TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Rasmus Mirma

Haiguste sageduste visualiseerimine OMOP CDM andmebaasi põhjal

Bakalaureusetöö (9 EAP)

Juhendaja:
Jaak Vilo, PhD

Tartu 2025

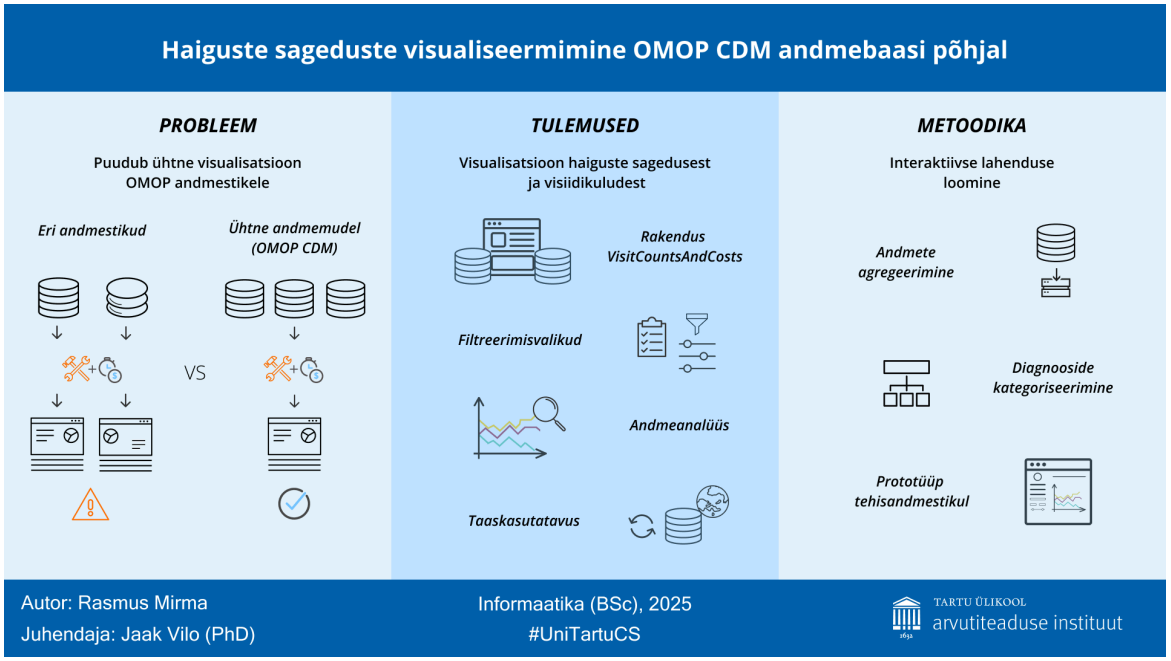
Haiguste sageduste visualiseerimine OMOP CDM andmebaasi põhjal

Lühikokkuvõte:

Terviseandmete töötlemine annab võimaluse rahvastiku tervise edendamiseks, uute ravimeetodite loomiseks ning ennetus- ja arendustegevuste planeerimiseks. Olemasolevad lahendused terviseandmete visualiseerimiseks keskenduvad kindlatele andmetüüpidele ega paku ühtlustatud kujul visualiseerimise võimekust. Terviseandmete valdkonnas üha enam rakendust leidvaks andmemudeliks on Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), mis oma olemuselt määratleb nii andmemudeli vormingu kui ka standardiseeritud terminoloogiad. Bakalaureusetöö eesmärk oli luua interaktiivne andmete töölaud, mida saaks rakendada kõikide OMOP CDM kujul olevate andmebaaside puhul. Valminud töölaud annab võimaluse vaadelda haiguste sagedusi ning nendega kaasnenud kulusid diagnooside ja demograafiliste filtrite lõikes. Lisaks eelnevale on töös kirjeldatud taaskasutatav andmepäring, mille abil on võimalik tabelitest kokku lugeda unikaalsete patsientide arv koos neile osutatud tervishoiuteenuste kogukuludega nii kalendriaasta, vanuse- kui ka soorühmade lõikes. Lahenduse edasiarenduseks on töös esitatud rakendamata funktsionaalsused ja parandusettepanekud.

Võtmesõnad: Terviseandmed, OMOP CDM, andmete visualiseerimine, töölaud

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika



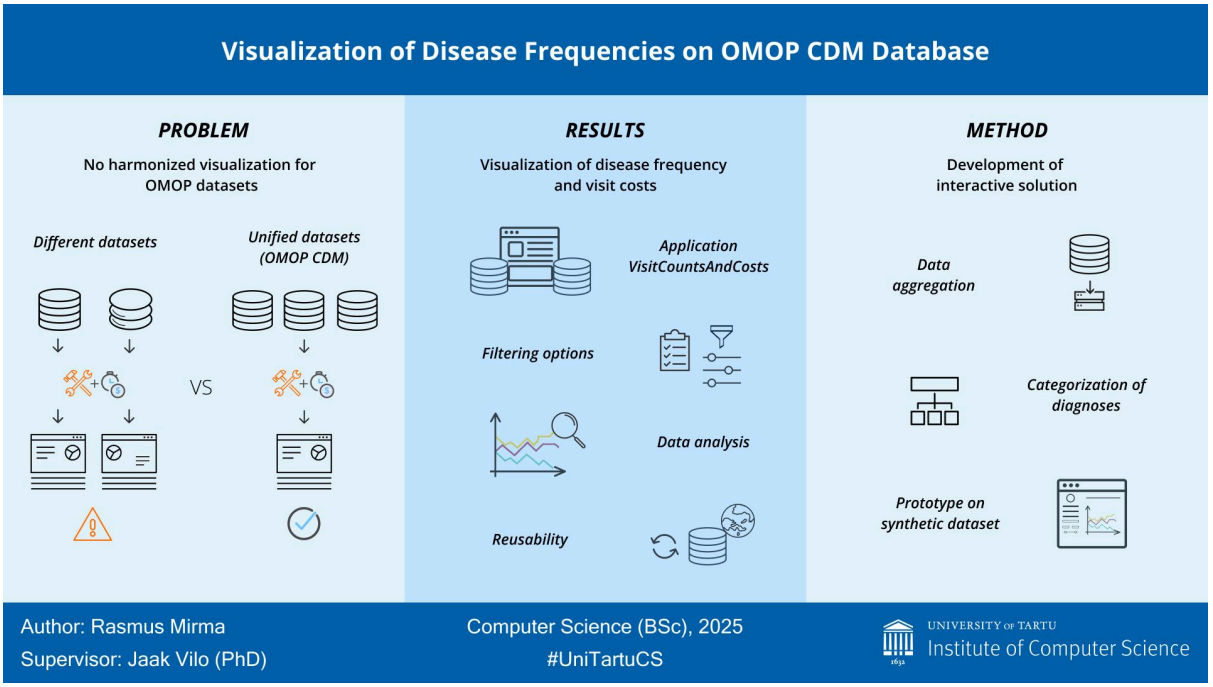
Visualization of Disease Frequencies on OMOP CDM Database

Abstract:

Processing health data paves the way to improve public health, the development of new treatment methods, and the planning of preventive and development activities. Existing solutions for visualizing health data focus on specific data types and do not provide harmonized visualization capabilities. Within the health-data domain, the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) is becoming an increasingly adopted data model, defining both the data-model format and standardized terminologies. The goal of this bachelor's thesis was to create an interactive data dashboard that can be applied to any database in OMOP CDM form. The completed dashboard makes it possible to examine disease frequencies and their associated costs by diagnosis and demographic filters. In addition, the thesis describes a reusable data query that counts the number of unique patients and the total cost of healthcare services provided to them by calendar year, age, and gender group. For further development of the solution, the thesis presents unimplemented functionalities and improvement proposals.

Keywords: Health data, OMOP CDM, data visualization, dashboard

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics



Sisukord

1. Sissejuhatus

Tänapäeval koguneb andmeid rohkem kui kunagi varem. Meditsiinivaldkonnas on palju salvestatavat teavet, nagu näiteks haiguslugude kirjeldused või mõõtmiste tulemused. Visiitide tulemusena täienevad andmestikud igapäevaselt. Rahvastiku tervise edendamiseks, uute ravimeetodite loomiseks ja ennetustegevuste planeerimiseks on tarvis kogutud terviseandmeid ka töödelda. Üheks võimalikuks andmete töötlemise viisiks on nende kirjeldamine läbi visualiseerimise, mis muudab andmestiku mustrid ja trendid paremini mõistetavamaks (Sügis jt, 2024). Andmete visualiseerimine on andmeanalüüsi oluline osa, tänu millele saame anda ülevaate näiteks haiguste sagedustest ja visiitidega kaasnenud kuludest.

Tehnoloogia arenduskeskuste (TAK) programmist rahastatud tarkvaraettevõtte STACC poolt on varasemalt küll arendatud lahendus haiguste sageduste visualiseerimiseks, kuid see on loodud toimima vaid kitsalt fikseeritud andmestikul. Nende loodud lahendus annab ülevaate eriarstiabi saanute arvust ja kuludest vanuse ning põhidiagnooside lõikes toetudes Tervisekassa raviarvetele. Visualisatsiooni aluseks võeti haigustrajektoore kirjeldavas artiklis (Jensen jt, 2014) esinenud joonis. Kuna andmeid ei uuendatud peale TAK programmi lõppu, siis on andmed vaadeldavad vaid perioodist 2013–2019.

Rahvusvaheliselt on terviseandmete ühtlustamise tarbeks arendatud ühtne andmemudel *Observational Medical Outcomes Partnership Common Data Model* (OMOP CDM). Ühtlustatud andmemudeli kasu väljendub eelkõige rahvusvaheliste uuringute tulemuste valideerimisel ning hõlbustab ka soovituste rakendamist Eestis. Ühtsel kujul olevad andmestikud aitavad vältida muidu vajalike ümberteisendusi eri struktuuriga andmestike vahel (Solvak jt, 2022). Eestis on nüüdseks valdav osa riiklikest terviseandmetest standardiseeritud OMOP andmemudelile, mis võimaldab laialdasemat koostööd eri andmeomanike vahel. Tartu Ülikooli terviseinformaatika uurimisrühm on tegelenud Eesti andmekogude viimisega OMOP kujule. Lisaks eelnevale on uurimisrühm välja töötanud mitmeid taaskasutatavaid teenindusskripte ja analüüsimeetodeid ühtsel kujul olevatel andmetel rakendamiseks (Reisberg jt, 2024).

Käesoleva bakalaureusetöö eesmärk on arendada taaskasutatav ja avatud lähtekoodiga interaktiivne andmete töölaud, mida saaks rakendada kõikide OMOP CDM kujul olevate andmebaaside puhul. Rakenduse põhifookuses on olla tööriistaks eelkõige andmeteadlastele. Reaalandmestiku puhastamise tulemusena oleks käesoleval tööl potentsiaali olla ka avalikult

kasutatav tööriist. Selgelt visualiseeritud andmed aitavad inimestel paremini mõista enda tervist ja haigusriske. Kui on konkreetsed võrreldavad numbrid, millelt nähtuvad tervisetrendid ning haigustega kaasnevad kulud, siis tekib suurem motivatsioon võtta vastutust oma tervise eest (Reisberg, 2016).

