebaasidest, pärit info talletamist standardiseeritud ja analüüsivalmis kujul. Mudeli töötas välja OHDSI kogukond, et toetada skaleeritavat, korratavat ja rahvusvaheliselt võrreldavat andmepõhist teadustööd.

OMOP CDM uusimas versioonis 5.4 on kokku 39 tabelit, mis jagunevad kuude põhikategooriasse:

1. Standardiseeritud kliinilised andmed – näiteks `person`, `visit_occurrence`, `condition_occurrence`, `drug_exposure`, `procedure_occurrence`, `observation`, `measurement`, `death`;
2. Standardiseeritud sõnastikud (ingl *vocabularies*) – näiteks `concept`, `concept_relationship`, `concept_ancestor`, `vocabulary`;
3. Standardiseeritud metaandmed ja administreerivad struktuurid – nt `care_site`, `provider`, `location`, `metadata`;
4. Analüüsi töövoogu toetavad tabelid – sh `cohort`, `cohort_definition`, `episode`;
5. Kliinilised lisaandmed – nt `note`, `device_exposure`, `specimen`, `cost`;
6. Kaardistamise ja toetavad tabelid – nt `source_to_concept_map`. Kõik tabelid on seotud patsiendi unikaalse identifikaatori (`person_id`) kaudu, mis võimaldab seostada erinevates domeenides esitatud andmed ühe indiviidi kaupa ning luua ajaliselt järjestatavaid sündmusahelaid. Selles töös kasutatud tabelid ja nende väljad on kirjeldatud täpsemalt peatükis (vt peatükk 1.5).

Üheks andmemudeli peamiseks eeliseks on see, et kõik kliinilised mõisted on esitatud standardiseeritud kontseptsioonidena, mis on seotud rahvusvaheliste terminoloogiatega, nagu SNOMED CT (diagnoosid ja seisundid), RxNorm (ravimid), LOINC (laboritestid ja mõõtmised) ning paljud teised. Standardiseeritud kontseptsioonid võimaldavad eri andmeallikate, riikide ja keelte vahel sisulist ühtlustamist ning võrreldavate analüüside läbiviimist. Samal ajal võimaldab seos algsete koodidega (ingl *source values*) säilitada vastavuse algandmetega, mis on oluline kohapealsete analüüside ja valideerimise jaoks.

Standardiseeritud andmestruktuur loob aluse ka ühtsete analüüsivahendite arendamiseks ja taaskasutamiseks. OHDSI kogukonnas on selleks loodud mitmeid rakendusi, millest käesoleva töö kontekstis on kõige olulisem ATLAS[1] – veebipõhine kasutajaliides, mis võimaldab luua kontseptsioonikomplekte, defineerida kohorte, koostada päringuid ja sooritada esmaseid andmeanalüüse. ATLASe abil saavad kasutajad määratleda uuritava populatsiooni, valida ajavahemikud (nt enne või pärast kohorti sisenemist), seada filtrid vanuse, soo ja diagnooside alusel ning genereerida automatiseeritud analüüsiskripte, mida saab jooksutada lokaalses andmebaasis. Tööriist võimaldab visuaalselt hallata keerukaid päringuid ja toetab erinevaid analüüsitüüpe, nagu kohordi kirjeldamine, riskiarvutused, seoseanalüüsid ja prediktsioonimudelid. ATLAS on antud töö kontekstis vajalik kohortide defineerimiseks.

OMOP CDM võimaldab läbi viia hajusanalüüsi – andmeid ei ole vaja asutusest välja saata. See lähenemine on oluline eeskätt rahvusvaheliste uuringute puhul, kus tagatakse andmekaitse ja privaatsus, kuid tulemused on siiski koondatavad ja võrreldavad. OMOP CDM-i rakendamine Eestis, sh MAITT projekti raames (vt peatükk 1.5), on näidanud, et Eesti andmestikud sobivad rahvusvaheliste standardite kasutuselevõtuks ja suudavad toetada kaasaegset teaduspõhist terviseandmete analüüsi. Käesolevas töös kasutatud andmestik ja selle põhjal loodud rakendus tuginevad OMOP CDM andmemudelile.

1.2 RHK-diagnoosikoodid

Rahvusvaheline haiguste klassifikatsiooni kümnes versioon ehk RHK-10 (ingl *ICD-10 – International Classification of Diseases, 10th Revision*) on Maailma Terviseorganisatsiooni (WHO) loodud standardne klassifikatsioonisüsteem, mida kasutatakse üle maailma haiguste, sümptomite, vigastuste ja muude terviseprobleemide süstematiseeritud kodeerimiseks [9]. Tegemist on laialdaselt levinud süsteemiga, mille kaudu dokumenteeritakse diagnoosid elektroonilistes terviseandmetes, sealhulgas Eestis.

RHK-10 koodid võimaldavad arstidel, teadlastel ja andmeanalüütikutel kirjeldada patsientide seisundeid struktureeritud ja rahvusvaheliselt võrreldaval viisil. Süsteem [9] on jaotatud peatükkideks vastavalt haiguste tüüpidele või kehasüsteemile (nt südame-veresoonkonna haigused, hingamisteede haigused, psüühikahäired). Iga peatükk sisaldab alamkategooriaid ning vastavaid koodide vahemikke.

[1] https://atlas-demo.ohdsi.org/

Näiteks:

- **I10–I15**: Hüpertensioon
- **I20–I25**: Isheemiline südamehaigus
- **E10–E14**: Diabeet
- **F32–F33**: Depressioon
- **C00–C97**: Pahaloomulised kasvajad

Käesolevas lõputöös kasutatakse RHK-10 koode patsientide kaasuvate haiguste tuvastamiseks, analüüsides nende olemasolu enne ja pärast kohorti sisenemist. Näiteks Khan jt artikli põhjal [10] on südamepuudulikkusega patsientide kohorti uurides asjakohane hinnata, kui paljudel neist esines eelnevalt hüpertensiooni (I10–I15), isheemilist südamehaigust (I20–I25) või diabeeti (E10–E14), kuna need seisundid võivad mõjutada ravi kulgu ja prognoosi.

RHK-10 on ametlikult kasutusel Eesti tervishoiusüsteemis ja on aluseks meditsiiniliste dokumentide koostamisel, sh retseptid, epikriisid ja haiguslugude kokkuvõtted [11]. Süsteem on intuitiivne ja arusaadav nii arstidele, andmeteadlastele kui ka otsustajatele; see toetab analüüsitulemuste tõlgendatavust ja rakendamist tervishoiuotsustes.

1.3 Kohortanalüüs terviseandmetes

Järgnev selgitus põhineb Aaron Kandola artiklil [3]. Kohortanalüüs on üks enimkasutatavaid uurimismeetodeid epidemioloogias ja terviseandmete analüüsis. Selle keskmes on kindlaks määratud isikurühm ehk kohort, mille liikmeid ühendab konkreetne sündmus või tunnus, näiteks haiguse esmane diagnoos, ravi alustamine või tervishoiuteenuse kasutamine. Kohordi liikmeid jälgitakse ajas, et uurida hilisemaid sündmusi, näiteks diagnooside lisandumist, tüsistusi, hospitaliseerimisi või surmajuhtumeid. Kohortuuringute eesmärk on tuvastada seoseid varasemate seisundite, riskitegurite ja hilisemate kliiniliste tulemite vahel. Erinevalt kontrollitud uuringutest toimuvad kohortuuringud päriselulises keskkonnas, ilma teadlikult määratud sekkumiseta.

Järgnevas kahes lõigus on välja toodud kohortanalüüsi eelised ja puudused. Selgitused põhinevad raamatul *Cohort Studies in Health Sciences*, mille autorid on Samer Hammoudeh, Wessam Gadelhaq ja Ibrahim Janahi [12]. Kohortuuringud on eriti väärtuslikud olukordades, kus eksperimentaalne lähenemine ei ole eetiline või teostatav. Näiteks ei saa inimesi juhuslikult määrata haigust põdema, kuid on võimalik jälgida isikuid, kellel on juba kindel diagnoos,

ning analüüsida nende edasist tervisekäiku. Kuna kohortuuringud jälgivad inimesi ajas edasi alates mingist sündmusest (nt esmane diagnoos) kuni mingisuguse tulemuse kujunemiseni, võimaldavad need hinnata ajarelatsiooni ning arvutada olulisi epidemioloogilisi näitajaid, nagu haigestumuskordaja ja suhteline risk. Oluline eelis seisneb ka selles, et kohortuuringutes saab uurida mitut erinevat tervisega seotud tulemit, mis võivad ilmneda pärast konkreetset sündmust või seisundit. Näiteks saab jälgida, millised kaasuvad haigused tekivad isikutel pärast teatud diagnoosi (nt südamepuudulikkus, diabeet või vähk). Samuti sobivad need uuringud hästi harva esinevate seisundite või sündmuste uurimiseks, kuna osalejad valitakse sageli just konkreetse kliinilise tunnuse alusel.

Samas on kohortuuringutel mitmeid metoodilisi piiranguid, mida tuleb tulemuste tõlgendamisel hoolikalt arvesse võtta. Esiteks ei võimalda see uuringutüüp põhjuslike seoste lõplikku tõestamist, kuna tegemist on vaatlusuuringuga, milles uuritavad ei ole juhuslikult jaotatud rühmadesse erineva kokkupuute (nt haiguse, ravi või sekkumise) alusel. Põhjuslikkuse hindamine eeldab aga eksperimentaalset lähenemist, mida kohortuuringud oma olemuselt ei paku. Paljud tervisega seotud tulemused, näiteks kroonilised haigused, kujunevad välja pika aja jooksul. Seetõttu vajavad kohortuuringud sageli pikki jälgimisperioode ja ulatuslikke valimeid, mis muudab need ajamahukaks, keerukaks ja kulukaks võrreldes paljude teiste uuringudisainidega. Samuti võivad kohortuuringud harva esinevate tulemuste uurimiseks vajada väga suurt osalejate hulka. Oluline metoodiline probleem on segavate tegurite olemasolu – muutujad, mis on seotud nii algse seisundi kui ka hilisema tulemusega, kuid ei ole osa põhjuslikust ahelast. Kui neid tegureid ei tuvastata ega kontrollita piisavalt, võivad tulemused olla eksitavad.

Kohortuuringu üks määravaid etappe on täpne kohordi defineerimine. Valesti või ebatäpselt määratletud kohort võib viia eksitavate järeldusteni, mistõttu on selle etapi kvaliteet kogu uuringu usaldusväärsuse seisukohalt kriitilise tähtsusega. Viimastel aastatel on kohordi määratlemise kvaliteedi ja kiiruse parandamiseks hakatud laialdaselt rakendama *study-a-thon* meetodit – koostööformaati, kus mitmekesise taustaga eksperdid (nt arstid, andmeanalüütikud, andmeteadlased jt) kogunevad ühise eesmärgiga defineerida uuritavad kohordid [13]. Eesti meditsiinisüsteemis kasutatakse haiguste kodeerimisel rahvusvahelist RHK-10 klassifikatsiooni, sellepärast on arstide analüütilised küsimused just RHK-10 kesksed. Seetõttu tekkis Tartu Ülikooli terviseinformaatika uurimisrühmas vajadus tööriista järele, mis võimaldaks defineeritud kohorti RHK-10 põhiselt analüüsida.

