TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Patrick Lomp

Diagnoosi trajektooride visualiseerimine

Bakalaureusetöö (9 EAP)

Juhendajad: Raivo Kolde
Maarja Pajusalu

Tartu 2021

Diagnoosi trajektooride visualiseerimine

Lühikokkuvõte:

Lõputöö eesmärk on luua interaktiivne tööriist, mis aitab visualiseerida graafide abil meditsiiniliste andmete analüüsi tulemusi. Töös loodud rakendus kuulub R tarkvarapaketi Trajectories alla. Trajectories paketis asetseb algoritm, mis tuvastab tihedalt koos esinevad sündmused terviseandmetes. Selliste diagnooside, ravimite väljakirjutamise ja protseduuride trajektoore on toorandmetes keeruline mõista. Sellepärast on vaja hästi visualiseerivaid vahendeid, et leida andmestikust relevantseid mustreid. Antud töös luuakse Shiny rakendus, mille abil saab sirvida elektroonilistes terviseandmetes leitud mustreid ja teha edasisi vaatlusi soovitud trajektooridel.

Võtmesõnad:

R, Shiny, andmete visualiseerimine, graafid

CERCS: P170 Arvutiteadus, süsteemid

Visualising health trajectories

Abstract:

The goal of this thesis is to produce an interactive tool to visualize medical data analysis results using graphs. The interactive tool is a part of a larger R package called Trajectories. Trajectories package has an algorithm that identifies common sequences of events in medical data. Such trajectories of diagnoses, drug prescriptions and procedures can get rather complicated and good visualisation methods are needed for finding relevant patterns in the data. In this thesis, a Shiny application is made, which can be used to browse the results of patterns found in medical data and do further characterization of trajectories of interest.

Keywords:

R, Shiny, data vizualisation, graphs

CERCS: P170 Computer science, systems

Sisukord

1. Sissejuhatus

Elektroonilisi terviseandmeid (lüh. EMR, ingl. *electronic medical record*) kasutatakse teadustöödes sagedasti. Tänu terviseandmetele on võimalik läbi viia erinevaid uuringuid haigustest, mida poleks võimalik uurida klassikaliste kliiniliste uuringutega [1]. Erinevad riiklikud terviseorganisatsioonid koguvad oma patsientide kohta infot, nagu arstivisiidid, retseptide väljastamised ja protseduurid. Antud andmete õige kasutamine võib aidata ennustada või isegi ennetada raskemaid haigusi ja terviserikkeid. EMR'd kirjeldavad tegelikku ravipraktikat ning pakuvad võimaluse leida uusi seoseid, mis olid enne jäänud märkamatuks [2]. Antud juhul pakub huvi haiguste jadade ehk trajektooride leidmine elektrooniliste terviseandmete põhjal. Need võimaldavad kirjeldada trende populatsioonis või ennustada tulevikus tekkivaid haigusi seni põetud haiguste põhjal [3].

Kuigi on tehtud erinevaid uuringuid riiklikul tasandil erinevates riikides, näiteks Taanis ja Rootsis, mille abil on leitud huvitavaid seoseid, siis ei saa eeldada, et kõik nendes uuringutes leitud seosed kehtiksid ka teistes keskkondades [1]. Kuna eelnevad uuringud toetusid ühele andmestikule, siis uuringutele mõjuvad erinevad tegurid nagu tervishoiusüsteemi eripära. Tänu sellele ei pruugi saadud järeldused teistes keskkondades kehtida. Selleks, et leida relevantset informatsiooni haiguste trajektooride kohta, mis ei oleks kindla süsteemiga seotud, on vaja korraldada uuringuid suuremal või kombineeritud andmete võrgustikul.

Antud töö kuulub Trajectories paketi alla. Trajectories pakett loob standardiseeritud raamistiku statistiliselt tähtsate sündmuste trajektooride avastamiseks. Trajectories raamistik ei ole optimeeritud ühe andmeallika analüüsiks, vaid töötab erinevate etteantud andmetega.

Töös kasutavate terviseandmete töötlus tehakse algoritmiga, mille lõpptulemus on andmestik, mis koosneb erinevate sündmuste paaridest. Üks sündmuste paar sisaldab muu hulgas sündmuse domeeni, sündmuste vahel olevate päevade arvu ja sündmuste paari väärtust. Kätte saadud sündmuste paare saab kujutada graafina ning paare kombineerides saab kätte pikemaid seoseid ehk trajektoore.

Avastatud paare ja trajektoore on tabelina keeruline uurida. Kuigi paare on võimalik kirjalikul kujul näha, siis paaride vahelisi seoseid mitte. Sellepärast on vaja kasutada visualiseerimist, et mitmesammulised seosed välja tuua. Graafide abil visualiseerides on võimalik näha haiguste arengute mustreid ning tänu filtreerimisvõimalustele on võimalik relevantne info kiirelt kätte saada.

Lõputöö eesmärk on luua R programmeerimiskeeles tarkvarapakett, mis teeb Trajectories tarkvarapaketi alla kuuluva algoritmi poolt kogutud andmepaarid visuaalselt loetavaks ja arusaadavaks. Andmepaare visualiseeritakse graafidena, kus igale sündmusele vastab graafi üks tipp. Graafe on võimalik salvestada sobivasse formaati koos graafi alla kuuluvate andmetega. Tarkvarapaketti saab vaadata lingitud koodirepositooriumis Githubi keskkonnas.

Käesolev bakalaureusetöö koosneb kokku 6 osast. Teises peatükis kirjeldatakse kasutatavaid mõisteid ja termineid. Kolmandas peatükis antakse ülevaade lõputöö taustast ja võrreldakse tööd varem valminud projektidega. Neljas peatükk keskendub rakenduse implementatsioonile. Viies peatükk sisaldab tulemuste analüüsi ja viimane peatükk kokkuvõtet.

2. Mõisted ja terminid

OMOP (Observational Medical Outcomes Partnership) Common Data Model (CDM) on mudel, mis lubab teha süstemaatilist analüüsi erinevate jälgitavate andmebaaside andmete kohta. Andmebaaside andmed viiakse ühisele formaadile ning samuti antakse andmetele sama kirjeldus.[1]

RHK-10 ehk ICD-10 on rahvusvaheline haiguste ja nendega seotud terviseprobleemide, epidemioloogia, sümptomite ja ebanormaalsete leidude ning kaebuste statistilise klassifikatsiooni kümnes versioon. [2]

Suhteline risk (ingl. *Relative Risk*) – oletatava teguriga kokkupuutunud indiviidide haigestumuse ja teguriga mitte kokkupuutunud isendite haigestumuse suhe ; kohortuuringutes tihti kasutatava seose mõõt. [3]

Graaf G = (V,E) koosneb tippude hulgast V ja servade hulgast E, kus nii V kui ka E on lõplik. Graafi struktuur ehk seos V ja E vahel määratakse mingi intsidentsusfunktsiooni E etteandmisega, mis igale servale e ∈ E seab vastavusse mingi tippude paari {u, v}.[4]

Suunatud graaf on graaf, mille tippude vahelised servad on suunatud. [5]

Komorbiidsus (ingl. *comorbidity*) on hulgihäirelisus, kahe või mitme (iseseisva) haiguse samaaegne esinemine [6].

[1] https://www.ohdsi.org/data-standardization/the-common-data-model/

[2] https://www.who.int/standards/classifications/classification-of-diseases

[3] http://www.eau.ee/~viltrop/EpiS_nastik.pdf

[4] https://research.cyber.ee/~peeter/teaching/graafid03s/graafid.pdf

[5] https://research.cyber.ee/~peeter/teaching/graafid03s/graafid.pdf

[6] Meditsiinisõnastik, 2004

3. Taustainfo

3.1 OMOP formaat

Meditsiini andmeid kogutakse erinevatel eesmärkidel ja erinevates formaatides, mis varieeruvad riigiti, haiglati ja isegi haigla sees erinevate osakondade kaupa. See on ajalooliselt teinud terviseandmete suuremahulise analüüsi üle paljude andmeallikate keerukaks. Standardite puudus on takistanud ka standardiseeritud uurimismetoodika ja koodi arendamist.

Üks võimalik lahendus on andmete viimine ühtsele andmemudelile. Selleks on võimalik kasutada andmemudelit OMOP CDM, mis defineerib ära ühtse andmestruktuuri ning standardiseerib ka kõigi andmeväljade esituse.