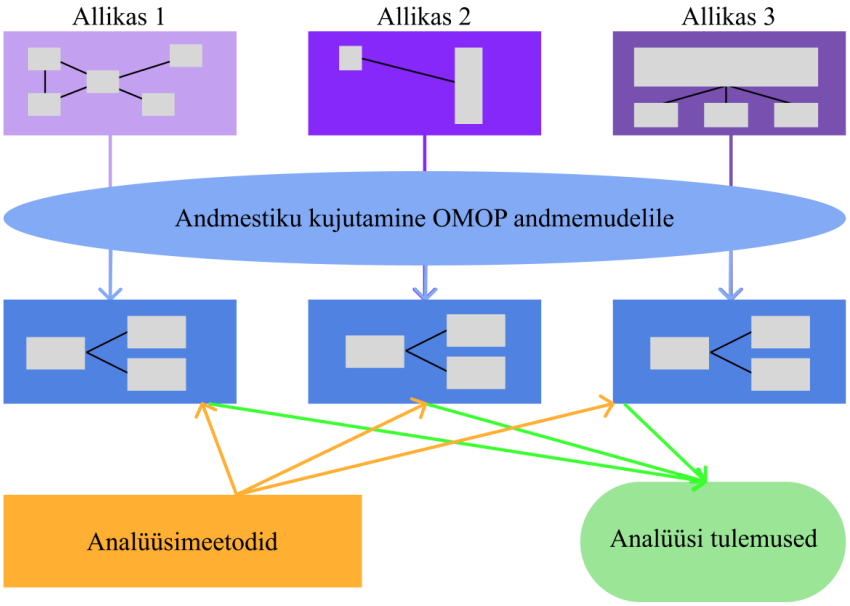
Bakalaureusetöö koosneb kolmest peatükist. Esimeses antakse ülevaade OMOP andmemudelist, *Observational Health Data Sciences and Informatics* (OHDSI) kogukonnast ning kirjeldatakse senist terviseandmete ühtlustamise ja visualiseerimise protsessi Eesti kontekstis. Teises peatükis on kirjeldatud kasutatud andmestikke ja nende rakendamist käesolevas töös. Viimases peatükis kirjeldatakse autori loodud rakenduse arhitektuuri ja toimimist ning analüüsitakse edasiarendamise võimalusi. Töö lisa I sisaldab linki programmi lähtekoodile ning lühikirjeldust, milline rakenduse versioon kasutamiseks valida.

2. Taustainfo

Selleks, et efektiivselt rahvastiku tervist edendada ja tõenduspõhiseid otsuseid teha, on tarvis kasutada andmete kogumiseks ja hoiustamiseks standardiseeritud meetodeid, muutes andmete analüüsimist kättesaadavamaks ja paremini tõlgendatavaks. Ühtlustatud ja laialdaselt kasutatav andmemudel on sobiv alus võimaldamaks üksikute visualisatsioonikomponentide korduvkasutust erinevate andmestike korral. Järgnevalt antakse ülevaade OMOP standardiseeritud andmemudeli ja OHDSI olemusest ning nende senisest rakendamisest Eestis.

2.1 Ülevaade OMOP andmemudelist

Üheks terviseandmete analüüsimisel kasutatavaks standardiseeritud andmemudeliks on Observational Medical Outcomes Partnership (OMOP) koostööprojekti raames arendatud ühtne andmemudel (ingl k *common data model*). OMOP CDM määratleb nii kasutatava andmemudeli vormingu kui ka standardiseeritud terminoloogiad, võimaldades kasutada konkreetset analüüsimeetodit korraga tervel hulgal eri OMOP CDM kujul olevatel andmestikel (Overhage jt, 2011). Jooniselt 1 selgub, et selline ühtlustamine lahendab andmete heterogeensuse probleemi — muudab eri allikatest pärinevad terviseandmed omavahel võrreldavateks ja analüüsitavateks ühises kontekstis.



Joonis 1. Andmemudel võimaldab eri struktuuriga andmestikud ühtsele kujule teisendada. Algallikast tõlgitud (OHDSI, 2025).

Ajakirjas Eesti Arst avaldatud artikli kohaselt on OMOP andmemudeli peamiseks eesmärgiks salvestada isikupõhiselt võimalikult palju fakte ja sündmusi, seejuures järgides rangelt kindlaksmääratud struktuuri. Artiklis on lisaks välja toodud, et terminoloogiasõnastikena kasutatakse diagnooside märkimisel peamiselt SNOMED CT ja ravimite puhul RxNorm sõnastikke (Reisberg jt, 2024). SNOMED CT on rahvusvaheline, enam kui 340 000 mõistet hõlmav kliiniline terminoloogia. Selle ülesehitus võimaldab üheselt määratleda diagnoose, protseduure ning kliinilisi leide. RxNorm ravimisõnastik normaliseerib toimeaine tugevuse ja ravivormi kombinatsioone ning seob neid püsivate tunnustega (Bodenreider jt, 2018). Allika järgi toetab OMOP mudel ka automaatseid teisendusi eelnimetatud sõnastike peale rahvusvahelise haiguste klassifikatsiooni (RHK) ja toimeainete klassifikatsiooni (ATC) sõnastikest (Reisberg jt, 2024).

2.1.1 OHDSI kogukond

Aastal 2018 sai alguse Euroopa Terviseandmete ja -tõendite võrgustiku programm (EHDEN), mille eesmärk oli arendada välja ühtne üleeuroopaline andmeplatvorm, mis oleks standardiseeritud OMOP CDM baasil ja võimaldaks terviseandmete alast laialdasemat koostööd. Programmiga liitusid 187 andmepartnerit 29 riigist üle terve Euroopa. Paralleelselt osalesid tehnoloogiat arendavate partneritena 64 väike- ja keskmise suurusega ettevõtet. Programmis olid esindatud nii Tartu Ülikool kui ka Eesti ettevõtted STACC ja Quretec. Programmis osalemisega anti ulatuslik panus nii OMOP andmemudeli populariseerimisele kui ka terviseandmete ühtsele kujule viimisele. EHDEN projekti raames rahastati omakorda veel *Observational Health Data Sciences and Informatics* (OHDSI) võrgustiku jätkusuutlikkust (EHDEN Foundation, 2022). Ryan ja Hripcsak on oma raamatus kirjeldanud OHDSI võrgustikku kui terviseteadustele pühendunud rahvusvahelist kogukonda, mille eesmärgiks on utiliseerida avatud lähtekoodiga tarkvara parendamaks inimeste tervist ja heaolu. OHDSI kasvas välja OMOP projektist ja tugineb seejuures OMOP andmemudelile. OHDSI kogukond ühendab eri ekspertiisiga teadlasi ja spetsialiste, toetades kestvat innovatsiooni ja edasiarenguid terviseuuringute valdkonnas läbi andmeteaduse ja informaatika. OHDSI kommuunis on uurimuste läbiviimisel olulisteks tööriistadeks nii ühtses formaadis andmete talletamine kui ka kohortide kasutamine (Ryan ja Hripcsak, 2021). Arendamaks koostööd ja edendamaks terviseandmete analüüsi Eestis, loodi aastal 2023 Sulev Reisbergi juhtimisel ka OHDSI Eesti haru (OHDSI Europe, 2017).

2.2 Terviseandmestike ühtlustamine Eestis

Eestis on mitmeid keskseid riiklike andmekogusid ja eri tervishoiuteenuste infosüsteeme. Nendes süsteemides olev andmestik sisaldab terviklikku infot patsientidest ning üldisest olukorrast, kuid nii Eesti-sisestes kui ka rahvusvahelistes uuringutes osalemine on siiski keeruline. Seda eelkõige seetõttu, et andmete omavahel kokkuviimist ja töötlemist on tehniliselt raske teostada tänu erinevustele semantikas ja struktuuris, mida nende eri andmekogude loomisel on kasutatud (Solvak jt, 2022). Kui OMOP CDM puhul on juhtivaks SNOMED CT terminoloogiastandard, siis Eestis tuginevad terviseandmed valdavalt rahvusvahelise haiguste klassifikatsiooni kümnenda versiooni (RHK-10[1]) terminoloogiastandardile.

Aastatel 2019–2022 läbi viidud RITA teadusprogrammi projekti "Masinõppe ja AI toega teenused (MAITT)" tarbeks loodi andmebaas, mis sisaldas 10% juhuvalimit kõigist Eesti isikukoodiga isikute andmetest kolmest riiklikust terviseandmekogust. Projekti käigus loodi struktrueeritud ja puhastatud andmete jaoks korduvkasutatavaid teisendusskripte, mis algandmeid automaatselt ühtsele OMOP andmekujule viiksid. Projekti raames uuritavates andmekogudes olid esindatud väljakirjutatud ravimite info retseptikeskuses, raviarved Eesti Haigekassa andmekogus ja terviselugude kokkuvõtted ning laborianalüüside tulemused Tervise Infosüsteemis. Ühendatud andmestik teisendati ühtsele OHDSI OMOP CDM andmekujule, misjärel see sai hoiustatud Tartu Ülikooli teadusarvutuskeskuse spetsiaalses turvakeskkonnas. Loodud andmestikus esinesid pseudonüümitult 20.7 miljonit erinevat tervisedokumenti ehk terviseandmeid 150 811 patsiendi kohta perioodist 2012–2019 (Solvak jt, 2022).

Harry-Anton Talvik on oma magistritöös kirjeldanud tervisedokumentide üldist teisendamist nende transpordikujult ühtse andmemudeli (OMOP CDM) kujule, mis võimaldab efektiivsemat andmete analüütilist kasutust (Talvik, 2022). Selle standardiseerimisprotsessi kirjeldamine oli Eesti terviseandmete ühtlustamisel ja nii siseriikliku kui ka võimaliku rahvusvahelise koostöö jaoks oluline samm. Aastal 2023 avaldas Tartu Ülikooli terviseinformaatika uurimisrühm Talviku magistritööle tuginedes ka eraldiseisva publikatsiooni kirjeldamaks metoodikat Eesti terviseandmete üleviimisel OMOP CDM andmemudelile. Töös püstitatud eesmärkide demonstreerimise tarbeks kasutati eelnevalt kirjeldatud MAITT andmestikku. Artikli kohaselt on varasemad uuringud ulatuslikult

[1] https://rhk.sm.ee/

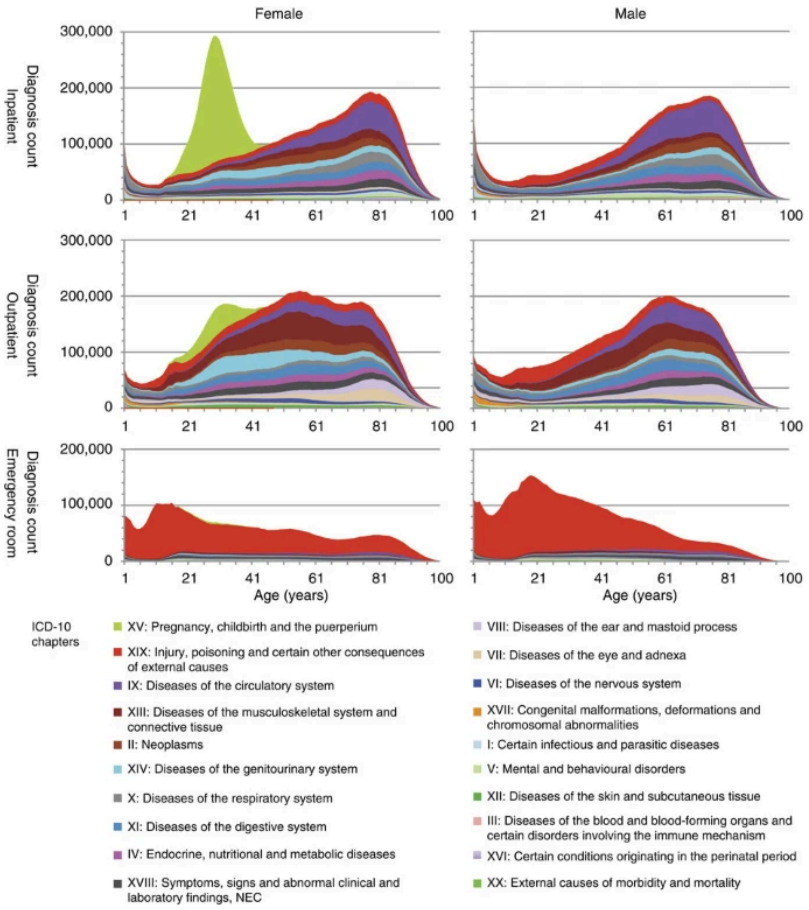
kirjeldanud andmebaaside ümberkujundamist OMOP CDM kujule. Aastal 2023 oli tänu OHDSI kogukonnale viidud Euroopas rohkem kui 453 andmebaasi üle OMOP CDM kujule. Koostöö edendamiseks ja ühtsel kujul olevate andmebaaside hulga suurendamiseks rõhutatakse artiklis andmebaaside ümberkujundamise protsessi jätkuva jagamise ja kirjeldamise olulisust (Oja jt, 2023).