1.4 Varasemad lahendused

Antud alapeatükis tutvustatakse kahte OHDSI arendatud tööriista, mis on loodud samuti kohortide kirjeldamiseks nagu töös loodud CohortExplorerICD. Alguses tutvustatakse põgusalt kahte paketti ja lõpuks tuuakse välja peamised erinevused loodud rakendusega.

1.4.1 R-pakett CohortDiagnostics

Järgnev kirjeldus põhineb paketi kodulehel avaldatud teabel [14]. CohortDiagnostics on OHDSI kogukonna poolt loodud R-pakett, mille eesmärk on toetada fenotüübialgoritmide arendamist ja hindamist. Fenotüübialgoritm tähendab reeglite ja kriteeriumite kogumit, mille alusel määratakse, kas inimene kuulub uuritavasse kohorti. Sellised reeglid võivad põhineda näiteks diagnoosikoodidel, ravimiretseptidel või sooritatud protseduuridel. Pakett pakub süsteemset raamistikku kohortide omaduste ja struktuuri analüüsimiseks, võimaldades olemasolevate andmete põhjal koostada ülevaadet, mida saab uurida interaktiivse R shiny liidese kaudu. Analüüsid hõlmavad muu hulgas kohorti kaasamise reeglite mõju hindamist, kohorti kuuluvate isikute omaduste kirjeldamist ning kohortidesse viinud sündmuste analüüsimist. Samuti võimaldab tööriist tuvastada koode, mis võiksid kohordis olla, kuid pole sinna mingil põhjusel kaasatud. CohortDiagnostics võimaldab arvutada kohordi liikmete esinemissagedusi aastate, vanuse ja soo lõikes ning analüüsida kohortide kattuvust ja erinevusi. Lisaks pakub pakett võimalust uurida juhuslikult valitud kohorti kuuluvate isikute detailseid profiile. Paketi kasutamine toimub kahes etapis: esmalt arvutatakse vajalikud näitajad olemasolevate andmete põhjal, seejärel saab tulemusi analüüsida kaasasoleva shiny rakenduse abil.

1.4.2 R-pakett CohortCharacteristics

Järgnev kirjeldus tugineb paketi kodulehele [15]. Pakett CohortCharacteristics sobib hästi suuremahuliste kirjeldavate uuringute tarbeks, võimaldab tulemusi võrrelda kontroll- ja sihtkohortide vahel ning aitab analüüsida ja visualiseerida patsiendikohorte, mis on määratletud OMOP CDM andmestikus. Paketi eesmärk on anda ülevaade, millised inimesed kohorti kuuluvad – näiteks milline on nende vanus, sugu, haigused või ravimikasutus. See töötab kolmes etapis: kõigepealt kogub see andmed otse andmestikust, siis loob neist struktureeritud kokkuvõtted, lõpuks võimaldab neid tulemusi näidata tabelite või joonistena. Selleks kasutatakse kolme tüüpi funktsioone: *summarise* funktsioonid loovad kokkuvõtted, *table* funktsioonid muudavad need tabeliteks ning *plot* funktsioonid joonisteks. Kohorte saab kirjeldada mitmete parameetrite alusel. Näiteks saab näidata, milline on kohorti kuuluvate inimeste vanuseline jaotus, kui paljud neist on mehed või naised, kas nad on uuringu ajal surnud, milliseid haigusi neil on esinenud, milliseid

ravimeid nad on tarvitanud või milliseid protseduure on neile tehtud. Samuti saab vaadata, milliseid mõõtmisi või laboritulemusi neil on registreeritud ja milliseid tervishoiuteenuseid on nad kasutanud. Kõiki neid andmeid saab vaadata nii ühe kohordi kohta eraldi kui ka mitut kohorti võrreldes. See muudab paketi kasulikuks kliinilistes uuringutes, kus on vaja hinnata, kui sarnased või erinevad on vaatluse all olevad kohordid.

1.4.3 Võrdlus loodud paketiga CohortExplorerICD

Töös loodud rakendus sarnaneb suuresti eelnevalt kirjeldatud pakettidega, kuid erineb mitmel olulisel viisil ning pakub kasutajale spetsiifilisemaid võimalusi, mis standardsete tööriistadega ei ole vahetult kättesaadavad. Esiteks pakub loodud rakendus täielikult interaktiivset kasutajaliidest, kus kasutaja saab reaalajas määrata selliseid sisendparameetreid nagu vanusevahemik, sugu, esmased ja kaasuvad diagnoosid ning ajavahemik seoses kohorti sisenemise kuupäevaga. CohortCharacteristics ei sisalda kasutajaliidest.

Teiseks integreerib rakendus andmekaitselise kontrollmehhanismi: kõik tulemused, mille isikute arv on alla viie, peidetakse automaatselt või tähistatakse märgisega *Not enough people (N ≤ 5)*. See mehhanism on oluline tundlike terviseandmete analüüsil, eriti väiksemate kohortide või haruldaste seisundite puhul. OHDSI tööriistades sellist funktsionaalsust vaikimisi ei rakendata, mistõttu tuleb andmekaitselised piirangud käsitsi lisada.

Kolmandaks on oluline rõhutada, et kogu loodud rakenduse loogika ja analüüs on üles ehitatud RHK-10 koodide baasil. See võimaldab kliiniliselt tähenduslikumat ja paindlikumat diagnoosipõhist analüüsi. OHDSI pakettides kasutatakse vaikimisi SNOMED CT kontseptsioone ja kontsepthierarhiaid, mis võivad terviseandmete lokaalses kasutuses osutuda vähem intuitiivseteks, eriti kui kohalik dokumentatsioon või arstiotsused põhinevad RHK-koodidel.

1.5 MAITT andmestik

Antud lõik põhineb Oja jt artiklil [16]. MAITT on Tartu Ülikooli arvutiteaduse instituudi terviseinformaatika töörühma juhitud projekt, mille üheks eesmärgiks on Eesti riiklikest terviseandmetest koosneva, standardiseeritud ja analüüsivalmiks tehtud andmestiku loomine. Projekti fookuses on olnud andmete teisendamine OMOP CDM struktuuri. Antud töös kasutatud andmestik põhineb 10% juhuvalimil Eesti elanikkonnast, milleks on 150 824 inimest aastatest 2012–2019. Selline pikaajaline katvus võimaldab uurida haiguse kulgu, ravi mustreid ja tervishoiuteenuste kasutust kuni seitsme aastase retrospektiivse perspektiiviga. MAITT andmestik on loodud, koondades info erinevatest Eesti andmeallikatest. Kõik

andmed on isikustamata ja seotud pseudonüümsete patsiendiidentifikaatoritega, mis tagavad andmekaitse ja privaatsuse. Pseudonüümitud ID-d võimaldavad siiski andmete sisulist sidumist ja analüüsi ühe indiviidi lõikes – näiteks diagnooside, ravimite ja visiitide ajalise järjestuse alusel. MAITT andmestik loob Eestis tugeva aluse andmepõhiseks terviseuuringuks ning on kasutatav mitmesugustes uurimisvaldkondades alates kohortanalüüsist kuni trajektoorimudelite ja teenusekasutuse analüüsideni.

Käesolevas töös kasutati mitmeid OMOP CDM andmemudeli tabeleid, mille kaudu viidi läbi kohortanalüüsi jaoks vajalikud andmete laadimised, seoste tuvastamised ning filtreerimised. Allolevas tabelis (vt Tabel 1) on esitatud andmemudeli tabelid koos olulisemate väljadega, mida töö käigus kasutati. Väljad olid seotud patsientide demograafilise info, diagnooside, raviprotseduuride, külastuste, teenuseosutajate ja asutustega. Lisaks kasutati kontseptsioonitabelit analüüsiks vajalike terminite sidumiseks.

Tabel 1. OMOP CDM tabelite väljad ja nende selgitus

Tabel	**Väli**	**Selgitus**
person	person_id	Isiku identifikaator
	year_of_birth	Sünniaasta
	gender_concept_id	Sugu
cohort	subject_id	Kohordi isik
	cohort_definition_id	Kohordi määratlus
	cohort_start_date	Kohordi alguskuupäev
condition_occurrence	condition_source_value	Diagnoosikood
	condition_occurrence_id	Diagnoosi sündmuse id
	condition_start_date	Diagnoosi kuupäev
	condition_type_concept_id	Diagnoosi tüüp
	visit_occurrence_id	Seotud visiit
procedure_occurrence	procedure_date	Protseduuri kuupäev
	procedure_concept_id	Protseduuri kontseptsioon
	visit_occurrence_id	Seotud visiit
observation	observation_date	Vaatluse kuupäev
	observation_concept_id	Vaatluse kontseptsioon
	visit_occurrence_id	Seotud visiit
visit_occurrence	visit_occurrence_id	Visiidi ID
	visit_concept_id	Visiidi tüüp
	visit_source_value	Külastuse lähtekood algandmetes
	visit_start_date	Visiidi alguskuupäev
	care_site_id	Tervishoiuasutus
	provider_id	Teenuseosutaja
death	death_date	Surmakuupäev
	subject_id	Surnud isik
concept	concept_id	Meditsiiniline kontseptsioon

Kõik nimetatud tabelid on seotud patsiendi unikaalse identifikaatori `person_id` kaudu. See võimaldab andmeid ajaliselt järjestada ja integreerida sündmusteahelateks, kus on võimalik jälgida patsiendi terviseseisundit enne ja pärast teatud sündmust (nt diagnoos).

Töö autor pääses andmestikule ligi, kuna see viidi läbi vastavalt Tartu Ülikooli 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.

2. Implementatsioon

Selles peatükis kirjeldatakse täpsemalt töövoo tehnilist teostust. Rakenduse loomiseks kasutati programmeerimiskeele R versiooni 4.4.0. Kasutajaliidese loomiseks kasutati R-paketti shiny[2]. Funktsionaalsuse lisamiseks ja andmete paremaks töötlemiseks kasutatakse ka teisi R-pakette ja teeke, millest saab täpsemalt lugeda järgnevatest alapeatükkidest. Viimases alapeatükis kirjeldatakse rakenduse visuaalset ülesehitust, sealhulgas kuvatavaid vahelehti ning igal vahelehel esitatavaid andmeid, jooniseid ja tabeleid. Kõik rakenduses kuvatav info ja joonised on valitud vastavalt töö juhendaja ja autori ühisele otsusele, põhinedes sellele, mida võiks terviseinformaatiku töös kõige enam vaja minna.