Trajectories pakett valmib EHDEN projekti raames ning kasutab OHDSI loodud OMOP andmeformaati. OHDSI ja EHDEN proovivad saavutada kirjeldatud eesmärke läbi ühtsete andmestandardite kasutuselevõtu edendamise. OMOP CDM formaat aitab erinevate andmebaaside andmed viia ühisele formaadile ning samuti antakse andmetele sama kirjeldus. Nii on võimalik standardiseerida meditsiini andmed erinevatest riikidest ja organisatsioonidest.

Tänu sellele on omakorda võimalik luua standardset tarkvara taoliste andmete analüüsiks. Tänaseks päevaks on palju erinevaid institutsioone erinevatest riikidest viinud oma andmed OMOP CDM kujule üle. OHDSI andmete järgi on formaat seotud üle 200 erineva andmebaasiga [4]. Nii on võimalik projekti katsetada erinevate andmestikkudega. See teeb erinevates keskkondades saadud andmete tulemused võrreldavaks.

Nii EHDEN kui ka OHDSI, kes formaati kasutavad, on loodava rakenduse kasutajasihtgrupid. Organisatsioonidel on võimalik erinevaid OMOP formaadis andmeid rakendusega analüüsida. See tähendab, et pole vaja iga projekti jaoks eraldi analüüsivat tarkvara luua.

OHDSI (Observational Health Data Sciences and Informatics) on mitut sidusrühma hõlmav interdistsiplinaarne koostöö programm, mille eesmärk on terviseandmete väärtuse väljatoomine suuremahulise analüüsi abil. Kõik OHDSI lahendused on avatud lähtekoodiga [5]. OHDSI on antud töös kasutatud OMOP CDM formaadi looja.

EHDEN (European Health Data & Evidence Network) loodi tegelemaks modernsete meditsiiniliste väljakutsetega luues ülevaadet ja pakkudes tõendeid päris maailma suure mahuga kliiniliste andmete järgi [6]. EHDEN teeb tihedat koostööd OHDSI organisatsiooniga.

3.2 Trajectories

Antud ülevaade on koostatud Künnapuu, Ioannou, Ligi jt loodud Trajectories paketti tutvustava artikli põhjal [1].

Trajectories R paketi eesmärk on olla raamistik statistiliselt tähtsate sündmuste trajektooride tuvastamiseks ja nende visualiseerimiseks OMOP formaadis andmestikust. Nii on võimalik leida sündmuste järjestusi, trajektoore, mis kirjeldavad tüüpilisi raviteekondi, haiguste progresseerumist ja võimaldavad avastada uusi haiguste riskifaktoreid. Rakendus töötab ette antud terviseandmete andmestiku peal, mille valib rakenduse kasutaja.

Trajektooride tuvastamise osa lahendab algoritm. Peale algoritmi analüüsi on võimalik analüüsitulemust visualiseerida, kasutades antut töö raames loodud rakendust.

Loodud raamistik on vabatarkvara koos avatud lähtekoodiga, mis kasutab OMOP CDM formaati ja kasutab standardiseeritud sõnastikku. Nagu eelmises alapeatükis selgitati, võimaldab formaat uuringuid teostada erinevatel andmestikel ning pärast tulemusi võrrelda. Seega on võimalik analüüsiks kasutada erinevate riikide või süsteemide andmeid.

OMOP formaadi ümber on tekkinud aktiivne vabavaralise tarkvara arenduse kogukond, mis on loonud tööriistu erinevat tüüpi uuringute sooritamiseks. Suurem osa sellest tarkvarast on loodud kombineerides keeli R ja SQL. Kuna antud töö kuulub EHDEN projektide alla, on samuti oluline, et töö oleks loodud kasutades samu keskkondi. Nii on võimalik tehtud tööd kasutada ka muudes projektides.

Algoritmi eesmärk on leida terviseandmetest statistiliselt oluliste sündmuste järjestused. Näiteks, et diagnoosile X järgneb ravimi Y väljakirjutamine või haiguse progressioon toimub läbi diagnoosikoodide X, Y ja Z. Et selliseid seoseid leida, tuleb leida sündmuste paare, kus esimese sündmuse E1 toimumine suurendab oluliselt sündmuse E2 toimumise tõenäosust. Selleks leitakse andmestikust kõik E1 juhud ning leitakse vanuse ja soo järgi sobitatud kohort patsiente kellel E1 ei toimunud. Nüüd hinnatakse binoomjaotusega p-väärtus, et E2 esinemiste arv on peale diagnoosi E1 on oluliselt suurem kui võrdluskohordis. Riskide esinevust väljendatakse suhtelise riskina, mis on E2 juhtumise tõenäosuste jagatis kahes grupis.

$$sr = P(E1 -> E2)/P(!E1 -> E2)$$

Kui E1 ei mõjuta sündmuse E2 toimumist, siis peaks suhteline risk olema lähedal ühele. Kui E1 mõjutab sündmuse E2 toimumist positiivselt, siis on suhteline risk suurem ühest ja väiksem ühest kui mõju on negatiivne. Suurema suhtelise riskiga sündmuse järjestused on üldiselt ka statistiliselt olulisemad. Kuid seda seost mõjutab ka andmestikus olevate sündmuste arv. Seega haruldaste sündmuste puhul leitakse ainult väga suure suhtelise riskiga olulisi sündmuste paare, kuid levinumate sündmuste puhul osutuvad statistiliselt oluliseks ka väiksema efektiga paarid. Seetõttu on oluline jälgida andmete interpreteerimisel nii suhtelist riski, kui ka toimunud sündmuste arvu.

Enne analüüsi tegemist saab määrata vaadeldava populatsiooni grupi tunnusomadusi. On võimalik analüüsida kogu andmehulka, kuid siiski võib osutuda kasulikuks määrata kriteeriumeid, mida vaadeldav patsientide grupp täitma peab. Antud raamistikus kasutatakse paindlikke definitsiooni printsiipe OHDSI/OMOP võrgust, kus vaadeldav grupp on hulk inimesi, kes rahuldavad vähemalt ühte kriteeriumi vaadeldava aja vältel.

Samuti on võimalik täpsustada uuringu parameetreid. Saab täpsustada, milliseid sündmusi analüüsis kasutatakse. Eelnevad uuringud on analüüsiks kasutanud peamiselt haiguste diagnoose, siis antud tööriistaga on võimalik kasutada sündmuseks ka muid terviseandmeid nagu ravimite määramised või muud protseduurid. Kuna kasutatakse OMOP CDM formaati, siis kõik sündmused on kodeeritud standardiseeritud OMOP sõnastikus. Paljude sündmuste paaride mõju võib olla väga väike, kus RR on ühe lähedal ning sel juhul need paarid pakuvad vähe praktilist väärtust. Seega olenevalt analüüsi eesmärgist on võimalik uurijal RR vahemik ise määrata, mis jätab paarid, mis vahemikku ei kuulu, analüüsist välja.

3.3 Rakenduse sisendandmed

Töös kasutatakse sisendfailiks Trajectories analüüsi poolt loodud faili. Fail on tabeli kujul .xlsx formaadis (Joonis 1).

E1_CONCEPT_ID	E2_CONCEPT_ID	E1_NAME	E2_NAME	E1_DOMAIN	E2_DOMAIN	RR	E1_AND_E2_TOGETHER_COUNT_IN_EVENTS
140673	1501700	Hypothyroidism	levothyroxine	Condition	Drug	3.231695	54
4079750	4114585	Osteoarthritis of knee	Primary gonarthrosis, bilateral	Condition	Condition	2.084352	85
77074	1506270	Joint pain	methylprednisolone	Condition	Drug	1.608131	69
194133	4196341	Low back pain	Radiography of spine	Condition	Procedure	1.432741	71
4216397	1201620	Nerve root disorder	codeine	Condition	Drug	1.385978	61
4216397	1125315	Nerve root disorder	acetaminophen	Condition	Drug	1.385978	61
80809	1307046	Rheumatoid arthritis	metoprolol	Condition	Drug	1.361409	106
80809	4324693	Rheumatoid arthritis	Mammography	Condition	Procedure	1.360532	108
1150345	4324693	meloxicam	Mammography	Drug	Procedure	1.335875	70
4083556	4163872	Seronegative rheumat	Plain chest X-ray	Condition	Procedure	1.326914	60
4035611	904453	Seropositive rheumat	esomeprazole	Condition	Drug	1.314105	65
46272790	4013636	X-ray of limb	Magnetic resonance imaging	Procedure	Procedure	1.312315	79
37016877	4004517	Plain x-ray of upper lir	Ambulatory surgery	Procedure	Procedure	1.30839	58
77074	46272790	Joint pain	X-ray of limb	Condition	Procedure	1.308099	107
4114585	4305221	Primary gonarthrosis, l	US scan of abdomen and pelvis	Condition	Procedure	1.301792	60
4163872	4158569	Plain chest X-ray	Emergency procedure	Procedure	Procedure	1.30172	93
80809	781039	Rheumatoid arthritis	alprazolam	Condition	Drug	1.297611	55
37016877	4324693	Plain x-ray of upper lir	Mammography	Procedure	Procedure	1.296859	88

Joonis 1. Näide sisendandmetest.