2.3 Varasemalt loodud lahendused

Eesti suurim tervisega seotud avalik andmekogu on Tervise Arengu Instituudi (TAI) poolt hallatav Tervisestatistika ja terviseuuringute andmebaas. TAI tegeleb riiklikul tasemel tervisestatistika loomisega, haiguste ennetamise programmide ja tervisearendusega. Aastate jooksul on TAI poolt loodud arvukalt eri riiklikul tasemel olulisi ülevaatlike andmebaase ning visualisatsioone (Tervise Arengu Instituut, 2022). Tervise Arengu Instituudi andmebaas sisaldab laialdast valikut eri andmebaasidest, kuid puudub ühtne visualisatsioon, mis annaks kasutajale hoomatava ülevaate arstiabi saanute arvust ja kuludest kõikide põhidiagnooside lõikes.

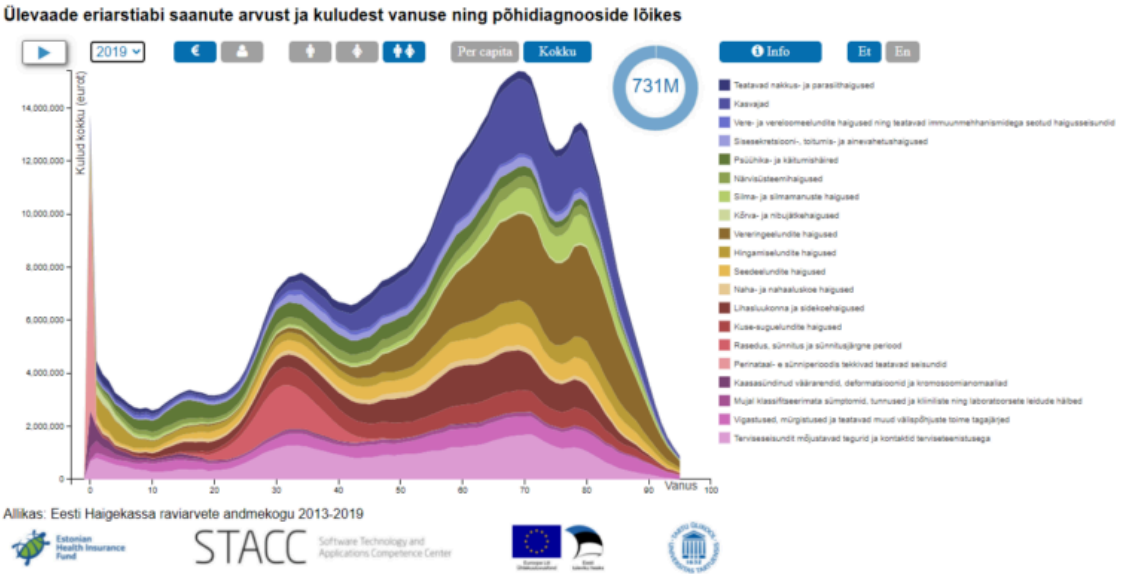
Aastal 2016 lõi tarkvaraettevõttes STACC Tormi Reinson interaktiivse graafiku, mille eesmärk on anda ülevaade eriarstiabi saanute arvust ja kuludest vanuse ning põhidiagnooside lõikes Eestis. Käesoleva töö juhendaja Jaak Vilo sõnul oli lahendus inspireeritud ajakirjas Nature Communications aastal 2014 ilmunud artikli visualisatsioonidest, mis on kujutatud joonisel 2. Artiklis kirjeldatakse kõnealuse visualisatsiooni abil haiguste esinemisi Taani näitel. Joonisel on kujutatud statsionaarse, ambulatoorse ning erakorralise meditsiini osakondade vahel jaotunud haiguste sagedused kasutades agregeerimiseks RHK-10 diagnoosirühmasid (Jensen jt, 2014).

Figure 1: ICD-10 diagnoses from the National Danish Patient Registry covering the entire Danish population in the period 1996–2010.



Joonis 2. Haiguste esinemine Taanis ravivormide ja diagnoosikategooriate järgi. Kuvatõmmis artiklist (Jensen jt, 2014).

Eeltoodud visualisatsioonist inspireerituna loodud interaktiivse graafiku kasutaja saab valida, kas visualiseerida andmeid raviarvete kulude või patsientide arvu järgi. Joonisel 3 kujutatud graafik võimaldab huvilisel andmeid filtreerida läbi mitmete tunnuste, nagu näiteks aasta, diagnoosikategooriad ja sugu. Lisaks saab valida, kas esitleda kulud patsientide vanuste lõikes kokku summeeritult või ümber arvutatult ühe elaniku kohta. Ettevõtte STACC poolt loodud lahenduse puhul on vaadeldavad andmed perioodil 2013–2019 ning andmete allikaks on Eesti Haigekassa (alates 2023. aasta kannab nime Tervisekassa) raviarvete andmekogu (Reisberg, 2016).



Joonis 3. Kuvatõmmis Tormi Reinsoni arendatud lahendusest[2] eriarstiabi saanute arvu ja kulude visualiseerimiseks (Reisberg, 2016).

Terviseandmete kogumisel kasutatakse üldjuhul erinevaid eelnevalt kindlaks määratud klassifikaatoreid või standardeid, mis võivad erineda nii kasutatavate tarkvarade kui ka tervishoiuteenuse osutajate vahel. Teadur Tormi Reinsoni poolt arendatud tööriist on loodud Tervisekassa eriarstiabi raviarvetele tuginedes. Praegu on enamus andmeid juba kohandatud OMOP CDM andmebaasidesse, mistõttu ongi tarvis luua lahendus, kuidas sarnast visualisatsiooni saaks rakendada iga OMOP andmebaasi peale.

Teine väljakutse lisaks OMOP andmebaasile visualisatsiooni lahenduse arendamisele on patsientide andmekohortide ajalise mõõtme kujutamine. Kui andmed on üheaastases vaatlusaknas, siis saab kokku lugeda konkreetse vanusega patsientide diagnooside sagedused. Kuid kuna OMOP andmebaas sisaldab infot üle mitmete aastate, siis tekib küsimus, kuidas andmekohorte pikema perioodi jooksul selgelt visualiseerida. Võimalikud lähenemised oleksid, kas liikuda läbi aja aasta kaupa, kuvada kõik andmed üheaegselt või leida lahendus, mis võimaldaks korraga visualiseerida nii kohordi jälgimise kestust kui ka patsientide vanust.

Autor võttis loodava rakenduse lähtepunktiks äsja kirjeldatud interaktiivse graafiku. Iga OMOP andmebaasi peale rakendamiseks saab rakendus sisendiks eelagregeeritud kindlas vormis tabeli. Pikema perioodi visualiseerimiseks anname kasutajale valiku, kas kuvada andmed üle valitud perioodi kokku summeeritult või valitud aastate omavahelise võrdlusena.

[2] https://stacc.ee/ehif-stacked-area/

3. Kasutatud andmestikud ja metoodika

Käesolevas peatükis kirjeldame lahenduse arendamisel kasutatud kahte andmestiku ja nende rakendamist andmete visualiseerimisel. Peatüki esimeses osas tutvustame Sulev Reisbergi poolt kasutatud metoodikat sünteetilise andmestiku loomiseks. Seejärel anname ülevaate turvalise andmeanalüüsi keskkonnast SAPU ning seletame lahti andmestikest andmete väljalugemise protseduuri, mis on tarvilik saadud andmete kasutamiseks käesolevas töös.

3.1 Tehisandmed

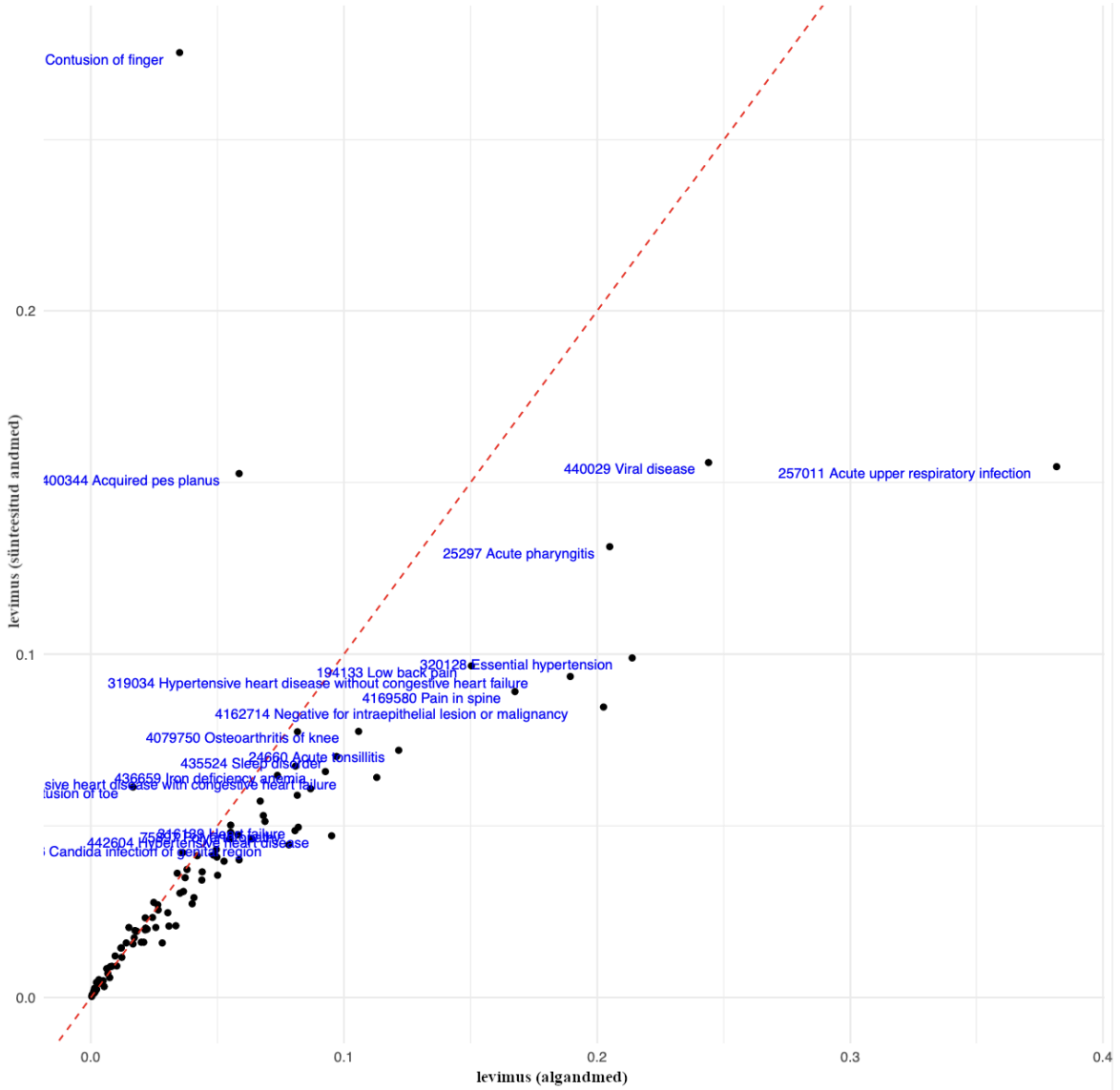
Tehisandmetega täidetud andmestik on kiiret juurdepääsu pakkuv, kuid seejuures ka patsientide anonüümsust säilitav vahelüli, mis hõlbustab lahenduste väljatöötamist ning koodi ja meetodite valideerimist. Peatükki kirjutades on lähtutud andmestiku dokumentatsioonis kirjeldatud metoodikast ja loodud andmestiku omadustest (Reisberg, 2024).