2.1 Metoodika

Rakenduse arendamisel lähtuti vajadusest kirjeldada OMOP CDM andmestikul põhinevat kohorte nii, et kasutaja saab dünaamiliselt määrata erinevaid sisendparameetreid ning saada sellele vastavalt intuitiivselt mõistetavaid visualiseeringuid. Shiny võimaldab jagada rakendust kaheks komponendiks (serveriloogikaks ja kasutajaliideseks), mis tagab paindlikkust graafiliste elementide ülesehituses ja funktsioonide realiseerimises. Kasutajaliidese struktuuri loomiseks kasutatakse lisaks paketile shiny ka shinydashboard[3], shinyWidgets[4], shinyBS[5] ning laadimisindikaatorite jaoks shinycssloaders[6].

Andmebaasiga suhtlemiseks kasutatakse OHDSI kogukonna poolt välja töötatud paketti DatabaseConnector[7], mis võimaldab R-keskkonnal ühenduda PostgreSQL andmebaasiga, kus MAITT projekti OMOP CDM andmestik on salvestatud. DatabaseConnector'i eelis seisneb võimaluses kasutada mitmeid andmebaaside ühendusliideseid ning selle tugevas integreeritavuses teiste OHDSI pakettidega. Lisaks kasutatakse pakette DBI[8] ja glue[9] SQL-päringute koostamiseks ja käivitamiseks.

---

[2] shiny, https://shiny.posit.co/
[3] shinydashboard, https://rstudio.github.io/shinydashboard/
[4] shinyWidgets, https://cran.r-project.org/package=shinyWidgets
[5] shinyBS, https://cran.r-project.org/package=shinyBS
[6] shinycssloaders, https://cran.r-project.org/package=shinycssloaders
[7] DatabaseConnector, https://github.com/OHDSI/DatabaseConnector
[8] DBI, https://cran.r-project.org/package=DBI
[9] glue, https://cran.r-project.org/package=glue

Rakenduse loogika näeb ette, et esmasel käivitamisel või uue kohordi valimisel tuuakse vajalikud andmed otse andmebaasist SQL-päringute abil. SQL-päringute tulemused tuuakse R-i ja töödeldakse edasi kasutades Tidyverse'i tööriistu – dplyr[10], purrr[11]. Nii on võimalik rakenduses teostada mitmesuguseid andmetöötlusoperatsioone: liitmine (ingl *join*), filtreerimine, muutmine (ingl *mutate*), grupiviisiline kokkuvõtete tegemine (ingl *summarise*) jms. Üks rakenduse olulisi eeliseid seisneb selles, et see võimaldab enamikku filtritest rakendada lokaalselt – seega toimub SQL-päring üksnes esmasel andmete laadimisel, pärast mida rakendub ülejäänud töövoog kohapeal, see teeb kasutajaliidese kiiremaks ja interaktiivsemaks.

Visualiseerimise jaoks kasutatakse kombinatsiooni ggplot2[12] (staatilised graafikud) ja plotly[13] (interaktiivsed graafikud). Plotly võimaldab kasutajal graafikuid suurendada, andmepunkte esile tõsta ning kursori peale liikumisel väärtusi kuvada. Tabelvaadete jaoks kasutatakse DT[14] paketti, mis võimaldab kuvada andmeid interaktiivselt ja sorteeritavalt.

2.2 Kohordi määratlemine ja filtrid

Selles töös kasutatakse eeldefineeritud kohorte, mis on koostatud OHDSI ATLAS[15] tööriistaga. Kohordi määratluse tulemusena salvestatakse vastavad indiviidid OMOP CDM andmebaasis `cohort` tabelisse, kus iga kirje esindab ühte isikut ja tema kohorti sisenemise ja võimaliku väljumise kuupäeva.

Rakendus kasutab andmebaasis olevat `cohort` tabelit kui sisendi lähtekohta – kõik analüüsid tehakse ainult kohorti kuuluvate isikute andmete põhjal. Kohordi identifitseerimiseks kasutatakse `cohort_definition_id` väärtust, mis määrab, millise loogika alusel indiviid kohorti kuulus. Kasutaja peab sisestama `db_connect.R` faili sobiva `cohort_definition_id` väärtuse. Lisaks kohordi valikule on rakenduses mitmeid filtreid, mille abil saab kasutaja täpsustada, millise alamkohordi andmeid soovitakse visualiseerida. Kõik filtrid on kujundatud interaktiivsetena ning nende muutmine viib jooniste ja tabelite uuendamiseni.

---

[10] dplyr, https://cran.r-project.org/package=dplyr
[11] purrr, https://cran.r-project.org/package=purrr
[12] ggplot2, https://cran.r-project.org/package=ggplot2
[13] plotly, https://cran.r-project.org/package=plotly
[14] DT, https://cran.r-project.org/package=DT
[15] ATLAS, https://atlas-demo.ohdsi.org/

Rakenduses on vanusefilter, soo filter, diagnooside filter ja ajaraam. Vanusefiltriga saab kasutaja määrata vanusevahemiku, milles kohorti kuuluvad indiviidid analüüsi kaasatakse. Vanus arvutatakse vastavalt isiku sünniaastale (väärtus tulbast `year_of_birth`) ja kohorti sisenemise kuupäevale (`cohort_start_date`). Filtri väärtus on reaktiivne ning mõjutab kõiki jooniseid, millel vanus on asjakohane.

Soo filtriga saab kasutaja valida, kas soovitakse analüüsida ainult mehi, ainult naisi, mõlemaid koos eristuseta ning valitud joonistel ka eristusega.

Kohordi liikmed võivad omada mitmeid diagnoose. Diagnoosi filtriga saab kasutaja määrata, millised diagnoosid loetakse peamisteks ning millised kaasuvateks. RHK-10 koodide põhjal koostatud valikukastides kuvatakse ainult need koodid, mis tegelikult selles kohordis esinevad. Kasutaja saab valida mitu diagnoosikoodi või kasutada *Select all / Deselect all* funktsiooni. See filter mõjutab kõiki diagnoosiga seotud analüüse (nt haiguste esinemissagedused, visiidid jne).

Paljude jooniste puhul (nt visiidid, surmad) saab kasutaja määrata ajavahemiku kohorti sisenemisest – näiteks *12 months before* kuni *12 months after*. Liugur võimaldab määrata täpset kuudepõhist akent (-12 kuni +12 kuud). Filtri abil välistatakse kõik sündmused, mis jäävad sellest ajavahemikust välja.

Rakenduses on sisse ehitatud kontrollmehhanism, mis takistab liiga väikese N (<= 5) andmete kuvamist. Iga joonise või tabeli andmetöötluse lõpus rakendatakse kontroll, mis filtreerib välja tulemused, kus arvulised väärtused jäävad alla viie. Sellistel juhtudel kuvatakse kasutajale teade: *Not enough people ($N \leq 5$)*.

2.3 Andmete laadimine ja töötlemine

Rakenduse tööprotsess jaguneb kaheks etapiks: andmete eellaadimine ning interaktiivne töötlemine kasutajaliideses. Selle ülesehituse eesmärk on vähendada koormust andmebaasile ning tagada interaktiivne kasutajakogemus.

Kohe rakenduse käivitamisel luuakse ühendus PostgreSQL andmebaasiga, kasutades R-paketti `DatabaseConnector`. Selle kaudu laetakse mällu kõik kohordiga seotud andmed, mis on vajalikud rakenduse erinevate komponentide toimimiseks. Tegevus toimub `db_connect.R` failis, kus määratakse ära:

`source_schema` – OMOP CDM skeem (nt `ohdsi_cdm`);

`cohort_schema` – ATLASega loodud kohordi skeem;

`target_schema` – kasutaja oma skeem, kuhu laetakse kohordiga seotud andmed;

`cohort_id` – valitud kohordi ID, mille alusel laetakse seotud andmed;

`con` – ühendusobjekt, mille kaudu tehakse kõik järgnevad päringud.

Rakenduse alguses koostatakse SQL-päringud, mille kaudu tuuakse mällu tabelid, mis on välja toodud peatükis (vt peatükk 1.5). Kõik need andmed salvestatakse lokaalselt R-mälus objektidena, et vältida korduvaid ja koormavaid andmebaasipäringuid rakenduse kasutamise ajal. Selline lähenemine tähendab, et rakenduse esmakäivitamine võib võtta aega umbes 0-3 minutit olenevalt kohordi suurusest, kuid seejärel muutub kasutajaliides sujuvaks ja reageerivaks. Rakenduse eeltöötlemise kiirust on käsitletud lähemalt viimases peatükis (vt 3.2 Rakenduse jõudluse hindamine).

Interaktiivses töötluses kasutatakse andmete ümbervormindamiseks ja analüüsimiseks:

dplyr – tabelite ümberstruktureerimiseks;

`shiny::reactive()` – et tagada andmete dünaamiline filtreerimine vastavalt kasutaja sisendile.

Selline kaheosaline arhitektuur võimaldab analüüsida suuri andmekogumeid ning vähendab oluliselt päringute arvu andmebaasi.

2.4 Rakenduse tehniline ülesehitus

Rakenduse arhitektuur on üles ehitatud moodulipõhiselt ning jaotatud loogilisteks failikomponentideks, et toetada paremat hallatavust, taaskasutatavust ja laiendatavust. Struktuur võimaldab hoida kasutajaliidese, serveriloogika ja andmebaasipöördumised selgelt eraldatud, mis lihtsustab rakenduse arendamist ja skaleerimist.

Failistruktuur koosneb käivitusskriptist `app.R`, kasutajaliidese komponentidest (`sidebar.R`, `body.R`) ning serveripoole loogikast, mis on jaotatud kolmeks funktsionaalseks osaks: `server_main.R`, `plot_helpers.R` ja `db_connect.R`.

- `app.R` – rakenduse käivitusskript, mis ühendab UI ja serveri loogika ning installib ning toob mällu vajalikud R-paketid.
- `sidebar.R` – kasutajaliidese vasakpoolne valikumenüü, mis sisaldab sisendvälju vanuse, soo, diagnooside ja ajavahemike filtreerimiseks.

- `body.R` – kasutajaliidese peamine sisuala, kus renderdatakse visualiseeringud, tabelid ja seletavad elemendid.
- `server_main.R` – sisaldab reaktiivset serveriloogikat, mis realiseerib kõikide moodulite filtrite käsitlemise ja visualiseeringute renderdamise funktsioonid. Seal defineeritakse ka kõik reaktiivsed andmestikud ning rakendatakse kasutajasisendipõhised filtrid (`input$ageFilter`, `input$gender_display`, `input$timeFilter`, `input$primaryDiagnosis` jmt).
- `plot_helpers.R` – graafikute loomise funktsioonid, mis töötlevad visualiseeritavat andmestikku ja tagastavad `plotly` objektid.
- `db_connect.R` – andmebaasiga ühenduse loomise loogika, mis kasutab `DatabaseConnector::connect()` funktsiooni koos eeldefineeritud ühendusparameetritega ning sisaldab kõiki vajalikke SQL-päringud ja funktsioonid andmete toomiseks.

Kõik SQL-päringud on realiseeritud failis `db_connect.R`, ning nende tulemused kantakse rakenduses reaktiivsetesse objektidesse `server_main.R` failis. Iga visualiseeringu loomiseks kasutatakse standardset kolmeosalise komponendi struktuuri:

- üks `reactive()` objekt, mis filtreerib ja vormindab andmestiku;
- üks `renderPlotly()` või `renderTable()` funktsioon, mis renderdab väljundi liideses;
- üks joonisefunktsioon, mis asub `plot_helpers.R` failis ja millele edastatakse töödeldud andmestik.