Etteantud faili iga rida koosneb kahest sündmusest, mis on tähistatud kui E1 ja E2. Andmestikus esineb sündmus E1 enne sündmust E2. Veergude kirjeldused on välja toodud tabelis 1.

Tabel 1. Valitud sisendandmete veerud koos kirjeldustega.

Veerg	Kirjeldus
E1_CONCEPT_ID	E1 sündmuse id kood
E2_CONCEPT_ID	E2 sündmuse id kood
E1_NAME	E1 sündmuse nimi
E2_NAME	E2 sündmuse nimi
E1_DOMAIN	E1 sündmuse domeen
E2_DOMAIN	E2 sündmuse domeen
RR	Paari suhtelise riski väärtus
E1_AND_E2_TOGETHER_COUNT_IN_EVENTS	E1 ja E2 koosesinemiste arv andmestikus

Mõlema sündmuse kohta on andmereas neid iseloomustavad andmed nagu id, nimi ja domeen. Sisendandmetes on erinevat tüüpi sündmused grupeeritud domeenidesse. Sündmuse domeeniks võib olla diagnoos, ravimi määramine, protseduur või muu määratud tüüp. Peale neid on sündmusepaare hindavad seosed nagu suhteline risk ning E1 ja E2 koosesinemiste arv.

Rakenduse loomisel kasutatakse ainult seosed, mille suhtelise riski väärtus suurem kui üks. Nii eemaldatakse kõik tippudevahelised seosed, kus E1 esinemine mõjub E2 esinemisele negatiivselt. See tähendab, et kui on esinenud sündmus E1, siis E2 esinemise tõenäosus on väiksem kui juhul kui E1 poleks toimunud. Kui sellised seosed jätta graafi, siis võib graaf jääda ebaselge. Eeldatakse, et graafis on näha ainult need seosed, mis mõjutavad järgmise sündmuse asetleidmist positiivselt ehk järgmine sündmus toimub suurema tõenäosusega.

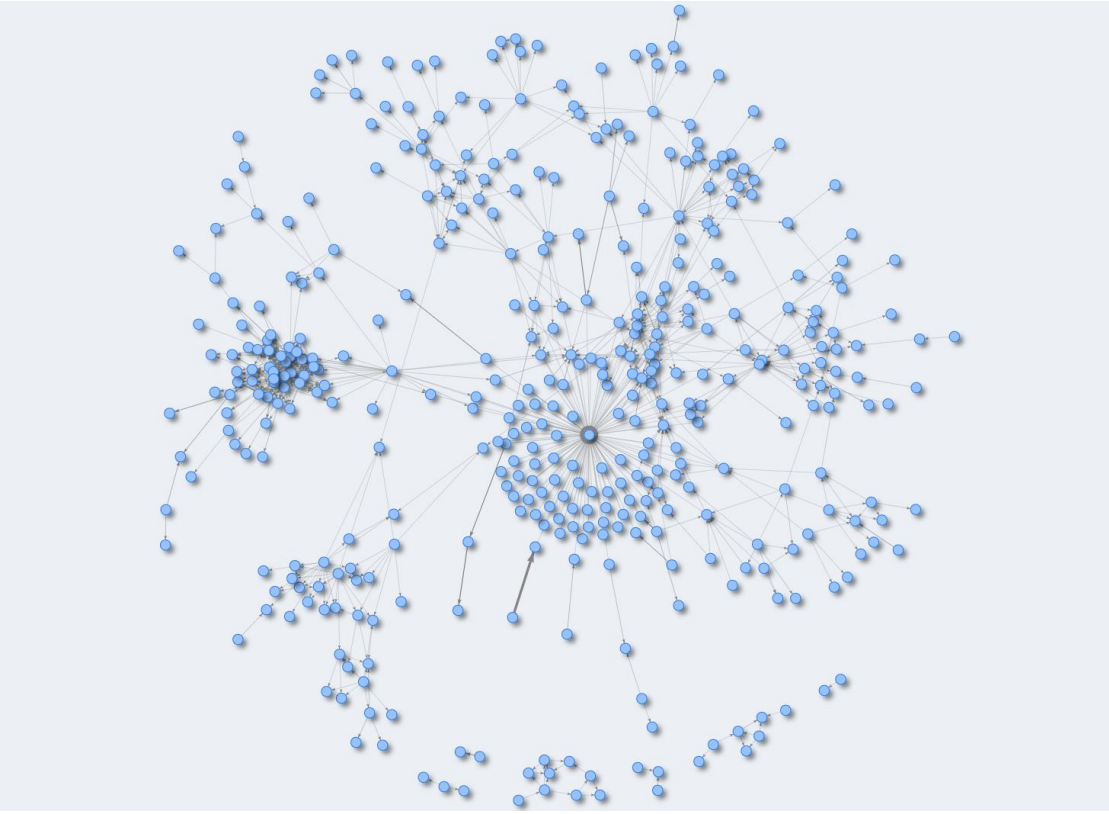
3.4 Visualiseerimise vajadus Trajectories paketis

Töö peamine eesmärk on leitud andmepaaride ja nende abil koostatud trajektooride visualiseerimine interaktiivsel kujul. Rakendus peab olema intuitiivselt arusaadav ning sisaldama vajalikke graafifiltreerimise omadusi.

Tabeli kujul näeb analüüsi tulemuses ainult andmepaare koos paari juurde käivate omadustega. Seega on tabel kasulik ühe andmepaari ülevaate saamiseks. See lubab näha antud paari väärtust ja võimaldab otsustada kas vaadeldav paar on statistiliselt huvitav või mitte, kuid pikemaid seoseid, kus trajektoor sisaldab üle kahe sündmuse sel viisil andmetest välja ei tule. Analüüsi tulemus võib sisaldada tuhandeid paare. Seega pole võimalik mitmesammuliste trajektooride kokkupanek käsitsi, kuid graafil paistab tekkinud trajektoor kohe välja. Kui on soov leida kõik seosed, mis puudutavad mingit diagnoosi, siis manuaalselt tabelist otsides võtaks see liialt aega, samas loodava kasutajaliidesega oleks see võimalik kiirelt leida.

Visualiseerimine on vajalik, et leida huvitavaid trajektoore kiirelt ja mõistetaval kujul. Andmepaare graafina visualiseerides tuleb välja seoseid, mida muidu tähele ei paneks. Visualiseerides on võimalik kiiremini vajalikku informatsiooni talletada. Visualiseerimise eesmärk on tähtsaid seoseid efektiivselt välja tuua. Visuaalne pilt, mis info kokku võtab, on parem viis tähtsate seoste ja mustrite märkamiseks. Isegi kui manuaalselt saaks seose kätte, siis on keerulisem seda seost näidata ilma visualiseerimata.

Ainult visualiseerimisest ei pruugi piisata, et efektiivselt infot edasi anda kuna graaf võib jääda raskesti mõistetav, sest sisaldab palju elemente (Joonis 2). Kasutajaliidese loomine lubab tekkinud graafi efektiivselt filtreerida lisatud kasutajakomponentide abil. Tänu sellele saab vähese vaevaga eemaldada graafilt vähetähtsad või ebahuvitavad seosed.



Joonis 2. Näide raskesti arusaadavast graafist.

Visualiseerimise tulemust on võimalik eksportida PDF formaadis, mis muudab selle kasulikuks tööriistaks uuringute näidete või põhjenduste väljatoomisel. Graafiline kasutajaliides võimaldab teha kiiremat koostööd, jagades leituid seoseid teiste osapooltega.