Lähtematerjalist, milleks oli RITA MAITT projekti andmestikust 10% juhuvalim aastavahemikust 2012–2019, võeti kõigi isikute ja 100 juhusliku haiguse statistika. Loodud statistika näitab, kui tõenäoline on mingi haiguse avaldumine eri vanuse-soo gruppides ning milline on haiguse kordumise muster ajas. Iga valitud haiguse kohta võeti arvesse haiguse esmaesinemise hetkel isiku vanus ja sugu. Piiratud ajaakna tõttu loeti haiguse esmakordseks esinemiseks juhtumit, kus isikul diagnoositi uus haigus ning haiguse tekkimise momendist arvestatult oli sama isiku kohta vähemalt kahe aasta jagu varasemaid andmeid või oli isik sündinud alates aastast 2012. Kui haigus oli tuvastatud esmakordselt, siis võeti vaatluse alla ka selle võimalik taastekkimine järgnevatel aastatel. Kuna RITA MAITT andmestik hõlmab maksimaalselt seitsme aasta pikkust ajaperioodi, siis on ka haiguste korduvuse info piiritletud vaid ajavahemikus 4–7 aastat.

Isikute ja haiguste statistika põhjal koostati mudel, mis võtab kokku RITA MAITT andmestiku reaalsete isikute haiguste avaldumise ja korduvuse sagedused, kuid säilitab isikute konfidentsiaalsuse. Mudeli koostamisel siluti haiguste esinemise sagedusi ja mustreid eemaldades haruldasemad soo- ja vanusegruppide kombinatsioonid. Selle tulemusena sai Reisberg üldisema, kuid tõepärase esinemissageduste maatriksi, millest ei ole võimalik üksikisikuid tuvastada. Saadud statistilise mudeli puhul puudub haigustevaheline korrelatsioon muude tegurite kui ainult vanuse ja soo lõikes. Ehk kui mudelis tuvastatakse inimesel üks haigus, siis see ei tõsta ega alanda teiste haiguste saamise tõenäosust peale selle, mis tuleneb vanuse ja soo jaotusest.

Seejärel sünteesis Reisberg mudeli isikute statistika põhjal 10 000 uut juhuslikku "inimest", kelle sugu ja vanus andmete algusaastal vastas ligikaudselt sama perioodi vanus-soo jaotusele. Igale sünteesitud isikule omistati eelnevast juhuslikust 100 haiguse mudeli statistikast haiguste esinemised ajavahemikus 2012–2019. Kuna haiguste omistamisel lähtuti eelnevalt koostatud mudelist, siis saadi haiguste esinemine järgnevast lähtudes: kui mudeli alusel isikul mõni haigus esines näiteks mitmel aastal järjest, siis lisati sünteesitud isikule uuesti sama haigus vastavatele aastatele; kui mudeli kohaselt on konkreetsele vanusele ja soole kõrge haiguse esmaesinemise tõenäosus, siis see peegeldub ka suurema esinemissagedusena sünteetilises populatsioonis. Kuna mudel põhines reaalsetel esinemissagedustel, siis on haiguste esinemissagedus sünteetilises andmestikus sarnane RITA MAITT algandmete üldisele trendile. Joonisel 4 näeme, et haiguste levimus kahe andmestiku vahel on võrdlemisi ühtlane. Loodud andmestikus on pealtnäha vaid lapseeas kogetavate haiguste levimus ebakorrektne. Reisbergil ei õnnestunud seda lahendada ilma ülejäänud andmestikus soovimatuid muudatusi põhjustamata. Sünteesitud isikud, haigused ja esinenud haigusjuhtude kirjed pandi kokku DuckDB[3] andmebaasiks. Tulemuseks saadi andmestik, mis sisaldas 10 000 inimest ja 100 erinevat diagnoosi, ning mille patsientide sooline-vanuseline jaotus järgib RITA MAITT uuringu andmestikku.

[3] https://duckdb.org/



Joonis 4. Diagnooside levimuse võrdlus sünteesitud andmete ja algandmete vahel (Reisberg, 2024).

Käesoleva töö autor kasutas Reisbergi loodud andmebaaasi versiooni 1.1. Erinevus võrreldes versiooniga 1.0 seisneb selles, et kõikidel andmebaasis esinevates tabelites on kõik OMOP CDM v5.4 standardis kirjeldatud veerud lisatud ning üleliigsed veerud on eemaldatud. Kumbki tehisandmetega täidetud andmestik ei sisaldanud endas patsientide visiitide kulusid. Seega rakenduse versioon, mis kasutab just seda tehisandmestikku — visiitide kulude visualiseerimist ei võimalda. Tehisandmestiku kompaktsus võimaldas töö autoril lugeda andmed edasiseks töötluseks mugavalt otse andmebaasist, kasutades Pythoni DuckDB teeki.

3.2 Töö reaalandmetega

Sensitiivsete andmete uurimise ja töötlemise eelduseks on alati ka relevantsete eetikakomiteede load (Solvak jt, 2022). Käesoleva bakalaureusetöö puhul võimaldati terviseandmetele ligipääs 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. Projektide läbi viimiseks on Tartu Ülikooli eetikakomitee luba nr 300/T-23 ning Eesti Bioeetika ja inimuuringute nõukogu luba nr 1.1-12/3088, pikendamise otsusega nr 1.1-12/613.

Käesoleva töö rakendamiseks reaalandmetel oli tarvilik läbida infotund, kus tutvustati tundlike andmete analüüsiplatvormi (SAPU) ning RITA MAITT andmestiku kasutamist turvakeskkonnas. Lisaks tuli autoril sõlmida konfidentsiaalsusleping Tartu Ülikooliga ja saada heakskiit eetikakomiteelt. Olles saanud vajaliku eetikakomitee loa, ligipääsu turvakeskkonda ja valideerinud lahenduse algversiooni toimimist sünteetilise andmestiku peal, oli lahendus valmis rakendamiseks reaalsetel andmetel. Selles alapeatükis annab töö autor esmalt ülevaate SAPU andmeanalüüsi keskkonnast ning seejärel kirjeldab rakenduse toimimiseks vajaliku materialiseeritud vaate loomise loogikat.

3.2.1 Turvaline andmeanalüüsi keskkond SAPU

Tundlike andmete analüüsiplatvorm (SAPU) on Tartu Ülikooli Teadusarvutuste keskuse (HPC) poolt pakutav turvaline keskkond tundlike andmete töötlemiseks. Andmetöötlus SAPU platvormi vahendusel pakub standardklastritel põhinevast lähenemisest rohkemat turvalisust vähendades riski, et tundlikke andmeid volitamata kopeeritakse või edastatakse. Selle saavutamiseks on SAPU täielikult võrgust isoleeritud, ligipääs toimib ainult virtuaalse töölaua kaudu ja kõik masinas teostatavad tegevused salvestatakse serveripoolses monitooringukihis. Samuti liigutatakse masina siseselt faile vaid objektipõhise salvestusruumi kaudu ning andmete masinast välja liigutamine nõuab andmeomaniku ülevaatamist ja nõusolekut. Lisaks isoleeritusele pakub SAPU ka sensitiivsete andmete töötlemiseks vajalikke eelvalideeritud protseduure ja reeglistiku andmete käsitlemiseks (Tartu Ülikool, 2024). SAPU keskkonnas võimaldati töö autorile ligipääs eelnevalt kirjeldatud RITA MAITT projekti raames loodud 10% juhuvalimit sisaldavatele andmestikele.

Rakenduse kasutamiseks SAPU-s asuval reaalandmestikul tuli töö autoril teha veel mõningane eeltöö. Kuna rakendus arendati algselt vaid tehisandmestikule tuginedes siis SAPU-s asuv reaalandmestik polnud rakenduse ülesehituse poolest koheselt toetatud tänu

tehisandmestiku ja reaalandmestiku vahelistele erisustele. Peamisteks erisusteks on andmebaasi tüüp ja kulude olemasolu andmestikus. Kui rakenduse kasutaja soovib visualiseerida andmeid mõnest teisest andmebaasist, siis selleks on tarvilik eelnevalt luua rakenduse tingimustele vastav tabel või materialiseeritud vaade. Loodud tabel peaks sisaldama eeltöötluse käigus kokku arvutatud sagedusi.

3.2.2 Sobiva struktuuriga tabeli loomine

OMOP CDM andmemudel sisaldab rohkelt dateeritud isikupõhist teavet. Kirjeldatud on nii mõõtmiste tulemused, teostatud protseduurid, informatsioon ravi teostaja kohta ning palju muud. Kuna OMOP andmemudeli puhul on autori poolt loodud rakenduse toimimiseks vajalik informatsioon eri tabelites ning andmeid on palju, siis rakendus aktsepteerib sisendina kindla formaadiga eelnevalt agregeeritud üksikut tabelit.

Töö autor kasutas andmestiku skeemiga tutvumiseks andmebaasirakenduse DBeaver graafilist kasutajaliidest ning uue tabeli loomiseks sama rakenduse SQL konsooli. Rakenduse kasutajal tuleb enne programmi käivitamist veenduda sobiva vaate või tabeli olemasolus. Töö autor lõi selle tarbeks kogu andmestiku pealt ühtse materialiseeritud vaate. Loodud SQL päring koosneb mitmest alampäringust ning ühest kokkuvõtvast päringust. Töö autor kasutas pika päringu paremini hallatavateks osadeks jaotamisel tavalisi ajutisi tabeliavaldisi (*common table expression*). Neist esimese *visit_info* eesmärgiks on kohordi moodustamiseks siduda iga patsiendi identifikaator isiku puhul esinenud visiitide infoga tuues välja nii visiidi toimumise aja, patsiendi vanuse ja soo.

```
SELECT vd.visit_occurrence_id, vd.person_id, vd.visit_start_date,
    EXTRACT(year FROM vd.visit_start_date) AS year,
    DATE_PART('year', AGE(vd.visit_start_date,
          MAKE_DATE(p.year_of_birth,
                COALESCE(NULLIF(p.month_of_birth, 0), 1),
                COALESCE(NULLIF(p.day_of_birth, 0), 1)))) as age,
    p.gender_concept_id
FROM ohdsi_cdm_202503.visit_occurrence vd
JOIN ohdsi_cdm_202503.person p ON vd.person_id = p.person_id
```

Loeme avaldise nimega *diagnosis_info* abil välja iga visiidi puhul määratud diagnoosikategooria, seejuures jättes välja topelt esinevad visiidi-diagnoosi paarid.

```
SELECT DISTINCT co.visit_occurrence_id,co.condition_concept_id AS diagnosis_concept_id
FROM ohdsi_cdm_202503.condition_occurrence co
```

Seejärel *costs* avaldise abil leiame iga visiidi kogukulu. Kui visiidil puudub kulu, siis asendatakse see nulliga, et kuludeta patsiendid agregatsioonist välja ei jääks.

```
SELECT vd.visit_occurrence_id, SUM(c.total_paid) AS total_cost
FROM ohdsi_cdm_202503.visit_detail vd
    JOIN ohdsi_cdm_202503.cost c ON vd.visit_detail_id = c.cost_event_id
GROUP BY vd.visit_occurrence_id
```