Andmekaitseline kontroll rakendatakse igas `summarise()` järel konstruktsiooniga `filter(persons > 5)`. Kui tulemuseks on tühi andmestik, kuvatakse liideses tekstiline teade, et kuvamiseks pole piisavalt inimesi. Selline struktuur võimaldab rakendust laiendada ka keerukamate analüütiliste töövoogude jaoks.

2.5 Kasutajaliides

Kasutajaliidese elemendid sisaldavad selgitavaid hüpikuid (ingl *tooltip*), mis ilmuvad kursoriga elemendi kohale liikudes. Hüpikud aitavad kasutajatel paremini mõista sisendväljade ja graafikute tähendust ilma eraldi juhendita. Need on lisatud valitud filtritele ning keerukamatele joonistele.

Kasutajaliidese värve valides on arvestatud ka ligipääsetavusega – kasutusel on kontrastsed sinised ja neutraalsed toonid, mis sobivad ka värvipimedatele kasutajatele. Vastavalt uuringutele [17] on sinised toonid kõige paremini eristatavad nii deuteranoopia kui ka protanoopia korral.

Rakenduses kuvatakse viis vahelehte, mis esitavad andmestikul põhinevaid kohordikirjelduste tulemusi. Vastavalt filtritele kuvatakse tulemused erinevate jooniste, infokastide ja tabelitena. Pilte kasutajaliidese vaadetest on näha järgmises peatükis, kus viiakse läbi näidisuuring (vt peatükk 3.1). Kõikide vahelehtede filtreid saab näha visuaalselt lisades (vt Lisa 1 ja Lisa 2).

2.5.1 Vaheleht *Demographic summary*

Vahelehel *Demographic summary* kuvatakse valitud kohordi üldine kirjeldus. Tulemused võimaldavad kasutajal hinnata kohordi suurust, demograafilist jaotust ning vanuselisi trende.

Filtrid:

- *Cohort defining diagnoses* – mitme valikuga rippmenüü, mille kaudu saab valida need diagnoosikoodid, mille põhjal indiviid kohorti kuulub (juhuks, kui kohort koosneb erinevate diagnoosidega isikutest ja soovitakse vaadelda mõnda alamgruppi).
- *Comorbid diagnoses* – kasutaja saab valida kohordi alamgruppe, mille liikmetel on olnud enne kohorti sattumist mingisugune kindel diagnoos.
- *Grouping age by* – valik (ingl *radio button*) vanuse grupeerimiseks (1, 5 või 10 aastat).
- *Choose age* – liugur vanusevahemiku määramiseks (vaikimisi 0–120).
- *Choose gender* – valik, kas analüüsis osalevad ainult mehed, naised, mõlemad koos eristatult või eristamata.

Infokastid:

- Inimeste arv kohordis
- Inimeste arv kohordis pärast filtrite rakendamist
- Meeste ja naiste arv kohordis pärast filtrite rakendamist
- Keskmine vanus pärast filtrite rakendamist
- Keskmine vanusevahemik 95% usaldusintervallis pärast filtrite rakendamist
- Mediaanvanus pärast filtrite rakendamist

Joonised:

- Joonis *Gender distribution* - Soolise jaotuse rõngasdiagramm
- Joonis *Cohort start year distribution* - Kohorti sisenemise tulpdiagramm aastate lõikes. Kursoriga tulba peale liikudes kuvatakse grupi täpne suurus ning protsent kohordist/alamkohordist.
- Joonis *Age distribution* - Vanuselise jaotuse tulpdiagramm 95% usaldusintervalliga, kus vanuserühmad on grupeeritud vastavalt valikule.

2.5.2 Vaheleht *Diagnosis sources*

Sellel vahelehel keskendutakse esmase diagnoosi allika analüüsile. Kasutaja saab vaadelda, kust pärineb esmadiagnoos (nt retsept, raviarve) ning kes selle väljastas.

Filtrid:

- *Cohort defining diagnoses* – mitme valikuga rippmenüü, mille kaudu saab valida need diagnoosikoodid, mille põhjal indiviid kohorti kuulub (juhuks, kui kohort koosneb erinevate diagnoosidega isikutest ja soovitakse vaadelda mõnda alamgruppi).
- *Comorbid diagnoses* – kasutaja saab valida kohordi alamgruppe, kellel on olnud enne kohorti sattumist mingisugune kindel diagnoos.
- *Choose age* – liugur vanusevahemiku määramiseks (vaikimisi 0–120).
- *Choose gender* – valik, kas analüüsis osalevad ainult mehed, naised või mõlemad koos eristamata.

Joonised:

- Joonis *Diagnosis sources* - Esmase diagnoosi allika jaotuse tulpdiagramm (retsept, raviarve või elekroonilised terviseandmed).
- Joonis *Diagnosis sources overlap* - Esmase diagnoosi allikat kajastav Venn'i diagramm
- Joonis *Claim source breakdown* - Esmase raviarve väljastaja tüüpide tulpdiagramm (perearst, ambulatoorne eriarstiabivastuvõtt, statsionaarne eriarstiabivastuvõtt, statsionaarne õendusabi ning EMO).

2.5.3 Vaheleht *Comorbid diagnosis*

Vaheleht *Comorbid diagnosis* võimaldab analüüsida kohordis esinevaid kaasnevaid haigusi, kas automaatselt (10 kõige levinumat) või kasutaja poolt valitud diagnooside kaupa.

Filtrid:

- *Cohort defining diagnoses* – mitme valikuga rippmenüü, mille kaudu saab valida need diagnoosikoodid, mille põhjal indiviid kohorti kuulub eesmärgiga need jooniselt kõrvale jätta
- Valikukast, mille valimisel grupeeritakse diagnoosid spetsiifilisemalt või laiemalt. Näiteks M10, mis on üldine kood podagra diagnoosile või M10.1, mis viitab plii põhjustatud podagra ja M10.3, mis viitab neerutalitluse kahjustusest põhjustatud podagrale.
- *Comorbid diagnoses* - kasutaja saab valida kohordis esinenud kaasuvaid haigusi, mis on diagnoositud 1 aasta enne või pärast kohorti sattumise kuupäeva. Kui kasutaja teeb valiku, siis kuvatakse joonisele need kaasuvad diagnoosid, mis valituks osutusid. Kui kasutaja valikut ei tee, kuvatakse joonisele 10 kõige rohkem esinenud diagnoosi.
- *Choose age* – liugur vanusevahemiku määramiseks (vaikimisi 0–120).
- *Choose gender* – valik, kas analüüsis osalevad ainult mehed, naised või mõlemad.

Joonised:

- Joonis *Comorbid conditions distribution* - 10 sagedasemat kaasuvat diagnoosi või kasutaja poolt ette antud diagnoosidega joonis, kus kursoriga peale liikudes näeb täpset inimeste arvu ja protsenti kohordist/alamkohordist.

2.5.4 Vaheleht *Visits*

Vahelehel *Visits* analüüsitakse tervishoiuvisiitide jaotust enne ja pärast kohorti sattumist.

Filtrid:

- *Comorbid diagnoses* – kasutaja saab valida kohordi alamgruppe, kellel on olnud enne kohorti sattumist mingisugune kindel diagnoos.
- *Visit-related diagnoses* – mitme valikuga rippmenüü, mille kaudu saab valida diagnoosikoodid, mille valikul kuvatakse joonisel visiidid, mis on diagnoosikoodidega seotud.

- *Choose age* – liugur vanusevahemiku määramiseks (vaikimisi 0–120).
- *Choose gender* – valik, kas analüüsis osalevad ainult mehed, naised või mõlemad koos eristamata.
- *Time before and after cohort start* – ajaraam, millal kasutaja soovib visiite vaadelda. Valikus on 1 aasta enne kuni 1 aasta pärast kohorti sattumist kuudes.

Infokastid:

- Inimeste arv kohordis
- Inimeste arv kohordis pärast filtrite rakendamist
- Inimeste arv, kelle visiidid on seotud *Visit-related diagnoses* valikus kinnitatud diagnoosidega.
- Kõikide visiitide koguarv eeldefineeritud kohordis, 1 aasta enne kuni 1 aasta pärast.
- Visiitide arv pärast filtrite rakendamist.
- Visiitide arv, mis on seotud *Visit-related diagnoses* valikus kinnitatud diagnoosidega.

Joonised:

- Joonis *Visits distribution* - Visiitide tüübi jaotuse tulpdiagramm. Visiidid on jagatud perearsti vastuvõttudeks, ambulatoorseteks eriarstiabivastuvõttudeks, statsionaarseteks eriarstiabivastuvõttudeks, statsionaarseteks õendusabi vastuvõttudeks ning EMO visiitideks. Kursoriga tulbale peale liikudes kuvatakse sellesse tulpa kuuluvate inimeste arv, protsent kohordist, visiitide arv, keskmine visiitide arv inimese kohta, visiitide mediaanarv inimese kohta ning maksimaalne ja minimaalne visiitide arv inimese kohta.

Tabelid:

- Tabel *Label data table* - kuvatakse sama info, mis *Visits distribution* joonisel, kuid tabelikujul, mida saab kopeerida, salvestada PDF- ja CSV-failina.

2.5.5 Vaheleht *Deaths*

See vaheleht keskendub suremuse analüüsile kohordis. Joonis ja infokastid võimaldavad jälgida suremuse ajastust ning hinnata, kui suur osa inimesi suri valitud ajavahemiku jooksul pärast kohorti sattumist.

Filtrid:

- *Comorbid diagnoses* – mitme valikuga rippmenüü, mille kaudu saab valida need diagnoosikoodid, mis esinesid inimesel enne kohorti sattumist.
- *Choose age* – liugur vanusevahemiku määramiseks (vaikimisi 0–120).
- *Choose gender* – valik, kas analüüsis osalevad ainult mehed, naised, mõlemad koos eristatult või eristamata.
- *Time after cohort start* – liugur kuude valikuks (0-48 kuud), kuvatakse surmad selles ajavahemikus.

Infokastid:

- Inimeste arv kohordis
- Kõikide surmade arv kohordis
- Surmade protsent terves kohordis
- Inimeste arv pärast filtreerimist
- Surmade arv pärast filtreerimist
- Surmade protsent filtreeritud alamkohordist

Joonised:

- Joonis *Deaths after cohort start* - Surmajuhtumite jaotus kuude lõikes. Kursoriga tulba peale liikudes kuvatakse protsent kohordist/alamkohordist, kumulatiivne protsent, mis arvestab ka eelneva surmaprotsendiga ning täpne surmade arv sellel kuul.