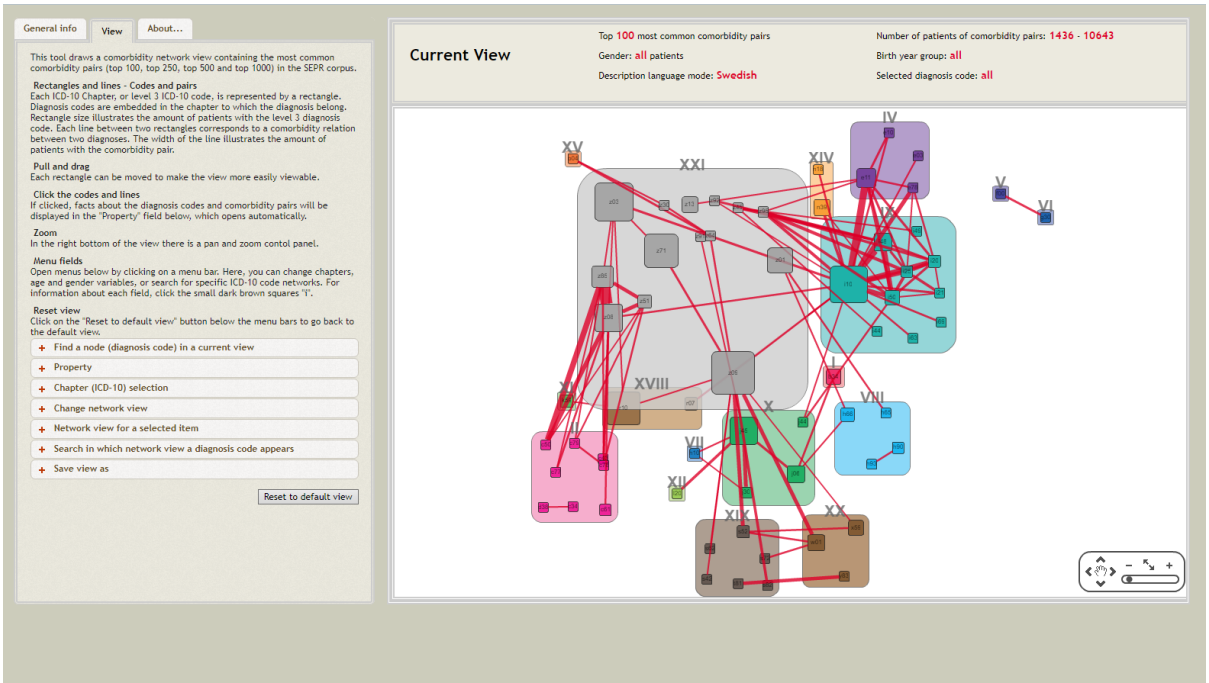
3.5 Sarnased tööriistad

Eelnevalt on loodud kaks sarnast rakendust, mis aitavad visualiseerida elektroonilisi meditsiini andmeid võrgustikena.

Comorbidity-view on esimesena loodud sarnane tööriist, mis omavahel seotud haiguste võrgustikke graafide abil visualiseerib (Joonis 3). Võrdlus on koostatud Hideyuki Tanushi, H. Dalianise ja G. Nilssoni kirjutatud artikli põhjal [7].

Comorbidity-View on visualiseerimise tööriist koosesinevate haiguste ehk komorbiidsete haiguste võrkudele. Rakendus näitab vaateid andmetest, mis on analüüsitud kasutades Stockholm Electronic Patient Records (SEPR) korpuse andmeid. SEPR korpus sisaldab endas pea 600 000 erineva patsiendi andmeid 900 erinevast kliinikumist. Haigused on põhiliselt kodeeritud ICD-10 abil.

Antud töö veebirakendus kasutab visualisatsiooni jaoks Flashi, mis on programmeerimise keel, mida tänaseks enam modernsed veebisirvijad ei toeta [8]. Tänu sellele pole võimalik enam rakendust veebisirvijas testida.



Joonis 3. Tööriist Comorbidity-View[7].

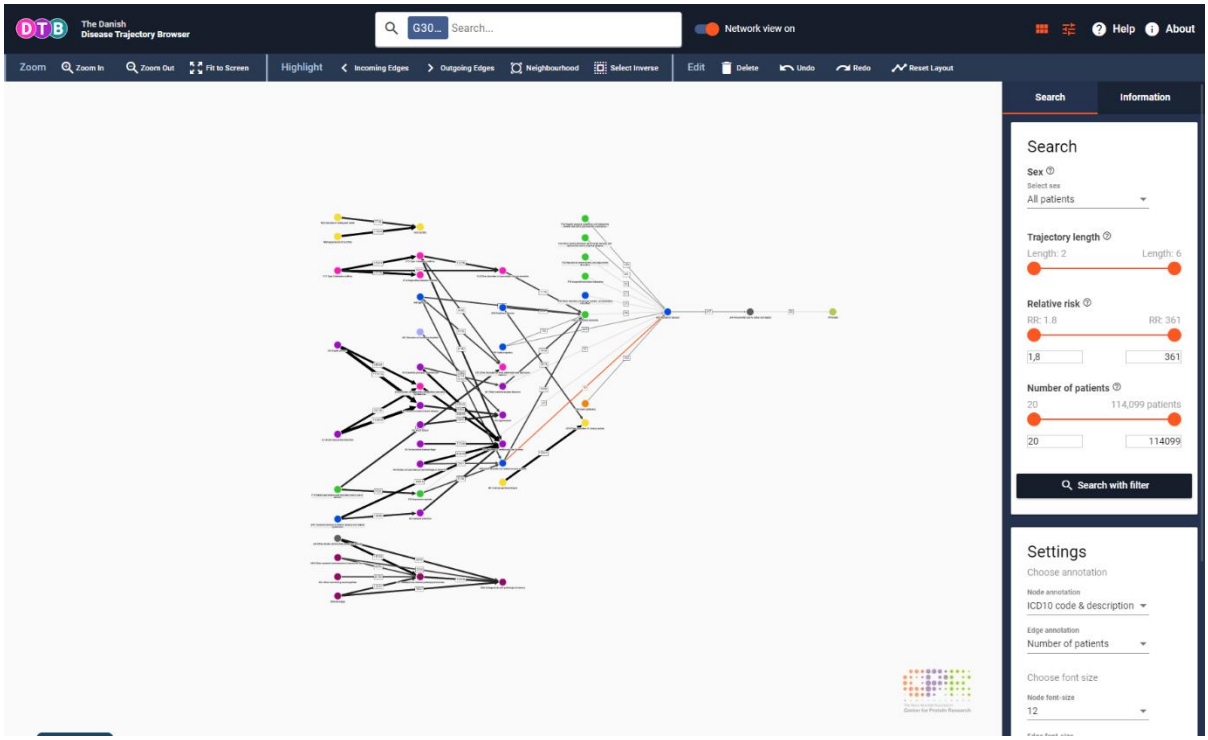
The Danish Disease Trajectory Browser on sarnane tööriist, mis lubab vaadelda haiguste trajektoore (Joonis 4). Järgnev võrdlus on koostatud Troels Siggaard, Roc Reguant,jt kirjutatud artikli põhjal [3]. DTB ehk Disease trajectory browser on tööriist, millega saab sirvida 25 aasta jooksul kogutud Danish National Patient Register'sse kogutud andmeid. Andmed sisaldavad 7,2 miljonit patsienti ja üle 122 miljoni haiguste diagnoosi kirje. Tööriista kasutajad

[7] https://www2.dsv.su.se/comorbidityview-demo/

saavad näha suunatud diagnoosipaare ja kombineerida neid lineaarseteks trajektoorideks. Saadud tulemused saab eksportida erinevatesse toetatud formaatidesse.

Tööriistas saab otsida haigusi ICD10 koodi järgi kas ühe haiguse koodi või ka mitme kaupa. Otsingutulemust saab filtreerida erinevate tunnusomaduste järgi nagu suhteline risk või seotud patsientide arvu järgi. Peale otsingu tegemist saab tulemuse eksportida sobivasse formaati (toetatud on JSON, PNG, JPEG ja CSV formaadid).

Antud sirvija töötab veebirakendusena ning on kättesaadav DTB kodulehel[8].



Joonis 4. Tööriist The Danish Disease Trajectory Browser.

Trajectories paketi eesmärk on samuti elektrooniliste meditsiini andmete abil leida huvi pakkuvaid haiguste seoseid. Nii nagu kahes välja toodud töös kasutatakse seoste visualiseerimiseks graafe. Välja toodud töödest leiti kasulikke aspekte visualiseeriva rakenduse koostamisel. DTB eeskujul luuakse rakendusse filtreerimiskomponent, kus saab graafi filtreerida RR ja patsientide arvu järgi. Suunatud graafi loomisel arvestatakse serva omadusi nagu paksus ja värv, et näidata ühendatud tippude seose tähtsust. Lisaks luuakse võimalus graafide salvestamiseks.

Kahel vaadeldaval rakendusel on andmestik ette määratud, st et kasutajal pole ise võimalik andmestikku valida. Seega, saadud tulemused on seotud etteantud andmestikuga, mida mõjutavad keskkond jm tegurid. Kuna etteantud andmestik on seotud kindla keskkonnaga, ei pruugi tulemused kehtida muudes keskkondades. Trajectories raamistikus saab analüüsiks kasutatavat andmestikku valida.