Avaldise nimega *combined* abil kombineerime üheks andmestikuks eelnenud ajutiste tabelipäringute tulemused.

```
SELECT vi.visit_occurrence_id, vi.person_id, vi.visit_start_date, vi.year, vi.age,
vi.gender_concept_id, di.diagnosis_concept_id, COALESCE(c.total_cost, 0) as total_cost
FROM visit_info vi
    LEFT JOIN diagnosis_info di ON vi.visit_occurrence_id = di.visit_occurrence_id
    LEFT JOIN costs c ON vi.visit_occurrence_id = c.visit_occurrence_id
```

Saadud *combined* tabeli pealt loetakse lõpuks kokkuvõtva päringuna nii unikaalsete patsientide arv kui ka kogukulu iga eraldiseisva diagnoosinimetuse, aasta ning täisaastates külastaja vanuse ja soo väärtuste lõikes. Patsiendi ja diagnoosi lõikes liidetakse kogu sama kalendriaasta kulud ja visiidid ühe konkreetse aasta alla.

```
SELECT comb.year, comb.age, comb.gender_concept_id, comb.diagnosis_concept_id as diagnosis, COUNT(DISTINCT rw.person_id) as patient_count, SUM(rw.total_cost) AS total_cost

FROM

(SELECT DISTINCT year, age, gender_concept_id, diagnosis_concept_id FROM combined) comb

LEFT JOIN combined rw ON

    rw.visit_start_date BETWEEN MAKE_DATE(comb.year::int,1,1) AND (MAKE_DATE(comb.year::int,12,31))

    AND rw.age = comb.age

    AND rw.gender_concept_id = comb.gender_concept_id

    AND rw.diagnosis_concept_id = comb.diagnosis_concept_id

GROUP BY comb.year, comb.age, comb.gender_concept_id, diagnosis;
```

Kogu päringu tulemuseks saame tabelis 1 kujutatud vaatele sarnaneva agregeeritud tabeli, kus iga rida esitab konkreetse kalendriaasta, vanuse- ja soorühma ning diagnoosi X lõikes esinenud unikaalsete patsientide arvu koos neile osutatud tervishoiuteenuste kogukuluga Y eurot.

Tabel 1. Päringu tulemusena loodud tabeli struktuur fiktiivsete andmetega.

Aasta	Vanus	Sugu	Diagnoos	Patsientide arv	Kogukulu
2019	31	8,507	440,432	3,587	488,421.62
2020	58	8,532	320,128	4,850	2,126,323.36
…	…	…	…	…	…

Tabelis on sugu ja diagnoos kirjeldatud neile vastavate koodide väärtustena. Koodide tõlgendused on leitavad vabavaralises OHDSI töörühma poolt hallatavas tarkvaras ATLAS (Ryan ja Hripcsak, 2021). Tulbas “Sugu” on koodi 8,507 puhul tegemist mehega ning koodi 8,532 puhul on tegemist naisega. Tabeli esimesel real näitena toodud diagnoosikood 440,432 viitab hambakatule või hambakivile ning järgneval real olev diagnoosikood 320,128 viitab

kõrgvererõhktõvele. Seega tabelis toodud andmete põhjal oli hambakatu tõttu aastal 2019 unikaalseid meessoost 31-aastaseid patsiente 3587 kogukuludega 488,421.62 eurot.

Kui rakenduse kasutajal on ligipääs töö autori poolt SAPU keskkonnas loodud tabelile (*user_rasmus_mirma.rw_count_and_total_cost*), siis tuleb programmi .env failis lisada vaid oma andmebaasile ligipääsemiseks kasutatav kasutajanimi ning parool. Olles loonud eelneva kirjelduse järgi vajalikke andmeid sisaldava uue tabeli, tuleb kasutajal täpsustada programmi .env failis ka uue andmetabeli asukoht. Uue tabeli kokku arvutamine ülaltoodud kirjelduse järgi võttis töö autoril temale värskeima kättesaadavaks tehtud andmestiku (*ohdsi_cdm_202503*) puhul aega umbes neli minutit. Tabeli loomine enne programmi käivitamist on kasutajale ühekordne tegevus, kuid andmete muutmisel algsetes andmehulkades tuleks värskendada ka rakendusele sisendiks antavat tabelit. Äsja kirjeldatud päring on täismahus kättesaadav rakenduse repositooriumi *init_db_query.txt* nimelises failis. Tabeli töös rakendamiseks kasutas autor Pythoni psycopg2 (Gregorio jt, 2021) teeki.

4. Töölauarakendus: *VisitCountsAndCosts*

Käesolev peatükk keskendub lahenduse loomise protsessi, ülesehituse ja kasutusvõimaluste kirjeldamisele. Autor arendas rakenduse kasutades programmeerimiskeskkonda PyCharm (v3.7+) ning seejuures kasutades keskkonda integreeritud funktsionaalsust alustatud sõnade automaatseks lõpetamiseks. Peatüki esimene osa kirjeldab rakenduse eesmärki ning seejärel tutvustame rakenduse loomisel kasutatud tehnoloogiaid ja programmi pakutavaid võimalusi terviseandmete visualiseerimisel. Peatüki lõpus toome välja ülevaatliku näite rakenduse kasutamisjuhust ja võimalikud edasiarenguvõimalused.

4.1 Eesmärk

Käesoleva bakalaureusetöö käigus arendas töö autor interaktiivseks andmete kujutamiseks graafilise kasutajaliidese *VisitCountsAndCosts*[4]. Loodud töölauarakendus võimaldab andmeid visualiseerida ja analüüsida mistahes andmebaasist, kus saab välja tuua kohortide lõikes järgnevad veerud: sündmuse esinemise aasta, patsiendi vanus, sugu, diagnoosimõiste, patsientide arv ning vastavalt soovile ka kogukulu. Rakenduse arendamisel võeti arvesse eelnevalt loodud lahenduse kitsaskohti nagu eri andmestikega ühildumine ning ajalise mõõtme kujutamine. Olles sõlminud konfidentsiaalsuslepingu, sai töö autor luua rakendusele OMOP CDM kujul olevatest andmestikest sobiva sisendi loomiseks vajaliku, peatükis 3.2.2 kirjeldatud skripti. Tänu ligipääsule reaalandmestikule sai töö autor valideerida ja arendada rakendust pidades silmas eelkõige OMOP CDM kujul olevaid andmestikke. Töölaud annab kasutajale ülevaate nii eriarstiabi saanute arvust kui ka sellega seotud kuludest ning võimaldab uurida ka täpsemalt huvipakkuvaid diagnoosikategooriaid ja kohorte. Töölaud loodi eesmärgiga täiendada Eesti andmeteadlaste olemasolevate töövahendite ampluaad kuid anda panus ka OHDSI rahvusvahelise kogukonna arengule ning sellest tulenevalt on nii rakenduse pealkiri, dokumentatsioon kui ka töölaud ise inglisekeelsed. Rakendus on kõigile huvilistele koos lähtekoodiga kättesaadav GitLab keskkonnas.

4.2 Rakenduse ülesehitus

OHDSI võrgustiku siseselt on varasemalt loodud tarkvarapakette enamasti just vabavaralises programmeerimiskeeles R. Kuigi ka see töö oleks olnud teostatav R-paketina, siis käesoleva töö lahenduse loomisel osutus valituks hoopis programmeerimiskeelt Python. Python on rohkete võimalustega vabavaraline universaalkeel, mis pakub rohkelt võimalusi nii

[4] https://gitlab.cs.ut.ee/rasmusmi/visitcountsandcosts

interaktiivsete töölauarakenduste arendamiseks kui ka andmeanalüüsi teostamiseks (Muddana ja Vinayakam, 2024). Rakenduse ülesehitus ja kasutatud tehnoloogiad muudavad selle hõlpsasti laiendatavaks ja kohandatavaks vastavalt võimalike edasiarendusi silmas pidades. Rakenduse avavaade annab kasutajale laia ülevaate arstiabi saanute arvust mõlema soogrupi kohta kogu andmestikus olemasolevate aastate suhtes akumuleeritult.

Töö autor lõi rakendusest kaks eraldiseisvat versiooni, mis on rakenduse repositooriumis eraldatud eri harudesse. Algne lahenduse arendamine toimus reaalandmetele ligipääsu puudumise tõttu sünteetilisel andmestikul, mis oli hoiustatud DuckDB andmebaasi. Enne töövoo alustamist tuleb kasutajal selgeks teha, kas tal on võimalik kasutada rakendusega koos reaalandmeid või soovib ta kasutada rakendust etteantud sünteetiliste andmete kujutamiseks. Vastavalt sellele tuleb valida töö repositooriumis haru *main* või *demo_dataset*. Töövoo alustamiseks tuleb kasutajal rakendus lokaalselt omale sobivas keskkonnas seadistada. Reaalandmestiku kasutamise puhul tuleb muuta andmebaasi väärtusi .env nimelises failis.

Kuna nii teisendusi sisaldav fail kui ka visualiseerimise tarbeks kasutatavad andmestikud on mahukad, siis veebilehe avanemisel tuleb kasutajal esmalt oodata, kuni vajalikud komponendid on kuvamiseks valmis laetud. Programmi ülesehitus on kasutajale intuitiivne ning selle kasutamine on väga lihtne. Programmi käivitamisel avanenud veebilehe vasakus servas on kuvatud kõik filtreerimisvalikud ning kogu ülejäänud ekraanialal kuvatakse filtrite tingimustele vastav visualisatsioon.

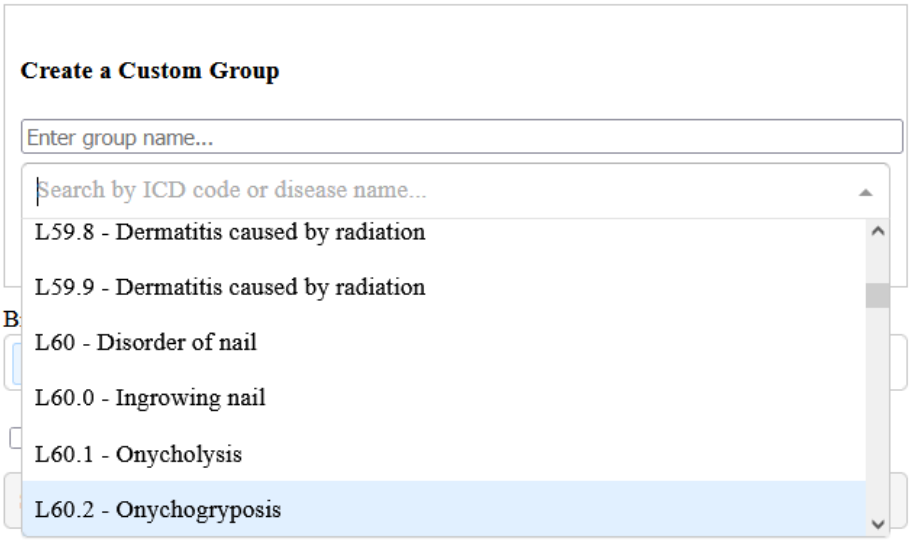
4.2.1 Filtersüsteem

Filtrites on kasutajal mitmeid valikuid, kasutaja saab lehe ülaosas valida, kas visualiseerida unikaalsete patsientide visiitide arvu või visiitide kogukulusid diagnoosikategooriate lõikes. Visiitide arvu ja kogukulude filtrivalik on tehisandmestikku kasutaval rakenduse osal kulude visualiseerimiseks vajalike andmete puudumise tõttu desaktiveeritud. Kasutaja saab valida, millised diagnoosikategooriad vaatluse alla võtta valides rippmenüüst omale meelepärased diagnoosikategooriad. Diagnoosikategooriatel on kaks eraldi astet, laiemad (ingl k *broad category*) ning kitsamad diagnoosirühmad (ingl k *subcategory*).