3. Tulemused

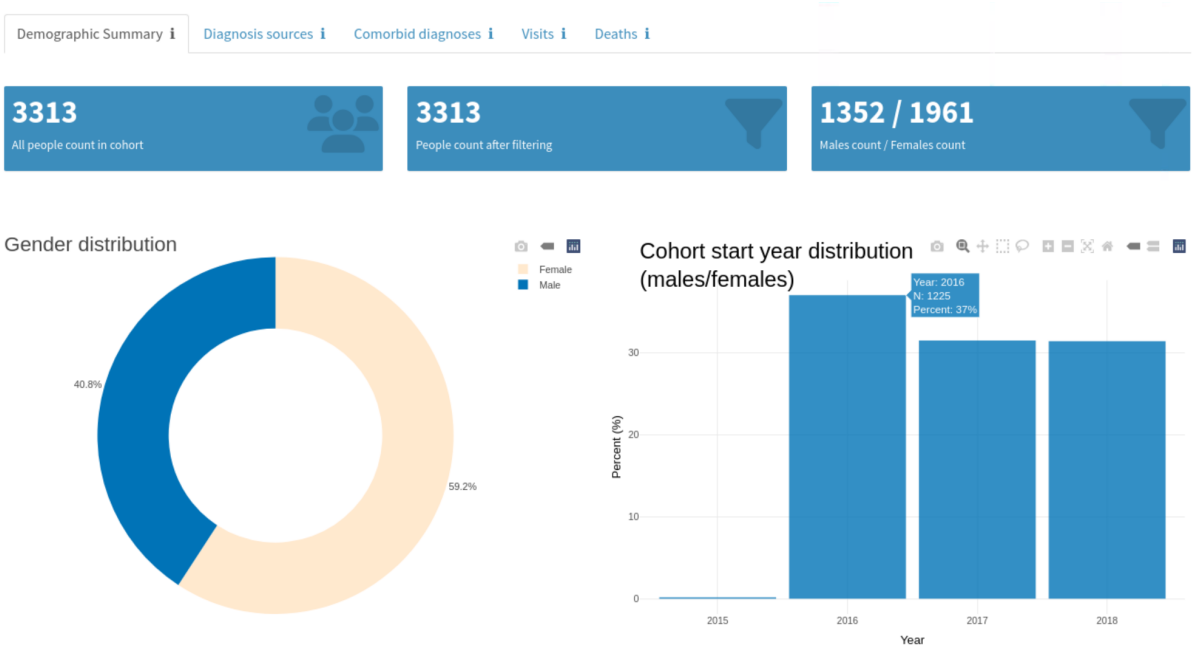
Tulemuste peatükk koosneb kolmest alajaotusest: näidisstsenaarium, rakenduse jõudluse hindamine ning edasised arendussuundad. Esmalt demonstreeritakse loodud rakenduse funktsionaalsust läbi konkreetse kliinilise näite – südamepuudulikkusega patsientide kohort. Seejärel testitakse rakenduse eellaadimiskiirust erinevate kohordisuuruste korral ja viimases alapeatükis käsitletakse rakenduse kitsaskohti ja võimalikke edasiarendusi.

3.1 Näidisstsenaarium

Näidisstsenaariumi eesmärk on loodud paketi CohortExplorerICD abil kirjeldada esmase südamepuudulikkuse kohorti. Kohort on eeldefineeritud tööriistaga ATLAS ja sinna kuuluvad inimesed, kellel on esmakordselt diagnoositud südamepuudulikkus. Isikud peavad vastama järgmistele kriteeriumitele:

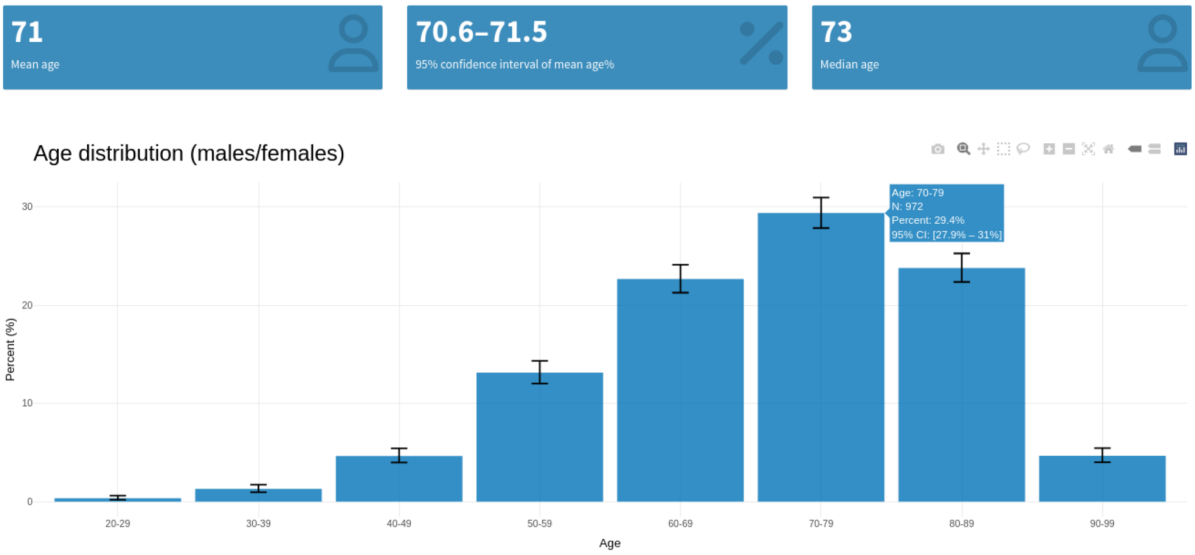
1. Vähemalt 18-aastased
2. Pole eelneva 90 päeva jooksul esinenud infarkti
3. Pole varasema nelja aasta jooksul diagnoositud südamekahjustusega hüpertooniatõbe südamepuudulikkusega (I11.0) ega südamepuudulikkust (I50)
4. On pärast diagnoosi saamist elus vähemalt 7-päeva

Kohordi kirjeldamisel kasutatakse rakenduse funktsionaalsust. Esimeselt vahelehelt on näha, et kohorti kuulub 3313 inimest, kellest 40.8% (1352 inimest) on mehed ja 58.2% (1961 inimest) naised. Kohordi algusaastad varieeruvad 2015. aastast 2018. aastani. Kõige rohkem inimesi on saanud esmase südamepuudulikkusega diagnoosi 2016. aastal. Neid oli 37% kohordist ehk 1225 inimest (vt Joonis 1).



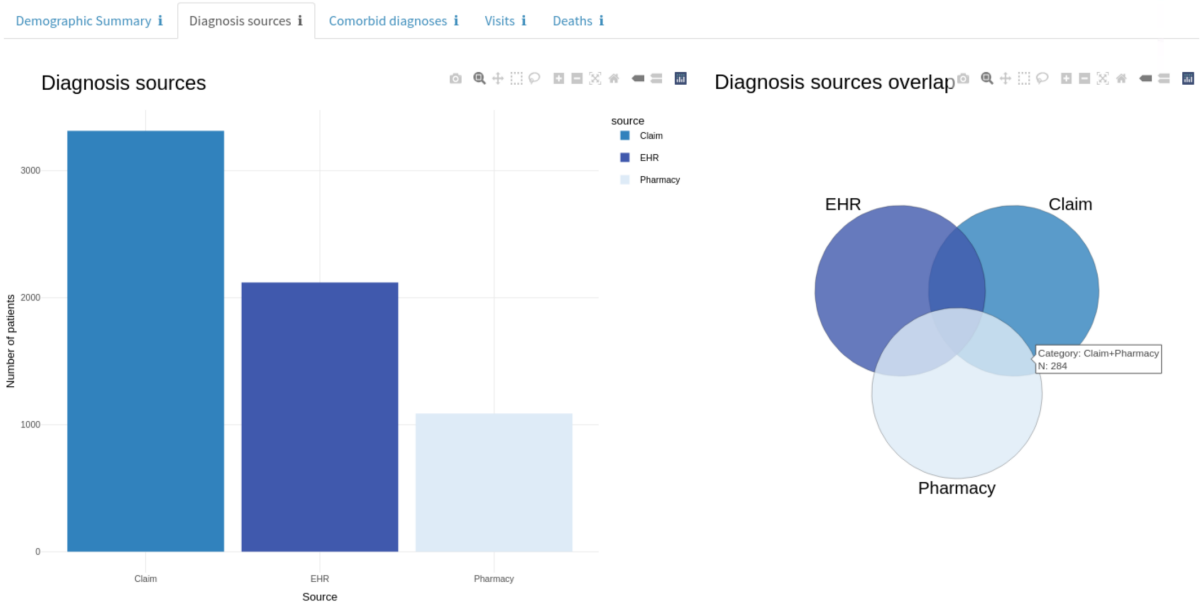
Joonis 1. Esimene osa vahelehest *Demographic summary*.

Patsientide vanuseline jaotus on esitatud alloleval joonisel (vt Joonis 2). Infokastidest on näha, et keskmine vanus on 71, samas kui mediaanvanus on 73 aastat. Keskmine vanus jääb 95% usaldusintervalliga 70.6 ja 71.5 vahele. Filtrites on valitud grupeerimiseks 10 aastat, seega joonisel grupeeritakse inimesed 10 aasta kaupa. Kõige suurem on 70-79-aastaste vanusegrupp, kuhu kuulub 29.4% kohordist ehk 972 inimest.



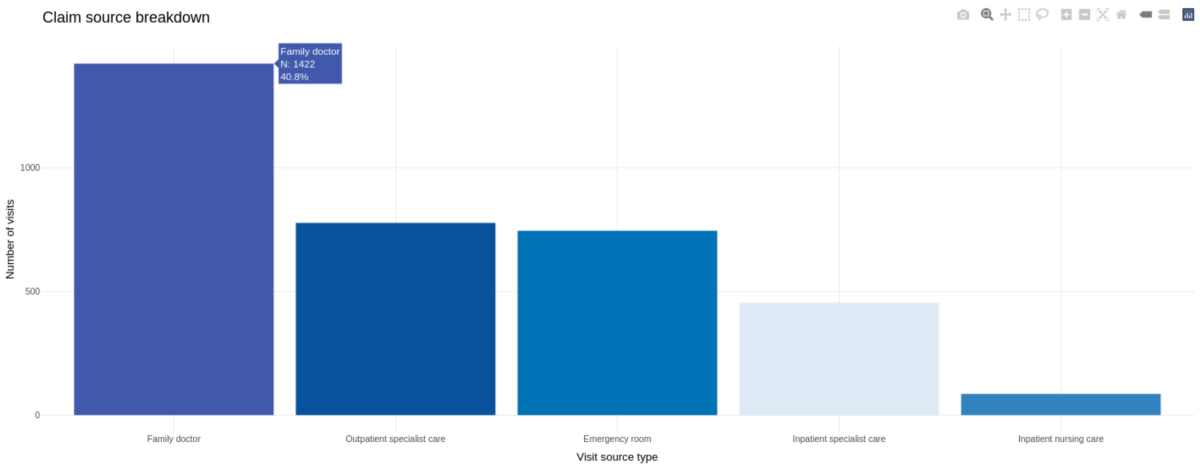
Joonis 2. Teine osa vahelehest *Demographic info*.