Vaadeldavad tööd kasutavad meditsiiniandmetest ainult haiguste diagnoose. Seega muude domeenide nagu ravimite, protseduuride jm seotuid sündmuseid ei ole võimalik näha.

[8] http://dtb.cpr.ku.dk/

Trajectories raamistik toetab ükskõik, mis sündmust senikaua kuni andmestik on OMOP CDM formaadis.

Mõlemad vaadeldud tööd on suletud lähtekoodiga, mis tähendab, et ei ole võimalik võtta antud töid alustalaks. Pole võimalik teiste osapoolte poolt rakenduste edasiarendus või koodi muudatuste tegemine, et toetada teisi andmestikke. Seevastu Trajectories raamistik on avatud lähtekoodiga. Avatud lähtekood lubab kasutada projekti eeskujuks või näha, kuidas keerukamad osad on lahendatud.

3.6 Graafide kasutamine visualiseerimiseks

Buldase, Laudi ja Villemsoni kirjutatud õpiku „Graafid" järgi on graaf G = (V, E) järjestatud paar mittetühjast hulgast V ja selle elementide paaride hulgast E [9, lk 5]. Hulga V elemente nimetatakse graafi tippudeks ja hulga E elemente graafi servadeks [9, lk 5]. Töös on graafi servad väärtustatud. Graafi servale antud väärtust nimetatakse kaaluks [9, lk 32].

Vaadeldavas andmestikus kirjeldab iga rida ühte sündmuste paari. Üks sündmuse paar koosneb kahest tipust, olgu need E1 ja E2. Iga sündmuste paar näitab seost kahe graafitipu vahel. Tippude seost kirjeldavad andmestikus näitajad nagu suhteline risk ja sündmuste koos esinemiste arv. Neid väärtusi kasutatakse serva kaaludeks. Mõlema kaalu abil on võimalik graafi filtreerida.

Andmete töötluse käigus eemaldatakse spetsiifiliselt välja paarid mille puhul mõlemat pidi on seosed olulised [1]. Seega algoritm otsib ajaliselt suunatud seoseid, kus E1 peab toimuma enne kui E2. Sellepärast on graaf suunatud. Sündmust E1 võib vaadata kui lähtetippu ja sündmust E2 kui lõpptippu. Kuna graafi servad on suunatud, siis on tegemist suunatud graafiga.

Graafi visualiseerimiseks on vaja graafi elemendid vaadeldavasse ruumi ära paigutada. Suure graafi elementide paigutus toob esile probleeme: kui tipud ja servad on halvasti paigutatud, pole võimalik elementidel enam vahet teha [10, lk 24]. Graafide paigutamiseks ei ole ühte optimaalset lahendust, vaid tuleb valida parim paigutusalgoritm, mis sobib vastava ülesande kontekstiga. Üldlevinud hea paigutusalgoritmi tunnused on graafi tippude ja servade ühtlane jaotus ruumis, sama pikkusega servad, servade minimiseeritud ristumine [10, lk 26].

Isegi kui on valitud sobiv paigutusalgoritm, siis ei pruugi sellest piisata. Suurte graafide visualiseerimine võib anda ülevaate struktuurist või sündmuse asukohast selle sees, kuid seda võib olla keeruline mõista [10, lk 24]. Sellepärast on vaja filtreerimisvõimalusi, mis tippude ja tekkinud servade arvu vähendaks.

Kuna tegemist on suunatud graafiga, siis on võimalik graafi genereerida hierarhilisel kujul. Loodud rakenduses on võimalik graafi genereerida hierarhiliselt vasakult paremale. Hierarhiline kuju lubab haiguse kulgu paremini jälgida.

Töös loodav võrgustik kasutab graafi elementide paigutamiseks jõududel põhinevaid algoritme. Igraphi poolt loodud paigutusfunktsioon valib graafi omaduste järgi automaatselt sobiva paigutusalgoritmi [11]. Implementatsiooni kohaselt kasutab funktsioon kuni 1000 tipuga graafi korral Fruchterman-Reingold paigutusalgoritmi. Fruchterman-Reingold algoritm kohtleb servi nagu vedrusid [12]. Rakendub kaks erinevat jõudu: üks jõud, mis tõmbab ühendatud tippe üksteise poole ja teine vastandikjõud, mis lükkab kõiki tippe (k.a omavahel ühendatud) üksteisest eemale [12]. Kui graafi tippe on üle 1000, siis funktsioon kasutab DrL paigutusalgoritmi. DrL on jõududel põhinev paigutusalgoritm, mis keskendub suurema skaalaga graafidele [13].

Teine rakenduses loodav võrgustik on Sankey võrgustik. Sankey diagramm aitab visualiseerida voo suurust tippude vahel [3]. Töös luuakse vood servade väärtuste järgi. Sankey vaade annab

lineaarsema ülevaate tippudevahelistest seostest. Vaade on kasulikum spetsiifilisemate otsingute puhul, sest liiga paljude elementide korral jääb graaf ebaselge. Võrgustikku on sobilik kasutada siis kui juba on rakendatud filtreerimise võimalusi, et leitud seoste hulk ei oleks liiga suur.

Kuna eesmärk on välja tuua trajektoore, siis graafilt eemaldatakse kõik isoleeritud tipud. Isoleeritud tipp on tipp, millel pole ühtegi serva ühegi teise tipuga. Kui graafilt eemaldada isoleeritud tipud, siis jäävad alles ainult tipud, millel on piisavalt tugev seos mingi teise tipuga ning alles jäävad ainult huvi pakkuvad tipud.

3.7 Shiny kasutamine R rakenduse loomisel

Loodava rakenduse jaoks on vaja graafilist kasutajaliidest. Üks levinumaid kasutajaliideseid R keelele on Shiny. Shiny on R studio poolt loodud pakett, mis lubab luua interaktiivseid veebirakendusi R keskkonnas [14]. Saab luua nii veebirakendusi kui ka ehitada koondpaneele. Shiny lubab tehnoloogiad nagu CSS, htmlwidgets ja javascript kasutada rakenduse osana. Shiny's toimub automaatne sisendite ja väljundite sidumine (ingl. *binding*) ning pakett sisaldab palju kasulikke kasutajaliidese elemente [15].

4. Rakenduse implementatsioon

4.1 Analüüs

Töö autorile esitati esialgne lähteülesanne koos ootustega loodavale rakendusele. Arenduse algfaasis kirjeldatud tingimustele lisandus nõudeid töö käigus tehtud konsultatsioonide käigus. Rakenduse täpsete nõuete analüüs ja kirjeldamine jäi bakalaureuse töö autori ülesandeks. Nõuete täpsustamiseks ja täiendamiseks uuriti sama valdkonna töid, lähtuti töö eesmärgist ja tähtsamatest tavadest. Koostöös Trajectories paketi loojatega korrigeeriti loodud nõudeid. Töö loomisel jälgiti agiilse tarkvaraarenduse põhimõtteid. Rakenduse arendus jaotati kahenädalasteks sprintideks. Sprint tähendab töö jaotamist väiksemateks iteratsioonideks [15, lk 37]. Iga sprindi lõpus vaadati üle töö hetkeseis, vajadusel lisati või muudeti nõudeid ning täpsustati järgmise sprindi eesmärk.

Algse tööeesmärgi kohaselt on vaja luua interaktiivne tööriist, mis lubaks sirvida Trajectoriese poolt loodud elektrooniliste meditsiiniandmete analüüsi tulemusi võrgustiku kujul. Rakendus valmib paketi osana.

4.1.1 Tähtsamad nõuded

Kokkulepitud nõuded on välja toodud alamteemade kaupa.

Keskkond

Rakendus peab töötama Trajectories paketi osana. Rakendus peab kasutama R ja Shiny raamistiku keskkonda.

Distributsioon

Rakendus on mõeldud jooksutamaks lokaalses keskkonnas kasutades R IDE.

Rakenduse kasutamine

Loodavat rakendust kasutatakse Trajectories paketiga andmeanalüüsi tegemise järel. Analüüsitulemused on sisendiks loodud visualiseerimispaketile.

Funktsionaalsed nõuded

Funktsionaalsete nõuete (ingl. *functional requirements*) abil on võimalik kirjeldada süsteemi oodatud käitumist [16]. Järgnevalt on välja toodud süsteemi tähtsamad funktsionaalsed omadused.