Rakenduse diagnoosinimede ja -koodide ning diagnoosikategooriate kaardistamiseks kasutas töö autor Tartu Ülikooli terviseinformaatika töörühma poolt loodud RHK-10 diagnooside vastendust SNOMED koodideks ja diagnoosinimedeks GitHub repositooriumist EstonianOMOPMappings (Terviseinformaatika uurimisrühm, 2024). Töö autor eemaldas olemasolevast vastendusi sisaldavast failist oma loodud rakenduse jaoks üleliigsed veerud,

diagnoosid, millel puudusid SNOMED koodideks vastendused ning read, mille puhul esinesid SNOMED koodides duplikaadid. Seejärel lisas iga diagnoosi (n=8822) juurde laiema ja spetsiifilisema rühma RHK-10 standardi põhjal. Seejärel teisendati vajalike vastendusi sisaldav fail ümber JSON-andmeformaati selle tõhusaks rakendamiseks programmis.

Lisaks eeldefineeritud kategoriseerimisele on töö autor implementeerinud ka kohandatud kategooriate loomise. Joonisel 5 nähtub, et kasutajal on võimalus valida endale sobivad diagnoosid ning nimetada nad uude kategooriasse. Rippmenüü valikutes on esindatud kõik diagnoosid, mis on olemas eelnevalt kirjeldatud vastendusfailis. Kohandatud kategooriad lisatakse lai-kategooriate rippmenüüsse, võimaldades neid kasutada koos eeldefineeritud kategooriatega.



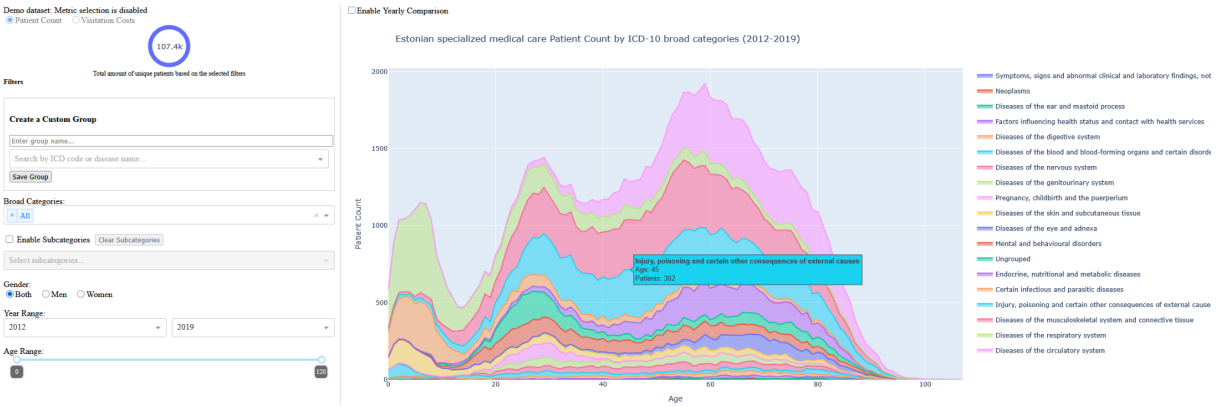
Joonis 5. Uue diagnoosirühma loomine RHK-10 koodide ja diagnoosinimede põhjal.

Visualiseeritava andmehulga täpsustamiseks on kasutajal filtreeritavateks elementideks patsientide soo, vanuse ning huvipakkuva ajavahemiku valimine. Loodud filtreerimisvalikud annavad võimaluse andmeid kujutada kasutajale sobivatel tingimustel.

4.2.2 Visualiseeritud graafikutüübid

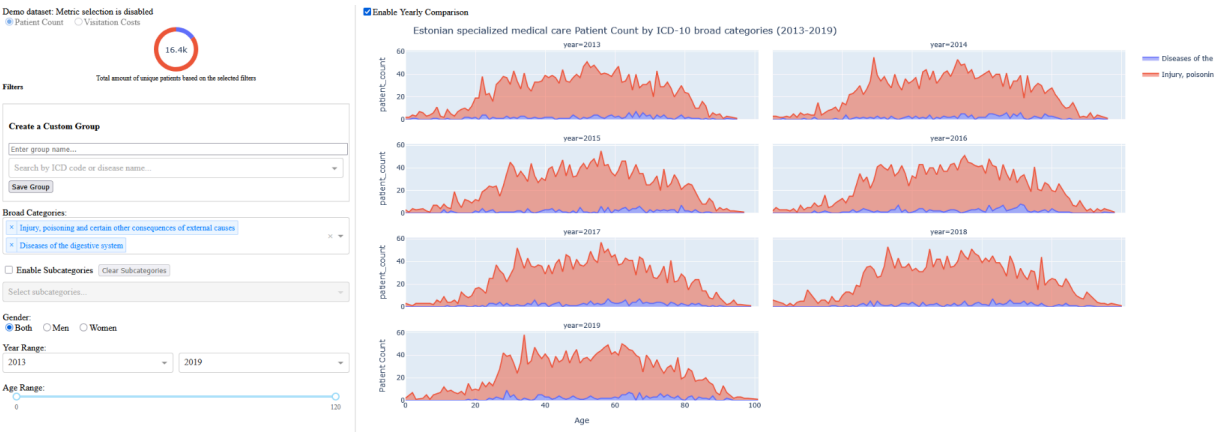
Eelnevalt peatükis 2.3 kirjeldatud lahenduse puhul visualiseeriti andmed virnastatud kihtdiagrammina iga üksiku aasta puhul eraldi. Käesoleva töö raames loodud rakendus kasutab andmete visualiseerimiseks kolme graafikutüüpi. Filtreerimisvalikute kohal asub sektordiagramm, mis kujutab valitud filtrite alusel visiitide arvu või kogukulu osakaalu kogu andmestikust. Peamiseks visualisatsiooni graafikutüübiks on kihtdiagramm. Esmalt on kasutajale kuvatud üksik graafik nagu joonisel 6, millel on kujutatud filtrite alusel andmed

üle valitud aastavahemiku kokku agregeeritult. Lisaks on kasutajal valik ümber lülituda kihtdiagrammide võrdluse vaatele, mis genereerib iga filtris märgitud aasta kohta eraldiseisva joonise üksikute aastate lõikes kõrvutatult.



Joonis 6. Rakenduse avavaade agregeeritud kihtdiagrammiga.

Kui sektordiagramm on kasutajale osakaalu kuvamiseks alati nähtav, siis kahe kihtdiagrammi vahel tuleb kasutajal teha valik, kas soovib andmeid agregeeritult visualiseerida või üksikute aastate kaupa nagu kujutatud joonisel 7. Selle tarbeks on visualisatsiooni kohal lüliti, mis muudab visualiseeringu tüüpi. Kihtdiagrammist vasakul asub joonise legend, mis kirjeldab esitletavad diagnoosikategooriad ja nende vastavad värvid. Rakenduse kasutaja saab hõlpsasti informatsiooni ka kursoriga joone peale liikudes, misjärel kuvatakse eraldi väljana joonele vastav diagnoosikategooria, vaadeldav vanuserühm ning sellele kohane patsientide arv või kulud vastavalt eelnevalt tehtud valikutele.



Joonis 7. Andmed on kujutatud iga vaatlusaasta lõikes eraldi.

Autor otsustas graafikul andmete kujutamisel, sarnaselt ettevõtte STACC lähenemisele, virnastatud kihtdiagrammide kasuks. Andmete virnastatus annab kasutajale hea ülevaate

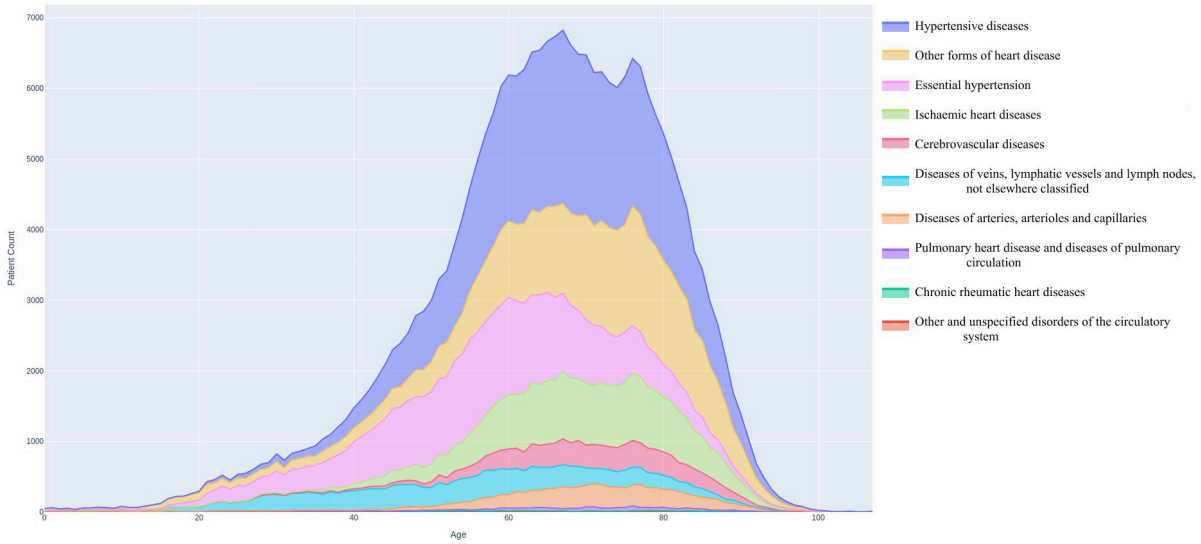
visiitide ning nendega kaasnevate kulude arvust kokku üle kõigi valitud diagnoosirühmade. Alternatiivseks variandiks oleks olnud visualiseerida üksteisest üle kattuvate joontena, mis oleks võimaldanud küll paremat diagnoosirühmade omavahelist võrdlust, kuid selleks, et saada ülevaade näiteks teatud vanuse kohal asuvate kahe või enama diagnoosirühma väärtustest kokku, peaks kasutaja ise vastavad väärtused kokku arvutama.

Loodud lahendus kasutab töölaua loomiseks ja graafikute kujutamiseks Pythoni Plotly Dash (Hossain, 2019) teeki. Andmete töötlemiseks on kasutatud teeki pandas (NumFOCUS, 2024). Pandas on NumPy mooduli põhjal loodud laialdaselt andmetöötluseks ja -analüüsiks kasutatav Pythoni pakett, mis pakub andmete loomiseks ja nende manipuleerimiseks kahte andmestruktuuri: seeria (ingl k *Series*) ja andmefreim (ingl k *Dataframe*). Seeria on vaid ühe veeruga andmeobjekt, kuid andmefreim on kahedimensiooniline andmeobjekt, mis võimaldab kasutada rohkemaid eri tüüpi veerge (Muddana ja Vinayakam, 2024). Rakenduse *VisitCountsAndCosts* puhul kasutati just andmefreimi andmebaasi materialiseeritud vaatest sisseloetud veergude rakendusesiseseks hoiustamiseks.