Joonisel 3 on näha esmase diagnoosi allika jaotus. Kõik kohorti kuuluvad patsiendid (100%) said esmase südamepuudulikkuse diagnoosi raviarve (ingl *Claim*) põhjal. Elektrooniliste haiguslugude (ingl *EHR*) kaudu said esmase diagnoosi ligikaudu 1900 inimest, samas kui apteegiandmete (ingl *Pharmacy*) põhjal diagnoositi neid vähem kui 1000. Venn-diagrammist (paremal) on näha, et osa patsiente kuulus mitmesse andmeallikasse, näiteks said nad diagnoosi samal päeval nii elektroonilisest haigusloost kui ka raviarvelt.



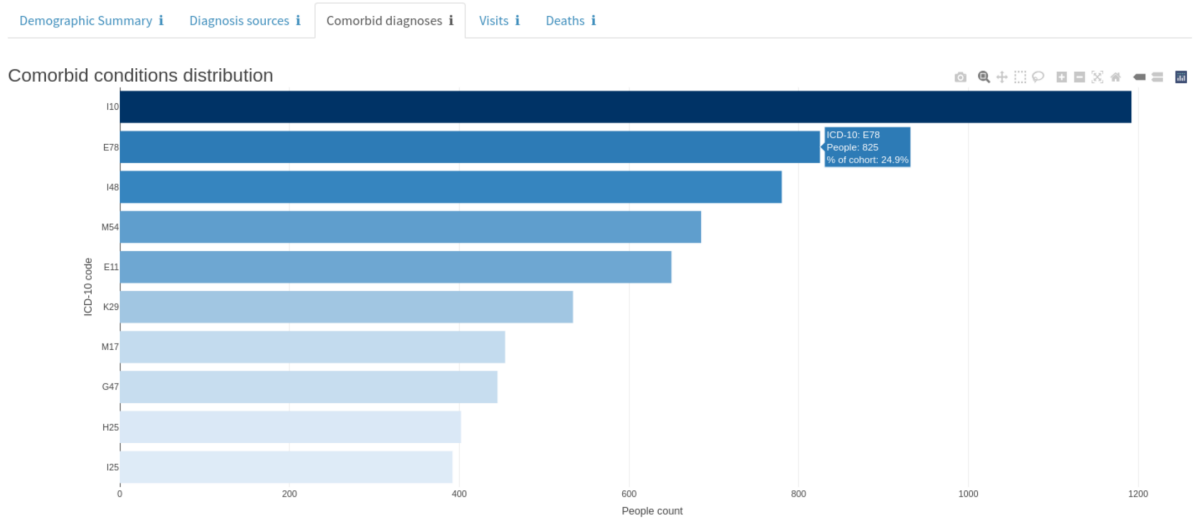
Joonis 3. Esimene osa vahelehest *Diagnosis sources*.

Allolev joonis (vt Joonis 4) annab ülevaate, kust väljastati raviarve neile, kelle esmane diagnoos tuli raviarvelt. Kõige enam (ligikaudu 1400 inimest) said diagnoosi perearsti kaudu. Seejärel järgnesid EMO, statsionaarsed ja ambulatoorsed visiidid. Väiksem osa diagnoose tuli õenduse (ingl *inpatient nursing care*) visiitidelt.



Joonis 4. Teine osa vahelehest *Diagnosis sources*.

Joonisel 5 on kujutatud kõige sagedamini esinenud kaasuvate haiguste (ingl *comorbid conditions*) jaotust RHK-10 koodide alusel. Kõige levinumad olid hüpertensioon (I10), düslipideemia (E78), kodade virvendus ja laperdus (südamerütmihäired, I48) ja 2. tüüpi diabeet (E11). Paljudel patsientidel esines ka luu- ja lihaskonna ning seedetrakti haigusi. Valik põhineb diagnoosidel, mis esinesid kuni üks aasta enne kohorti sisenemist ning on selles kohordis 10 kõige levinumat.



Joonis 5. Vaheleht *Comorbid diagnoses*.

Visiitide joonisel (vt Joonis 6) on kujutatud erinevate külastustüüpide (ingl *visit types*) jaotus ühe aasta jooksul enne ja pärast kohordi algust. Kõik inimesed kohordist käisid kokku 72961 visiidil. Sellel joonisel on filtreeritud just visiidid, mis on toimunud 1 aasta enne kohorti sattumist, neid oli kokku 39999. Kõige enam külastati perearsti (ingl *Family doctor*) – üle 3200 inimese

käisid rohkem kui 20 000 visiidil. Ka ambulatoorsete eriarstide vastuvõttudel käis enam kui 1000 inimest (ingl *Outpatient specialist care*). Iga tulba kohal kuvatakse detailne statistika: inimeste arv, visiitide koguarv, keskmine ja mediaan külastuste arv inimese kohta, maksimaalne ja minimaalne visiitide arv ning osakaal kogu kohordist.

Joonis 6. Esimene osa vahelehest *Visits*.

Tabelis (vt Tabel 2) on toodud üksikasjalik statistika iga külastustüübi (ingl *visit type*) kohta ka tabelkujul: inimeste arv, visiitide koguarv, keskmine, mediaan, maksimaalne ja minimaalne visiitide arv ning osakaal kogu kohordist (ingl *percent from cohort*). Näiteks EMO külastas 1 aasta enne kohorti sattumist 1619 inimest (48.9% kohordist) keskmiselt 1.6 korda inimese kohta.

Tabel 2. Teine osa vahelehest *Visits*.

Label data table

Copy CSV Excel PDF Print

Visit type	People	Visits	Average per person	Median	Max	Min	Percent from cohort (%)
Emergency room	1619	2645	1.6	1	17	1	48.9
Family doctor	3234	20729	6.4	6	32	1	97.6
Inpatient nursing care	443	937	2.1	1	24	1	13.4
Inpatient specialist care	1056	1779	1.7	1	24	1	31.9
Other	78	96	1.2	1	5	1	2.4
Outpatient specialist care	2506	13813	5.5	4	58	1	75.6

Showing 1 to 6 of 6 entries

Joonisel (vt Joonis 7) on kujutatud surmade jaotus ajas pärast kohorti sisenemist. Filtritest on valitud ajaraamiks 0-12 kuud. Enamik surmasid toimus esimese kuue kuu jooksul. Esimesel kuul suri ligikaudu 1.5% patsientidest. Kokku suri esimese 12 kuu jooksul 284 inimest, mis moodustab 8.6% filtreeritud kohordist. Tulpadel on kuvatud iga kuu kohta surmade arv, osakaal ning kumulatiivne protsent (ingl *cumulative percent*).

Joonis 7. Vaheleht *Deaths*.

Eeltoodud joonised annavad ülevaate südamepuudulikkuse esmase diagnoosiga kohordi demograafilistest ja kliinilistest omadustest, esmase diagnoosi päritolust, kaasuvatest haigustest, tervishoiuteenuste kasutusest ning suremusest. Nende tulemuste põhjal on võimalik tuvastada olulisi mustreid haiguskoormuses ja tervishoiusüsteemi koormuses, mis võivad toetada edasist teadustööd.

3.2 Rakenduse jõudluse hindamine

Rakenduste stresstestimine (ingl *stress testing*) on tarkvarasüsteemide testimismeetod, mille eesmärk on hinnata süsteemi töökindlust ja jõudlust olukordades, kus ressursikasutus või sisendsuurused lähenevad äärmusväärtustele. Tegemist on olulise sammuga skaleeritavate

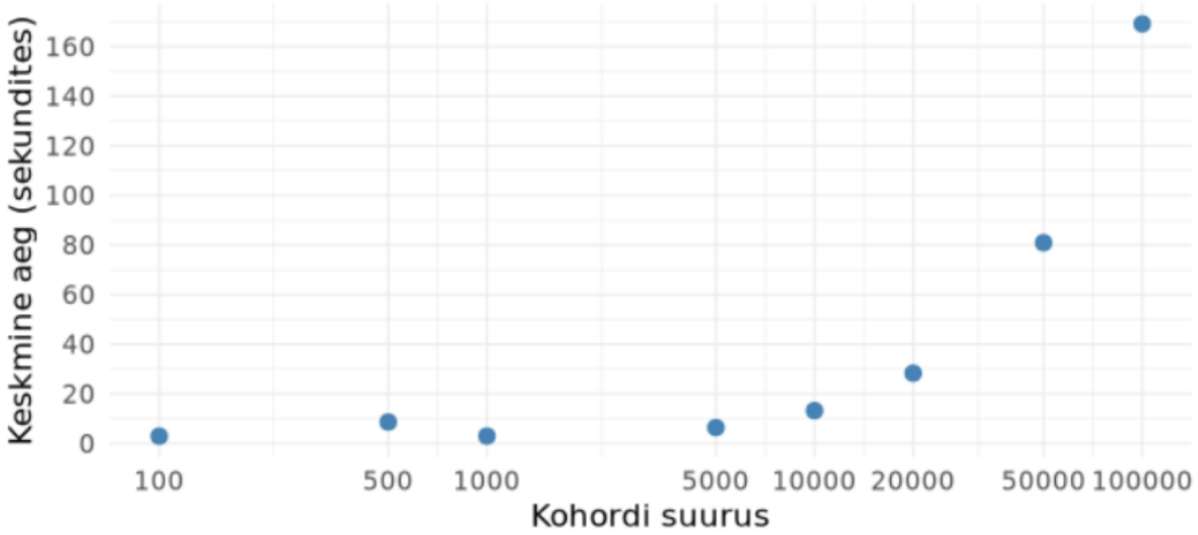
rakenduste valideerimisel, eriti interaktiivsetes süsteemides nagu veebipõhised analüüsivahendid [18].

Antud töös viidi läbi stresstest, et hinnata rakenduse käitumist olukorras, kus sisendkohordi suurus varieerus alates väikestest valimitest (100 isikut) kuni suurte valimiteni (100 000 isikut). Testi eesmärk oli mõõta, kui kiiresti rakendus suudab laadida ja töödelda andmeid enne kasutajaliidese kuvamist, ning kas teatud sisendsuuruste juures esineb jõudluslangust. Mõõtmisel kasutati R programmeerimiskeele funktsiooni Sys.time[16], millega määrati mõõtmisaja algus ja lõpphetk.

Testimisel fikseeriti kaks muutujat:

- **Kohordi suurus (N)** ehk sisendina kasutatud indiviidide arv;
- **Andmete laadimise kestus (sekundites)** alates päringute algusest kuni rakendusi kuvamiseni.