Graafi omadused:

1. Graafe koostatakse Trajectories analüüsi poolt loodud faili põhjal, mis on rakenduse sisendfailiks.
2. Graafi seosed vastavad sisendandmete seostele.
3. Rakendus sisaldab suunatud graafivaadet.
4. Graafide tippudeks on mingi sündmus.
5. Graafi sündmuste kirjeldused moodustatakse sisendandmete järgi.
6. Graafide servadeks on mingi kahe sündmuse mingi kaal.
7. Graafide servad toetavad kahte kaalu: RR ja sündmuste koos esinemiste arvu.
8. Rakendus sisaldab Sankey graafivaadet.
9. Suunatud graaf toetab hierarhilist vaadet.
10. Suunatud graafis on võimalik graafitippe esile tuua.
11. Suunatud graafi tipud kuuluvad gruppidesse.
12. Suunatud graafi tipud on värvitud vastavalt oma grupile.
13. Graafe on võimalik eksportida PDF formaadis.

14. Suunatud graafi tipu suurus peab olema seotud sündmuse esinemiste arvuga.
15. Suunatud graafi on võimalik sisse ja välja suumida (ingl. *zoom*).

Filtreerimine:

1. On võimalik valida spetsiifilisi sündmusi, mis jätab graafi alles ainult need sündmused, mis on valitud sündmustest kindla maksimaalse kaugusega.
2. Spetsiifiliste sündmuste filtreerimisel on kasutajal võimalik valida maksimaalset kaugust.
3. Graafi on võimalik filtreerida grupi kaupa. Grupid moodustatakse sisendandmete domeeni järgi
4. Graafi on võimalik filtreerida suhtelise riski väärtuse järgi.
5. Graafi on võimalik filtreerida koosesinemiste arvu järgi.
6. Graafi on võimalik filtreerida tipu tsentraalsuse väärtuse järgi.

Üldine:

1. Rakenduse sisendandmed on valideeritud.
2. Rakenduse väljakutsumine vale parameetriga annab veateate.
3. Rakenduse väljakutsumisel avaneb rakendus uues aknas ning jookseb lokaalsel pordil.
4. Rakenduse sisendandmeid on võimalik vaadata tabeli kujul.
5. Rakenduses kuvatud tabelis on võimalik kasutada otsingut.
6. Rakenduse põhilised tegevused on logitud.

Mittefunktsionaalsed nõuded

Mittefunktsionaalsed nõuded (ingl. *non-functional requirements*) kirjeldavad süsteemi atribuute nagu kasutuskõlblikkust, jõudlust ja töökindlust[17].

Graaf:

1. Visualiseeritava graafi tipu nime on selgelt näha.
2. Esile toodud graafitippe on võimalik teistest eristada.
3. Graafi serva laiusi on võimalik eristada.
4. Genereeritud graafi tippe ja servasid on võimalik omavahel eristada.

Filtreerimine:

1. Filtreerimine toimub piisavalt kiiresti: filtreeritud graafi genereerimise aeg on lühem kui 1 sekund.
2. Filtreerimise ajal ülejäänud kasutajaliides ei hangu. Rakendus on hangunud, kui ta ei reageeri sisendseadmete signaalidele[9].
3. Filtreerimiskomponentide kirjeldavad sildid (ingl. *labels*) on arusaadava tähendusega.

Kasutajaliides:

1. Kasutajaliides on üldiselt arusaadav.
2. Filtreerimiskomponentide mõju on mõistetav.
3. Kasutajaliides on minimalistlik.
4. Rakendus on koondpaneeli kujundusega.

[9] https://akit.cyber.ee/term/5739-tarduma-hanguma

4.2 Tippude ja servade andmestikkude loomine

Töös kasutatakse sisendfailiks tabelikujul faili, mis koosneb sündmuste paaridest. Etteantud faili iga rida koosneb kahest sündmusest, mis on tähistatud kui E1 ja E2, kus E1 esineb enne sündmust E2. Mõlema sündmuse kohta on andmereas kirjeldavad andmed nagu id, nimi ja domeen. Nagu alapeatükis 3.6 selgitati, defineerib graafi tema tippude ja servade hulk. Seega, et graafi koostada, on vaja sisendandmetest leida kõik tipud ja servad. Selleks moodustatakse saadud andmetest kaks eraldi andmestikku, kus esimene kirjeldab graafi tippe ja teine andmestik graafi servi.

Tippude andmestikus on iga tipu kohta lisatud tema iseloomustus. Praegusel juhul on iga tipu kohta vaja teada tema id koodi, nime ja domeeni. Peale tippude on vaja leida graafis olevad servad. Esiteks on iga serva kohta vaja teada lähte- ja lõpptippu selleks, et teada saada, kus serv asub. Teiseks on vaja teada serva kaale, mis iseloomustavad kui tugevat seost tippude vahel serv väljendab.

Sisendfailis on tipud kirjeldatud kahe sündmusena E1 ja E2. Selleks, et tippe kätte saada on vaja leida kõik unikaalsed sündmused, mis leidsid aset kas sündmusena E1 või E2. Lisatakse tipu kohta käivad veerud, nagu kirjeldus antud tipust ning grupp, mis saadakse tema domeeni järgi.

Selleks, et mõlemast hulgast väärtusi kätte saada, loodi kaks tabelit, kus esimene sisaldas kõiki E1 sündmuste andmeid ja teine tabel kõiki E2 sündmuste andmeid.

Peale tabelite loomist ühendatakse tabelid veergude järgi kokku üheks tabeliks. Lõpuks jäetakse ainult unikaalsed väärtused id järgi, mis eemaldab andmestikust kõik korduvad sündmused. Tippude lõplik andmestik sisaldab nelja veergu (Tabel 2).

Tabel 2. Tippude loomiseks tehtud andmestik

Veerunimi	**Veerunime kirjeldus**
Id	Tähistab tipu id koodi.
Title	Tähistab tipu nime
Group	Näitab, mis gruppi antud tipp kuulub
Label	Tähistab tipu nime. Kasutatud visNetworki poolt

Servade andmestiku loomiseks on vaja sisendfailist kätte saada kokku neli veergu:
1. Lähtetippu ja lõpptippu identifitseerivad veerud.
2. Serva kaale tähistavad veerud. On vaja leida nii suhtelise riski väärtus kui ka sündmuste koosesinemiste arv.

Tabel 3 kirjeldab seoseid servade andmestiku ja sisendandmete vahel.

Tabel 3. Servade loomiseks kasutatud veerud.

Sisendandmete veerunimi	Uus veerunimi	Veerunime kirjeldus
E1_CONCEPT_ID	FROM	Tähistab lähtetipu[10] id koodi.
E2_CONCEPT_ID	TO	Tähistab lõpptipu[12] id koodi.
E1_AND_E2_TOGETHER_COUNT_IN_EVENTS	E1_AND_E2_TOGETHER_COUNT_IN_EVENTS	E1 ja E2 koosesinemiste arv
RR	RR	Suhtelise riski väärtus

Nagu alapeatükis 3.3 mainiti, siis tuleb andmetest eemaldada kõik seosed, mille RR väärtus on väiksem kui 1. Selleks rakendatakse servadele esmane filtreerimine, kus eemaldatakse kõik servad, mille suhtelise riski väärtus jääb alla ühe. Kõik filtreerimisel tekkinud isoleeritud tipud eemaldatakse graafilt.

4.3 Andmete visualiseerimine

Trajektooride visualiseerimiseks kasutakse töös graafe ning graafide loomiseks R pakette. Loodud rakendus koosneb kahest graafi vaatest ja tabeli vaatest.

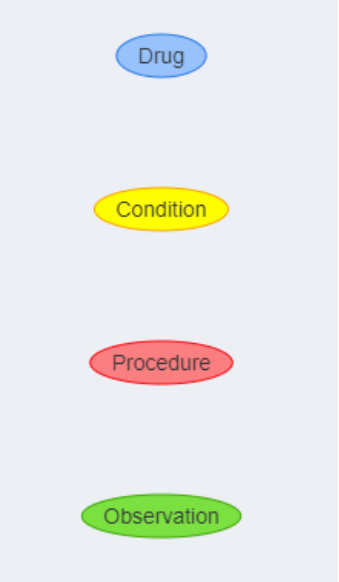
4.3.1 Suunatud graafivaade

Esimene graafivaade on suunatud graafivaade. Suunatud graafi puhul on iga serv suunaga, tänu millele on näha algsündmus ja sellele järgnev sündmus. Suunatud sündmuste jadadest on võimalik välja lugeda huvi pakkuvaid trajektoore. Iga tipu all on näha tipu nimi ning tipu peal hõljudes on näha tipu kirjeldust. Tipu värvi määrab grupp, mille alla antud tipp kuulub. Grupp on määratud tipu domeeni järgi.