4.3 Näide rakenduse kasutamisest

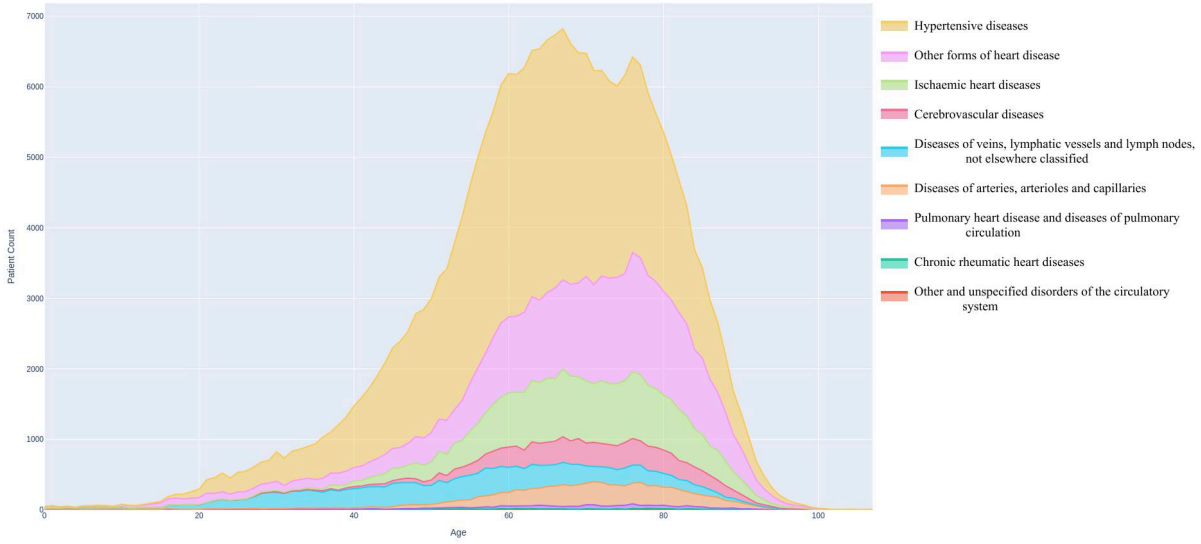
Käesolevas peatükis esitleme loodud rakendust. Toome välja ühe võimaliku kasutusjuhu kirjelduse ning näited rakenduse abil loodud visualisatsioonidest. Kasutusjuhu ilmestamiseks kasutame rakendust eelnevalt kirjeldatud reaalandmestiku *ohdsi_cdm_202503* peal. Väärib märkimist, et kasutatav andmestik representeerib Eesti rahvastikust vaid 10% juhuvalimit, kuid ettevõtte STACC poolt arendatud lahendus sisaldas kõiki raviarveid. Seetõttu pole siin esitletud kasutusjuhu tulemus otseselt võrreldav varasemalt loodud lahenduse tulemustega.

Oletame, et rakendust kasutavale isikule pakuvad huvi vereringeelundite haigused ja spetsiifilisemalt kõrgvererõhktõbi. Seega valime filtritest välja vereringeelundite haiguste kategooria (*Diseases of the circulatory system*) ja kuvame kõik selle alamkategooriad. Lisaks toome välja ka üksiku diagnoosina kõrgvererõhktõve, selleks saame luua uue kategooria, mis sisaldab vaid ühte diagnoosi. Sisestame otsingusse diagnoosi ingliskeelse nime *Essential hypertension* või kasutame RHK-10 klassifikatsiooni koodi I10. Kuna kategooria sisaldab vaid ühte diagnoosi, nimetame kategooria selles esineva diagnoosinime järgi — *Essential hypertension*. Joonisel 8 on kuvatud loodud filtrivalikutega visualisatsioon üle aastate 2013–2019, kus valikus on vaid meespatsiendid.



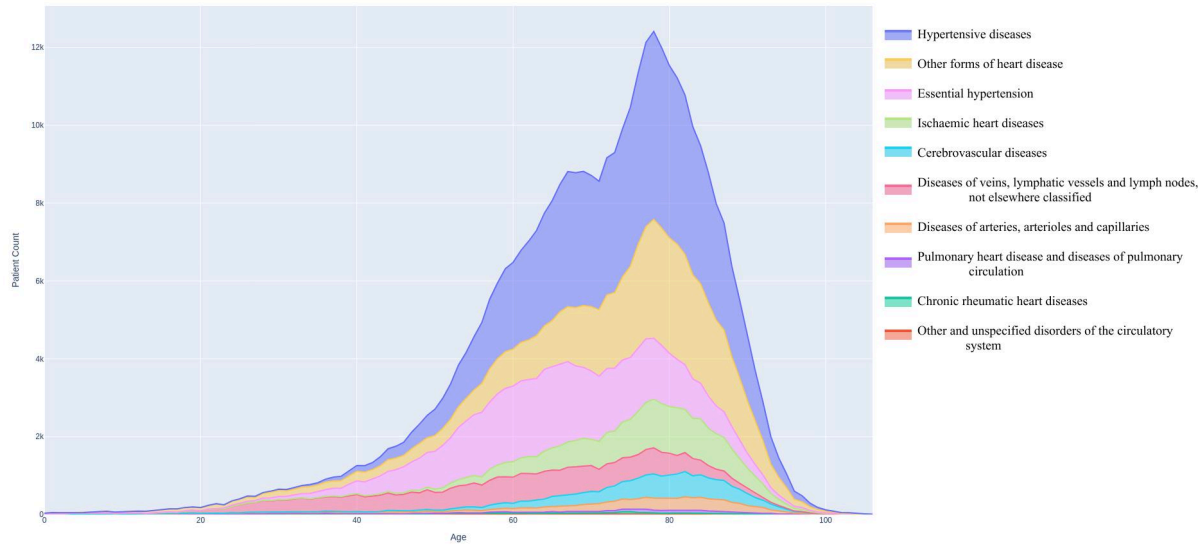
Joonis 8. Vereringeelundite haiguste esinemise arv meestel (2013–2019). Kuvatõmmis rakendusest.

Joonisel 9 on sama visualisatsioon, kuid filtritest on eemaldatud meie vastloodud diagnoosirühma kuvamine. Tuleb tähele panna, et kõrgvererõhkhaiguste (*Hypertensive diseases*) alamrühm ei sisalda diagnoosi koodiga I10 kui kategooriad on valitud visualiseerimiseks samaaegselt. Jooniseid 8 ja 9 võrreldes näeme, et diagnoos *Essential hypertension* on justkui ülejäänud kategooriast eraldi välja toodud. Seesugune lähenemine võimaldab lisaks eeldefineeritud kategooriate visualiseerimisele, kasutajal kujutada ka üksikuid diagnoose selle kategooriasse kuuluvusest hoolimata. Näeme, et kõrgvererõhktõve diagnoos moodustab tervelt kolmandiku kogu vereringeelundite haiguste diagnoosimistest.



Joonis 9. Vereringeelundite haiguste esinemise arv meestel ilma kõrgvererõhkhaiguste diagnoosita (2013–2019). Kuvatõmmis rakendusest.

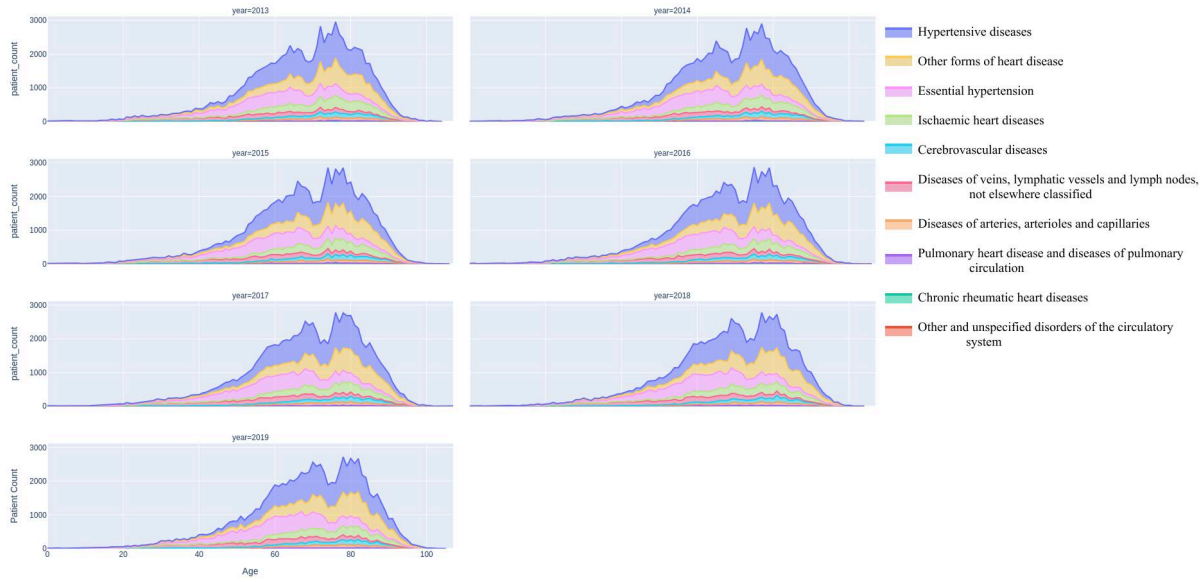
Kasutajale võib huvi pakkuda, kas vaatluse alla võetud kategooriate puhul esineb erinevusi meeste ja naiste vahel. Võtame võrdlusesse ka naissoost patsiendid.



Joonis 10. Vereringeelundite haiguste esinemise arv naistel (2013–2019). Kuvatõmmis rakendusest.

Joonisel 8 on kujutatud, et meestel kasvab vereringeelundite haiguste diagnoosimine kuni 60. eluaastateni ja seejärel püsib edasi võrdlemisi stabiilsena. Joonisel 10 aga näeme, et naiste puhul on selles vanuses terves kategoorias diagnoosimiste arv jätkuvas kasvutrendis. Mõlema soogrupi puhul on üksteisele sarnast diagnoosimiste langust täheldada 80. eluaastates. Jooniste omavahelises võrdluses tuleb esile, et meestel diagnoositakse kõrgvererõhktõbe rohkem varasemas eas kui naistel. Kui meestel on selle haiguse diagnoosimiste arv langustrendis juba vanuses 65, siis naistel tekib langustrend umbes 77 aasta vanuste juures.

Meil on võimalus ka võrrelda, et kas aastate lõikes on täheldada muutusi sellesse kategooriasse kuuluvate haiguste esinemistel. Selleks saame rakenduses aktiveerida aastate omavahelise võrdluse vaate ja valida kujutamiseks mõlemad soorühmad kokku. Joonisel 11 on näha, et vereringeelundite haiguste puhul on haiguste sagedus üle aastate ühtlane.



Joonis 11. Vereringeelundite haiguste esinemise arv aastate omavahelise võrdlusena. Kuvatõmmis rakendusest.

Rakenduse võimalike kasutusjuhte on veel mitmeid. Kogu eelnevalt kirjeldatud protsessi saab näiteks korrata ka visiitidega kaasnenud kulude visualiseerimiseks. Selle tarbeks tuleks lisaks vaid rakendada visiidikulude valik filtrites. Rakenduse interaktiivsus võimaldab uurida üksikasjalikult iga aasta ja vanuse puhul esinevaid väärtusi. Tänu sellele on võimalik vaadelda näiteks kõige kulukamaks osutunud diagnoose ja ka avastada andmestikus andmete sisestamisel tekkinud vigasid.

4.4 Edasiarendusvõimalused

Interaktiivse andmete töölaua põhifunktsionaalsus sai käesoleva töö puhul implementeeritud, kuid loodud lahenduse kasutusväärtuse tõstmiseks ning programmi kasutusmugavust silmas pidades on arenduseks mitmeid võimalusi. Hetkel saab rakenduse töövoogu kokkuvõtvalt kirjeldada nii, et kasutaja veendub sobiva andmetabeli olemasolus, kohandab .env failis parameetrid ning seejärel käivitab rakenduse. Üks võimalikke täiendusi oleks rakenduse käivitamise eeltöö rakendusele delegeerimine ehk sobiva andmetabeli kokku arvutamine teha programmisiseselt. See hõlmaks endas rakenduse esmasel käivitamisel tabeli kokku agregeerimist või ajakohastamist, kui nõuetele vastav tabel on eelnevalt juba olemas. Selline lähenemine võimaldaks vähendada kasutajapoolset vajalikku eeltööd enne rakenduse töövoo alustamist.