Mõõtmised viidi läbi ühesuguses keskkonnas. Tulemused logiti struktureeritult CSV-vormingus, võimaldades hiljem nende võrdlust ja visualiseerimist. Sisendsuurusi oli kokku 8 (100, 500, 1000, 5000, 10000, 20000, 50000 ja 100000 isikut), iga kohordisuurusega tehti 30+ kordusmõõtmist. Katse käigus registreeriti järgmised tulemused, mida kirjendab järgnev joonis (vt Joonis 8):



Joonis 8. Rakenduse eellaadimise aja sõltuvus kohordisuurusest

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

Jooniselt 8 on näha selget seost kohordi suuruse ja laadimisaja vahel. Väikeste kohordisuuruste (nt 100 või 1000 indiviidi) korral jääb keskmine laadimisaeg alla 3 sekundi. Kui sisendandmete maht kasvab 10 000 inimeseni, tõuseb laadimisaeg juba üle 13 sekundi. Suuremate kohordisuuruste puhul kasvab aeg oluliselt – 50 000 inimese korral ulatub keskmine laadimisaeg 80,9 sekundini ja 100 000 puhul juba 169 sekundini. See viitab logaritmilisele või isegi kiiremini kasvavale sõltuvusele, kus süsteemi jõudlus langeb märgatavalt suure sisendi korral. Järgmisena testitakse, kas lineaarne regressioon sobib mõõtmistele ning millise täpsusega see suudab kirjeldada laadimisaja sõltuvust kohordi suurusest.

Järgnev lõik põhineb statistikaõpikul [19]. Lineaarne regressioon on statistiline meetod, mida kasutatakse seose modelleerimiseks sõltuva muutuja y ja ühe või mitme sõltumatu muutuja x vahel. Lihtsa lineaarse regressiooni üldkuju on $y = \beta_0 + \beta_1 x + \varepsilon$, kus β_0 tähistab lõikepunkti, β_1 regressioonikordajat, ning ε on juhusliku vea komponent. Mudeli eesmärk on hinnata, kuidas muutub muutuja y sõltuvalt sellest, kuidas muutub seletav muutuja x. Lineaarne regressioon eeldab, et see seos on sirgjooneline ning et vead on juhuslikult jaotunud ning sõltumatud.

Tabel 3. Lineaarse mudeli tulemused: koefitsiendid ja mudeli sobivusnäitajad

Tüüp	**Näitaja**	**Väärtus**	**Standardviga**	**t-statistika**	**p-väärtus**
Koefitsient	Lõikepunkt (Intercept)	-87,502	5,603	-15,62	$<0{,}001$
Koefitsient	Logaritmiline kohordisuurus	15,497	0,681	22,76	$<0{,}001$
Mudeli sobivus	Jäägi standardviga	34,941	—	—	—
Mudeli sobivus	R-ruut	0,553	—	—	—
Mudeli sobivus	Parandatud R-ruut	0,552	—	—	—
Mudeli sobivus	F-statistiku	518,14	—	—	—
Mudeli sobivus	F p-väärtus	$<0{,}001$	—	—	—

Tabelis 3 on kirjeldatud lineaarse regressioonimudeli tulemusi ja need viitavad, et rakenduse laadimisaja ja kohordi suuruse vahel esineb tugev ja statistiliselt oluline seos. Logaritmilise kohordisuuruse koefitsient (15,497) näitab, et iga kümnekordne suurenemine sisendsuuruses (nt 1 000 → 10 000) toob kaasa laadimisaja pikenemise ligikaudu 15,5 sekundi võrra. Mudeli sobivust iseloomustavad näitajad – R-ruut (0,553) ja parandatud R-ruut (0,552) – viitavad sellele, et ligikaudu 55% laadimisaja varieeruvusest on seletatav kohordisuuruse logaritmiga. See osutab mõõdukale, kuid mitte täielikule sobivusele, mistõttu võib eeldada, et laadimisaega mõjutavad

ka teised tegurid, näiteks päringute keerukus. F-statistiku väärtus (518,14) ja sellele vastav p-väärtus (< 0,001) kinnitavad, et kogu mudel on statistiliselt oluline. Jäägi standardviga (umbes 35 sekundit) viitab sellele, et vaatamata heale seosele võib tegelik laadimisaeg mõnevõrra kõikuda ka mudelis mitteseletatud põhjustel. Kokkuvõttes toetavad mudelitulemused hüpoteesi, et rakenduse laadimisaeg sõltub sisendkohordi suurusest logaritmiliselt. Mudeli sobivus on piisav praktiliste järelduste tegemiseks, kuid edasised analüüsid võiksid käsitleda ka süsteemiväliseid tegureid, mis võivad jõudlust mõjutada suuremate sisendite korral.

3.3 Edasiarendus

Kuigi loodud rakendus võimaldab kirjeldada kohorti mitmekülgselt ning pakub kasutajale interaktiivset ja paindlikku kasutajakogemust, ilmnes arenduse käigus mitmeid arenguvõimalusi, millele võiks tulevikus tähelepanu pöörata.

Esiteks vajab täiendavat tähelepanu koostalitlusvõime ATLAS platvormi ja teiste OHDSI tööriistadega. Praegune lahendus eeldab käsitsi määratud kohordi liikmeid ning ei ole automaatselt ühendatud tööriistaga ATLAS defineeritud kohorti määravate sündmustega. Tugevam integratsioon võimaldaks dünaamilist kohordi importimist ja looks aluse suuremale automatiseeritusele ning töövoogude tõhusamale haldamisele.

Teiseks on soovitav laiendada toe mitmekordsetele isikukirjetele, st olukordadele, kus üks inimene kuulub mitme episoodiga samasse kohorti. Käesolev versioon keskendub ainult esmastele sündmustele, mistõttu korduvad juhtumid jäävad analüüsist välja. Sellise toe lisamine suurendaks rakenduse sobivust näiteks krooniliste haiguste või korduvate ravijuhtude analüüsimisel.

Kolmandaks on loodud rakendus hetkel arendatud konkreetse andmebaasi jaoks ning puudub veel täielik platvormiülene tugi. Edaspidi võiks olla rakendus võimeline töötama ka teiste OMOP CDM andmebaasidega, sõltumata nende asukohast või tehnilisest platvormist, kasutades standardiseeritud liideseid ja dünaamilist skeemihaldust. Selline laiendus suurendaks rakenduse kasutusvõimalusi.

Lisavõimalusena võiks rakendus pakkuda automaatse aruandluse funktsiooni, mis võimaldaks genereerida analüüsi tulemuste põhjal PDF-formaadis aruande koos valitud jooniste ja tabelitega. See lihtsustaks andmete dokumenteerimist, jagamist ning kasutamist teadustöös või igapäevases aruandluses.

Visualiseerimise täiustamiseks võiks kaaluda mitmesuguste täiendavate jooniste ja tabelite lisamist. Et kindlaks teha, milline kujundus, graafikud või info oleksid sihtgrupile (nt arstid, teadlased, terviseandmete analüütikud) kõige kasulikumad, võiks viia läbi kasutajauuringu või küsitluse. Selline lähenemine aitaks paremini mõista, millist teavet või funktsionaalsust kasutajad tegelikult ootavad, ning suunaks arendust kasutajakesksemaks.

Jõudluse vaates võib alati kaaluda optimeerimist, eriti suurte kohortide korral. Kuigi rakendus on võimeline töötlema mitmekümnetuhandelisi kohorte, võib mastaabi kasvades tekkida vajadus parema andmetöötlusloogika või hajutatud töötluse kasutuselevõtuks.

Tähelepanu väärib ka kasutajaliidese disain. Selle veelgi intuitiivsemaks muutmine, sh parema navigeeritavuse ja visuaalse selguse tagamine, aitaks suurendada rakenduse kasutusmugavust.

Kokkuvõte

Bakalaureusetöö eesmärk oli arendada töövoog ja interaktiivne rakendus, mis võimaldab kirjeldada OMOP CDM formaadis esitatud meditsiinilisi kohorte. Töö keskendus praktilise lahenduse loomisele, mis toetaks terviseandmete kohortanalüüsi, pakkudes võimalust visualiseerida kohordi omadusi.

Loodud pakett CohortExplorerICD võimaldab kasutajatel määratleda esmaseid ja kaasuvaid diagnoose rahvusvahelise haiguste klassifikatsiooni RHK-10 alusel, rakendada filtreid soo, vanuse ja ajavahemike järgi ning analüüsida kohortide statistilisi näitajaid. Rakenduses on tähelepanu pööratud andmekaitse- ja ligipääsetavusnõuetele: tulemusi ei kuvata, kui alamhulgad sisaldavad viit või vähem isikut, ning visualiseerimisel kasutatakse värve, mis sobivad ka värvipimedatele kasutajatele.

Implementatsiooni peatükis (vt peatükk 2) kirjeldati rakenduse ülesehitust, andmete laadimist ning visualiseerimist ggplot2 ja plotly pakettidega. Andmete eellaadimine mällu võimaldab rakendusel pakkuda sujuvat kasutuskogemust, kus filtrite rakendamine ja visualiseeringute uuendamine toimub ilma korduvate andmebaasipäringuteta. Rakenduse arendamisel ja testimisel kasutati MAITT andmestikku ja kohortide defineerimiseks OHDSI tööriista ATLAS.

Näidisstsenaariumina kirjeldati südamepuudulikkuse patsientide kohorti, mille alusel demonstreeriti rakenduse võimalusi analüüsida demograafilisi näitajaid, diagnoosiallikate päritolu, kaasuvate haiguste esinemist, tervishoiuvisiitide mustreid ja surmajuhtumite jaotust ajas. Lisaks viidi läbi rakenduse jõudlustest erinevate kohordisuuruste puhul, et hinnata süsteemi töökindlust. Testide tulemused näitasid, et rakendus on suuteline töötlema suuri kohorte (kuni 100 000 inimest) aktsepteeritava aja jooksul, säilitades interaktiivse kasutajakogemuse.

Töö lõpus toodi välja võimalused edasisteks arendusteks, sealhulgas funktsionaalsuse lisamine, täiendav platvormiülene andmetugi, täiendavad joonised ning rakenduse optimeerimine veelgi suuremate andmehulkade jaoks. Lõputöö tulemuseks on kohandatav tööriist, mis laiendab võimalusi Eesti terviseandmete analüüsimiseks ning toetab andmepõhiseid uuringuid OHDSI kogukonnas.

Viited

[1] Raghupathi W. ja Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2014, 2(1): 3, https://doi.org/10.1186/2047-2501-2-3 (12.04.2025).

[2] Observational Health Data Sciences and Informatics. Standardized Data: The OMOP Common Data Model. https://www.ohdsi.org/data-standardization/. (13.04.2025).

[3] Kandola A. Cohort studies: What they are, examples, and types. *Medical News Today*, 2023. https://www.medicalnewstoday.com/articles/281703/ (12.04.2025).

[4] Observational Health Data Sciences and Informatics. Who We Are. https://ohdsi.org/who-we-are/. (13.04.2025).

[5] Observational Health Data Sciences and Informatics Europe. Homepage. https://www.ohdsi-europe.org/ (13.04.2025).

[6] Reisberg S., Mooses K., Kolde R., Kõrgvee L.-T. ja Vilo J. Uudne lähenemine – OMOP-andmemudelil põhinevad terviseuuringud. *Eesti Arst*, 2024, 103(9): 420-429. https://ojs.utlib.ee/index.php/EA/article/view/24470 (15.04.2025).