Graafiservad on suunatud näidates sündmuste tekkimise järjekorda. Graafiserva paksus on seotud servadele määratud kaaluga. Servadele rakendatava kaalu määrab kasutaja: olgu selleks siis RR või E1 ja E2 koos esinemiste arv.

Graafist vasakul asub graafilegend, mis annab ülevaate kõikidest graafil esinevatest gruppidest (Joonis 4). Graafil asetsevad tipud on värvitud legendi järgi vastavalt seda värvi, kuhu gruppi tipp kuulub.

[10] https://research.cyber.ee/~peeter/teaching/graafid03s/graafid.pdf



Joonis 5. Suunatud graafi gruppe kirjeldav legend.

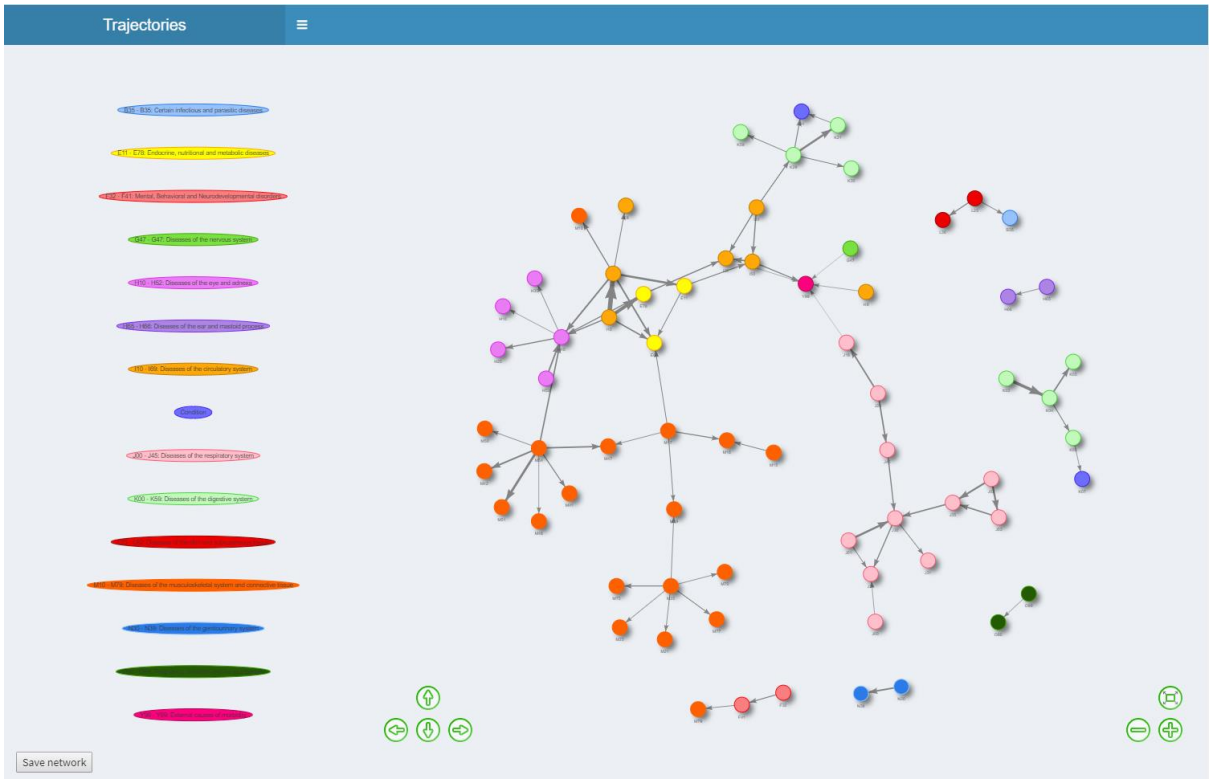
Graafi loomiseks on kasutatud raamistikku visNetwork. See on R pakett võrgustike visualiseerimiseks, kasutades javascripti teeki vis.js [18]. Järgnev info pärineb cran.r-project.org visNetworki tutvustusest [18]. Võrgustiku loomiseks on vaja kahte andmestikku: tippude ja äärte andmestikku. Tipu andmestik peab sisaldama vähemalt id veergu ja äärte andmestik from ja to veergu.

Praegusel juhul sisaldab tipuandmestik lisa veerge *label* ja *group*. Veerg *label* näitab graafil tipu all tipunime. Veerg *group* määrab tipu grupi ning selle järgi määratakse tipu värv.

Graafil olevaid tippe on võimalik valida. Valides mingi graafil oleva tipu, tuuakse see koos temaga lähedaste tippude ja servadega esile. Töös loetakse lähedaseks tipuks kõik tipud, mis on maksimaalselt kahe punkti kaugusel ja lähedasteks servadeks kõik servad, mis on lähedaste tippude vahel.

Graafi on võimalik vaadata kas hierarhiliselt või tavavaates. Suunatud graafi on võimalik genereerida hierarhilise paigutusega vasakult paremale. Hierarhiline kuju lubab haiguse kulgu paremini jälgida, kuid tavavaates toetab tõhusamalt suuremaid graafe. Mõlemas vaates graafi on võimalik sisse ja välja suumida (ingl. *zoom*).

Loodud graafi on võimalik alla laadida PNG-vormingus. Alljärgneval joonisel on näha suunatud graaf mittehierarhilisel kujul (Joonis 6).



Joonis 6. Suunatud graafivaade.

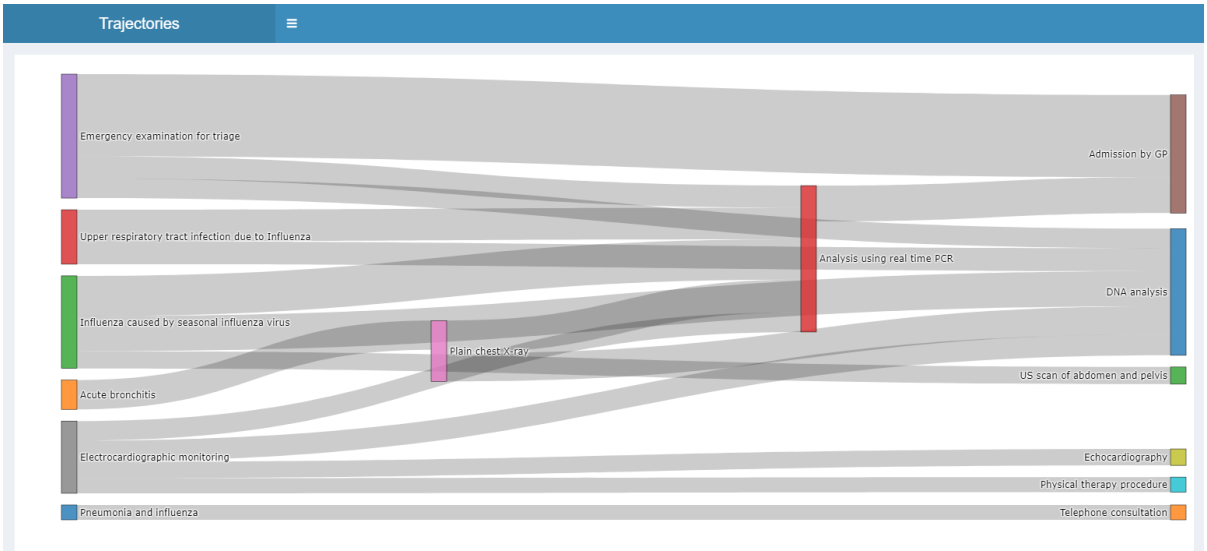
4.3.2 Sankey graafivaade

Teine rakenduses loodud vaade on Sankey vaade, mis aitab visualiseerida voo suurusi vastavalt servale rakenduvate kaalude järgi. Graafil on iga tipp tähistatud eraldi värviga. Tipu peal hõljudes on võimalik näha tipu kohta lisainformatsiooni. Iga tipu kohta näidatakse tipu nime, väljamineva voo suurust ja sissetuleva voo suurust. Voo suuruse määrab serva raskus. Mida suurem on voo suurus seda laiemalt on voog graafil esitatud.

Sankey võrgustiku paigutus on hierarhiline. Antud töös on implementeeritud hierarhilisus vasakult paremale.

Sankey võrgustik on loodud plotly raamistiku abil. Plotly raamistik lubab luua interaktiivseid graafe [19]. Võrgustiku loomiseks on vaja allikat, mis esindab lähtetippu, sihtmärki, mis tähistab lõpptippu, voo mahu määramiseks vajalikku väärtust ja silti, mis näitab tipu nime. Tekkinud graafi on võimalik alla laadida PNG kujul.