Arendatud rakendus on hetkel kasulik eelkõige andmeanalüütikutele, kellel on juba eelnevalt ligipääs ka OMOP CDM kujul olevale terviseandmestikule. Tehisandmestik on sobilik

vahelüli, mille abil rakendust testida ning see representeerib küll tavakasutajale üldiseid trende, kuid seda vaid piiratud mahus. Selleks, et ka tavakasutaja saaks rakendust efektiivselt kasutada ja tulemusi analüüsida, tuleks reaalandmestik puhastada ning seejärel ka see avalikkusele kättesaadavaks teha. Andmestiku välja toomine turvakeskkonnast nõuab vastavaid lubasid ja kooskõlastusi ning kindlustatust, et ükski indiviid poleks loodud andmestiku põhjal tuvastatav. Andmete anonüümsel kujul agregeerimiseks on võimalik kasutada peatükis 3.2.2 loodud SQL skripti. Tutvustatud päringule lisatingimuste seadmisel saab eemaldada harvemini esinevad kohordid. Näiteks saame välja jätta diagnoosirühmad, kus patsientide arv on väiksem kui viis. Tulemusena saadud tabeli põhjal ei ole üksikjuhtumid tuvastatavad, kuid andmestik sisaldab siiski üldist representatiivset informatsiooni. Järgmiseks soovitatavaks täienduseks oleks lisada väline populatsioonipüramiid näiteks Statistikaameti andmestikust. Populatsioonipüramiidi lisamine võimaldaks edastada ja võrrelda informatsiooni visiitide arvust või keskmistest kuludest arvutatult ühe isiku kohta eri diagnoosigruppide lõikes.

Loogiliseks täienduseks oleks täiendavate graafikutüüpide lisamine, mis võimaldaks kasutajal endal visualiseerida olemasolev andmestik just endale sobival kujul. Lisaväärtust saab pakkuda ka võimaldades agregeeritud andmestiku eksportimist või näiteks kogu rakenduse lokaliseerimisega ehk tõlkimisega eesti keelde. Juhul kui tõlgitud andmed lisada juurde, asendamata olemasolevaid ingliskeelseid kirjeldusi, siis lokaliseerimise tulemusena suureneb kindlasti ka rakenduse käivitamise aeg, sest programmi poolt töödeldavaid andmeid on rohkem. Tõlkimise tarbeks oleks tarvis lisada diagnoosinimede eestikeelsed vasted JSON-failis ning lisada ka eesti keeles kirjeldatud töölaua elemendid.

Töö mahtu ning ajaraamistikku silmas pidades ei jõudnud töö autor eelnevalt nimetatud funktsionaalsusi käesoleva töö jooksul implementeerida, mistõttu jäävad need tuleviku edasiarendusteks.

5. Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks oli luua interaktiivne andmete töölaud kujutamaks arstiabi saanud isikute arvu ja visiitidega kaasnenud kulusid OMOP CDM andmebaasi põhjal. Peamisteks varasemalt loodud lahenduste puhul käsitlemata kitsaskohtadeks olid rakendatavus iga OMOP CDM kujul oleva andmebaasi peale ning patsientide andmekohortide ajalise mõõtme kujutamine. Siinse töö autorile teadaolevalt ei ole varem sarnast töölauarakendust OMOP CDM kujul standardiseeritud andmete kujutamiseks loodud.

Ühtse andmemudeli rakendamine on oluline jätkusuutliku tervisealase koostöö tarbeks nii riiklike kui ka eri tervishoiuteenuste osutajate infosüsteemide puhul. Inimeste tervise edendamiseks luuakse ühtsele andmemudelile rakendatavaid taaskasutatavaid tööriistu ja metoodikaid ning nende haldamisega tegeleb peamiselt rahvusvaheline OHDSI kogukond.

Bakalaureusetöö raames valmis programmeerimiskeeles Python kirjutatud töölauarakendus, mis võimaldab visualiseerida ja analüüsida terviseandmeid mistahes OMOP CDM kujul olevatest andmebaasidest. Kasutajaliideses on võimalik filtreerida andmeid nii diagnoosirühmade, isikute soo ja vanuse järgi. Kasutajal on võimalus andmeid kujutada virnastatud kihtdiagrammidena nii üle aastate kokku agregeeritult kui ka aastaid omavahel võrreldes. Lisaks kasutajaliidese arendamisele lõi töö autor ka taaskasutatava andmebaasi päringu, tänu millele saab hõlpsalt eri tabelitest kokku lugeda unikaalsete patsientide arvu koos neile osutatud tervishoiuteenuste kogukuludega nii kalendriaasta, vanuse- kui ka soorühmade lõikes. Käesolevas töös loodud rakenduse lähtekood ja andmepäring on avalikult kättesaadavad töö autori GitLab repositooriumis.

Töö edasiarendusteks pakkus autor välja nii rakenduse kasutamiseks tarviliku eeltöö suunamise rakendusesiseseks protsessiks kui ka reaalandmestiku puhastamise ja kättesaadavaks tegemise avalikuks kasutamiseks. Samuti tõi autor välja välise populatsioonipüramiidi lisamise, eri graafikutüüpide implementeerimise kui ka rakenduse lokaliseerimise.

Viidatud kirjandus

Bodenreider, O., Cornet, R. ja Vreeman, D. J. (2018). Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearbook of medical informatics*, *27*(1), 129–139. https://doi.org/10.1055/s-0038-1667077

EHDEN Foundation. (2022). Kasutatud 30.04.2025, https://www.ehden.eu/vision-and-mission/ehden-legal-entity

Gregorio, F. D., Varrazzo, D. ja The Psycopg Team. (2021). Psycopg – PostgreSQL database adapter for Python—Psycopg 2.9.10 documentation. Kasutatud 02.05.2025, https://www.psycopg.org/docs/

Hossain, S. (2019). Visualization of Bioinformatics Data with Dash Bio. 126–133. https://doi.org/10.25080/Majora-7ddc1dd1-012

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J. ja Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, *5*(1), 4022. https://doi.org/10.1038/ncomms5022

Muddana, A. L., Vinayakam, S. (2024). Data Manipulations with Pandas. Python for Data Science, 1, 171–200. https://doi.org/10.1007/978-3-031-52473-8

NumFOCUS. (2024). pandas documentation. Kasutatud 02.05.2025, https://pandas.pydata.org/pandas-docs/stable/

OHDSI. (2025). Standardized Data: The OMOP Common Data Model. Kasutatud 08.05.2025, https://www.ohdsi.org/data-standardization/

OHDSI Europe. (2017). Kasutatud 01.12.2024, https://www.ohdsi-europe.org/index.php/national-nodes/estonia

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. (2023). Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: Lessons learned. JAMIA Open, 6(4). https://doi.org/10.1093/jamiaopen/ooad100

Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G. ja Stang, P. E. (2011). Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association, 19(1), 54–60. https://doi.org/10.1136/amiajnl-2011-000376

Reisberg, S. (2016). Eesti eriarstiabi kulud on nüüd detailselt ja interaktiivselt vaadeldavad. Tarkvara Tehnoloogia Arenduskeskus, Eesti Haigekassa. Kasutatud 01.12.2024, https://stacc.ee/et/eesti-eriarstiabi-kulud-on-nuud-detailselt-ja-interaktiivselt-vaadeldavad/

Reisberg, S. (2024). Sulevi tehtud sünteetilised andmed ver 1.0 ja ver 1.1. Kasutatud 30.04.2025, https://docs.google.com/document/d/1_4PGfCO2BA2bD1EbLuQBwZpwxM32HXCAGi-vCRBAIiw

Reisberg, S., Mooses, K., Kolde, R., Kõrgvee, L.-T. ja Vilo, J. (2024). Uudne lähenemine – OMOP-andmemudelil põhinevad terviseuuringud. Eesti Arst, 103(9), 420–429. https://doi.org/10.15157/ea24470

Ryan, P. Hripcsak, G. (2021). The Book of OHDSI: Observational Health Data Sciences and Informatics. Kasutatud 01.12.2024, https://ohdsi.github.io/TheBookOfOhdsi/

Solvak, M., Vilo, J., Reisberg, S., Tamm, S., Oja, M., Ligi, K., Unt, T., Võrk, A., Leets, P., Kamm, L., Ostrak, A., Kaminaga, H., Siil, T., Tammet, T., Vaarandi, R., Nõmm, S., Lepik, T., Lember, V., Nõmmik, S., … Kerikmäe, T. (2022). Programmi RITA tegevuse 1 projekti „Masinõppe ja AI toega teenused“ lõpparuanne. Eesti Teadusagentuur. Kasutatud 30.04.2025, https://www.etag.ee/wp-content/uploads/2022/05/RITA_MAITT_LOPPARUANNE_FINAL.pdf

Sügis, E., Tampuu, A., Aljanaki, A., Fišel, M. ja Kull, M. (2024). Praktiline andmeteadus. Kõrgkooliõpik. Tartu Ülikooli arvutiteaduse instituut. https://hdl.handle.net/10062/106497

Talvik, H.-A. (2022). Töövoog tervisedokumentide teisendamiseks OMOP CDM kujule. Magistritöö. Tartu Ülikool, andmeteaduse õppekava. Kasutatud 30.04.2025, https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=75134

Tervise Arengu Instituut. (2022). Tervisestatistika ja terviseuuringute andmebaas. Kasutatud 18.03.2025, https://statistika.tai.ee/Resources/Info/juhend.pdf

Terviseinformaatika uurimisrühm. (2024). Diseases-Related Code Sets Translations to OMOP CDM Standard Concepts. Tartu Ülikool. Kasutatud 16.03.2025, https://github.com/HealthInformaticsUT/EstonianOMOPMappings/tree/main/diseases

Tartu Ülikool. (2024). HPC Public Documentation: Sensitive data analysis platform. Kasutatud 30.04.2025, https://docs.hpc.ut.ee/public/services/SAPU/

Lisad

Lisa I: Rakenduse lähtekood ja versiooni valik

Bakalaureusetöö raames loodud rakenduse lähtekood ja paigaldusjuhised on kõigile huvilistele avalikult kättesaadavaks tehtud järgnevas GitLab keskkonna repositooriumis: https://gitlab.cs.ut.ee/rasmusmi/visitcountsandcosts. Repositooriumis on kaks eraldiseisvat haru koos neile kohalduvate paigaldusjuhistega. Kasutajal tuleb valida, kas soovib kasutada rakendust koos reaalandmestiku või tehisandmestikuga.

Kui rakenduse kasutajal on ligipääs töö autori poolt SAPU keskkonnas loodud andmetabelile *user_rasmus_mirma.rw_count_and_total_cost*, siis tuleks SAPU keskkonda viia repositooriumi harus *main* olev lähtekood. Kui kasutajal puudub ligipääs nimetatud andmetabelile aga on olemas mõni teine OMOP CDM kujul olev reaalandmestik, siis tuleb rakendusele sisendiks sobiv andmetabel enne rakenduse töövoo alustamist luua. Sobiva struktuuriga andmetabeli loomise päring asub rakenduse *main* haru *init_db_query.txt* nimelises failis. Kui sobiva struktuuriga tabel on loodud, saab kasutaja jätkata vaikimisi valitud *main* harus *README.md* failis kirjeldatud seadistamise protseduuriga.

Reaalandmestikule juurdepääsu puudumisel saab kasutada programmi *demo_dataset* harus asuvat koodi koos nimetatud haru *data* kaustas asuva tehisandmestikuga. Selle versiooni kasutamise puhul ei ole kasutajal tarvis täpsustada andmebaasi kasutajatunnust ja parooli. Sarnaselt *main* harule, on ka *demo_dataset* harus *README.md* failis kirjeldatud programmi käivitamise juhend.

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Rasmus Mirma ,
(autori nimi)

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

   Haiguste sageduste visualiseerimine OMOP CDM andmebaasi põhjal ,
   (lõputöö pealkiri)

   mille juhendaja on Jaak Vilo ,
   (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.

Rasmus Mirma
15.05.2025