[7] Observational Health Data Sciences and Informatics. OHDSI Symposium 'Scaling up Reliable Evidence across Europe' June 3rd, 2024. https://www.ohdsi-europe.org/index.php/symposium/45-archive-symposium-2024 (12.04.2025).

[8] Observational Health Data Sciences and Informatics. Common Data Model: Background. https://ohdsi.github.io/CommonDataModel/background.html (12.04.2025).

[9] World Health Organization. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). 2016. https://icd.who.int/browse10/ (12.04.2025).

[10] Khan M. S., Samman Tahhan A., Vaduganathan M., Greene S. J., Alrohaibani A., Anker S. D., Vardeny O., Fonarow G. C. ja Butler J. Trends in prevalence of comorbidities in heart failure clinical trials. *Medical Care*, 2022. https://pubmed.ncbi.nlm.nih.gov/32293090/ (12.04.2025).

[11] Tervise Arengu Instituut. RHK ehk rahvusvaheline haiguste klassifikatsioon. https://www.tai.ee/et/instituudist/meditsiiniterminoloogia-kompetentsikeskus/who-klassifikaatorid-rfk-ja-rhk/rhk-ehk (12.04.2025).

[12] Hammoudeh S., Gadelhaq W. ja Janahi I. Prospective Cohort Studies in Medical Research. *Cohort Studies in Health Sciences*. IntechOpen, 2018. https://www.intechopen.com/chapters/60939.

[13] Hughes N., Rijnbeek P. R., Bochove K. van, Duarte-Salles T., Steinbeisser C., Vizcaya D., Prieto-Alhambra D. ja Ryan P. Evaluating a novel approach to stimulate open science collaborations: a case series of ‘study-a-thon’ events within the OHDSI and European IMI communities. *JAMIA Open*, 2022, 5(4): ooac100. https://doi.org/10.1093/jamiaopen/ooac100 (12.04.2025).

[14] Gilbert J., Rao G., Schuemie M., Ryan P. ja Weaver J. CohortDiagnostics: Diagnostics for OHDSI Cohorts. R package version 3.4.1. https://github.com/OHDSI/CohortDiagnostics (12.04.2025).

[15] Catala M., Guo Y., Lopez-Guell K., Burn E., Mercade-Besora N. ja Alcalde M. CohortCharacteristics: Summarise and Visualise Characteristics of Patients in the OMOP CDM. R package version 0.5.1. https://darwin-eu.github.io/CohortCharacteristics/ (12.04.2025).

[16] 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. ja Reisberg S. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. *JAMIA Open*, 2023, 6(4): ooad100. https://doi.org/10.1093/jamiaopen/ooad100 (12.04.2025).

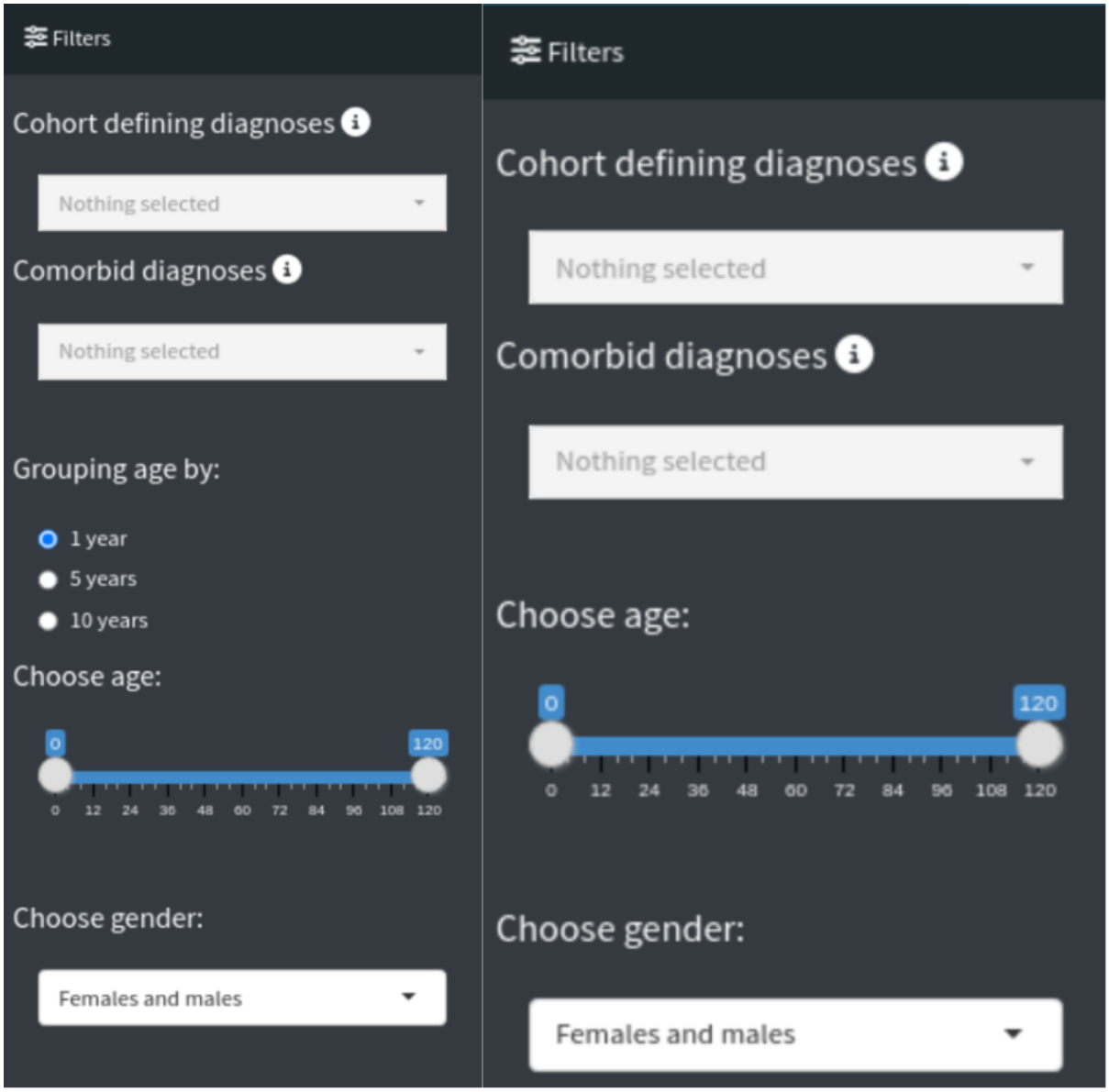
[17] Wong B. Points of view: Color blindness. *Nature Methods*, 2011, 8(6): 441. https://www.researchgate.net/publication/51506740_Points_of_view_Color_blindness (12.04.2025).

[18] Jorgensen P. C. Software Testing: A Craftsman’s Approach. 4. väljaanne. Auerbach Publications, 2013.

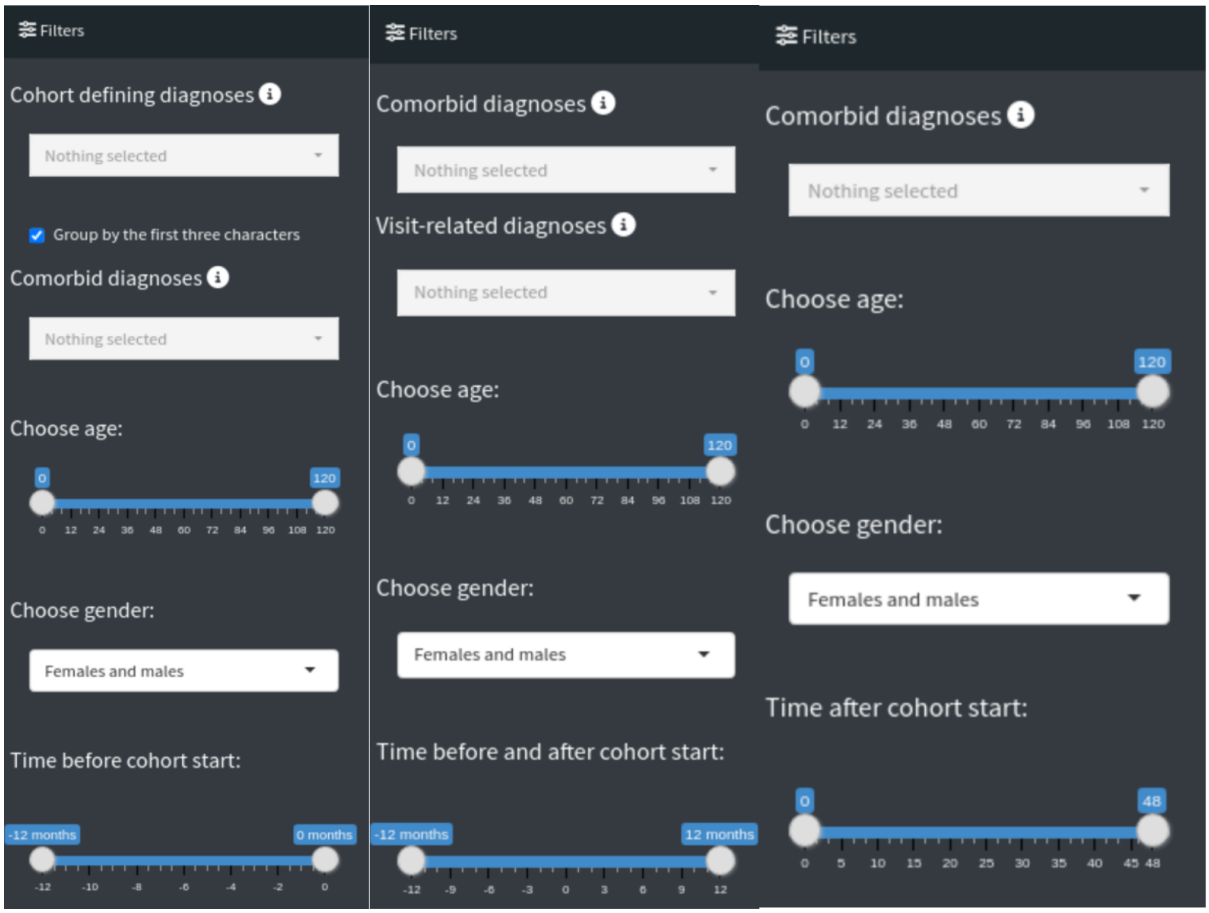
[19] James G., Witten D., Hastie T. ja Tibshirani R. An Introduction to Statistical Learning: with Applications in R. Springer, 2013.

Lisad

I. Vahelehtede *Demographic summary* ja *Diagnosis sources* külgribad



II. Vahelehtede *Comorbid diagnoses*, *Visits* ja *Deaths* külgribad



III. GitHubi repositoorium

Lõputöös loodud rakenduse leiab lingilt: https://github.com/KarinRosen/CohortExplorerICD.git

IV. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Karin Rosenberg,**

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „**R-pakett meditsiiniliste kohortide kirjeldamiseks**”, mille juhendaja(d) on **Kerli Mooses**, 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.

Karin Rosenberg
15.05.2025