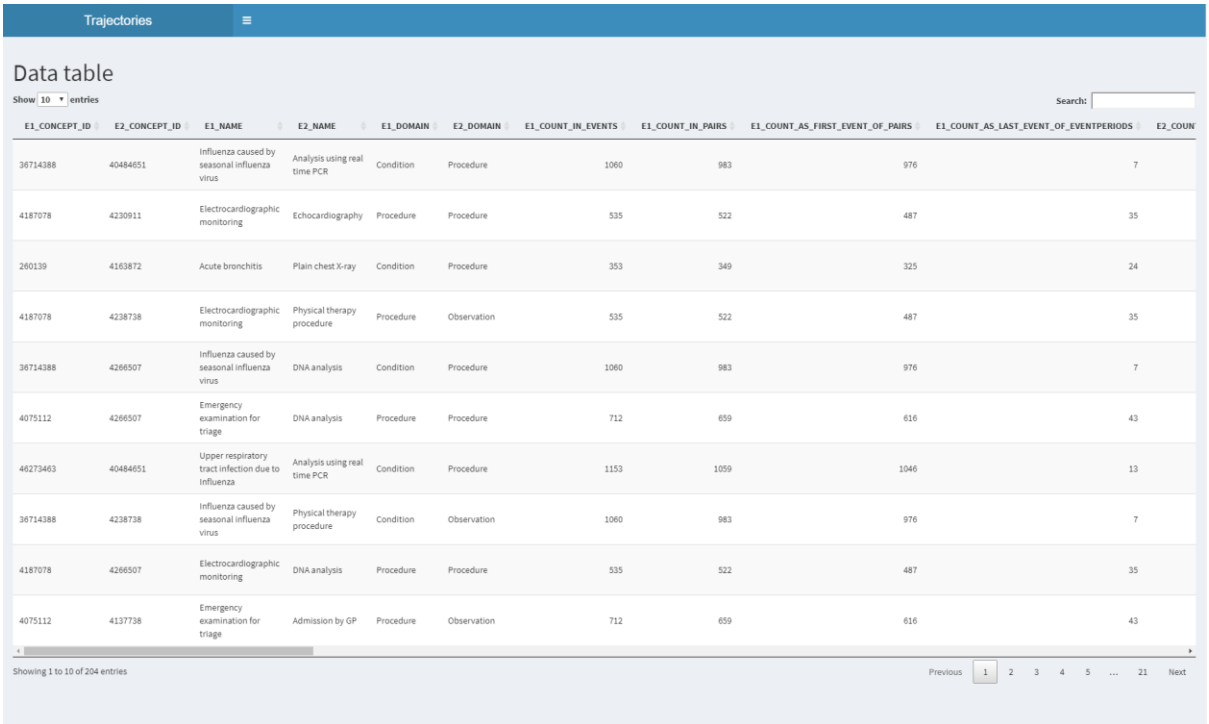
Alljärgneval joonisel on näide Sankey graafist (Joonis 7).



Joonis 7. Sankey graafivaade.

4.3.3 Andmete sirvimine tabeli kujul

Peale kahe loodud võrgustiku on võimalik andmeid vaadelda ka algsel tabeli kujul, mis lubab andmeid manuaalselt sirvida (Joonis 8). See on kasulik kindla paari täpse info kättesaamiseks või andmete kontrollimiseks. Lisaks saab kasutada otsingut, millega on võimalik soovitud andmerida üles leida sündmuse või muu omaduse järgi.



Joonis 8. Sisendandmed tabeli kujul.

4.4 Graafi filtreerimine

Kuigi on võimalik kasutada erinevaid paigutusalgoritme ja muid viise, et loodav graaf selgem välja näeks, siis mingist mahust alates pole see piisav ning graaf muutub raskesti mõistetavaks. Filtreerimise eesmärk on võrgustikust eemaldada võimalikult palju mitte huvipakkuvaid tippe ja servi. Nii muutub graaf uuesti arusaadavamaks ja tähtsad seosed paistavad rohkem välja. Ilma filtreerimata võivad tähtsad seosed jääda aga sootuks märkamata. Graafi filtreerimine on võimalik erinevate tippude või servade omaduste järgi. Alapeatükis kirjeldatakse filtreerimise loogikat ning kasutajaliidese elementide kasutamist filtreerimisel. Rakenduses on kokku neli peamist viisi kuidas võrgustikke filtreerida.

4.4.1 Filtreerimine sündmuse järgi

Enamasti ei paku huvi terve võrgustik korraga, vaid konkreetsed sündmused ning nendega seotud tipud. Et võimaldada sellist filtreerimist, saab kasutajaliideses valida üks või mitu tippu, mille tulemusel kuvatakse võrgustikus nende tippude lähimat naabrust.

Graaf filtreeritakse nii, et järgi jäävad ainult need tipud, mis asuvad valitud sündmuste tippudest kuni maksimaalse lubatud kauguseni. Kui on määratud mitu sündmust korraga, siis jäävad järgi valitud tipud ja tipud, mille kaugus vähemalt ühest valitud tipust on väiksem võrdne maksimaalse lubatud kaugusega. Maksimaalse lubatud kauguse saab määrata kasutaja ning selle algne väärtus on 3.

Kui määratud maksimaalseks kauguseks on 3, siis iga teine graafil esinev tipp on ühest vaadeldavast tipust maksimaalselt 3 sammu kaugusel. See tähendab, et ühest suvalisest tipust mingisse vaadeldavasse tippu peab jõudma tee, mis on kuni 3 sammu ehk serva pikk.

4.4.2 Filtreerimine sündmuste gruppide järgi

Sõltuvalt uurimisküsimusest ei pruugi rakenduse kasutaja olla huvitatud igat tüüpi sündmustest vaid ainult kindlatest klassidest. Näiteks võivad mingis uuringus huvi pakkuda ainult haiguste diagnoosid. Selle probleemi lahendamiseks on implementeeritud filtreerimine klassi järgi.

Iga tipp kuulub kindlasse gruppi. Tipp võib kuuluda ravimite, protseduuride, haiguste või muu sisendandmetes määratud grupi alla. Tipud on grupeeritud algandmete domeeni väärtuse järgi.

Kasutajal on võimalik graafi filtreerida grupi järgi. Nii on võimalik graafi alles jätta ainult need grupid, mis kasutajale huvi pakkuvad. Valides vaadeldavad grupid, eemaldatakse graafilt kõik tipud, mis vastavatesse gruppidesse ei kuulu.

4.4.3 Filtreerimine kaalu järgi

Tippude vahelised kaalud näitavad kui tugevalt on omavahel kaks tippu seotud. Kaalu järgi graafi filtreerides on võimalik sellelt eemaldada nõrgad seosed. Graafi on võimalik filtreerida kahe erineva kaalu järgi: suhteline risk (lüh. RR, ingl. *relative risk*) või sündmuste koos esinemiste arvu järgi.

Filtreerimisel kasutatakse liuguri (ingl. *slider*) komponenti, millel valitud väärtusega toimub filtreerimine. Kõik näidatust väiksema kaaluga servad eemaldatakse graafilt. Alles jäävad servad, mille mõlemad kaalud on suuremad kui kasutaja määratud minimaalne lubatud kaal. Kui kumbki kaaludest on väiksem kui lubatud, siis eemaldatakse vastav serv graafilt. Peale seda eemaldatakse graafilt tekkinud isoleeritud tipud.

4.4.4 Filtreerimine tsentraalsuse väärtuse järgi

Tipu tsentraalsus aitab mõõta kui tähtis mingi tipp võrgustikus on. Kuna tipu tähtsus võrgus pole täpselt defineeritud ja sõltub võrgustikust, siis esineb palju viise tsentraalsuse leidmiseks

[20]. Töös kasutatakse tsentraalsuse väärtuste leidmiseks *betweeness centrality* algoritmi, mis mõõdab tipu mõju võrgus. Mõõdiku järgi on tipp tsentraalne kui ta jääb kahe teise tipu paari lühima tee peale [21]. Töös kasutatud algoritm pärineb tidygraphi raamistikust.

Vahelmisuse väärtus leitakse iga tipu kohta. Kui tipu vahelmisuse väärtus on väiksem kui määratud minimaalne piirväärtus, siis tipp eemaldatakse graafilt. Kasutajal on võimalik piirväärtus määrata liuguri abil.

4.5 Shiny kasutajaliides

4.5.1 Shiny kasutus rakenduses

Antud töös on kasutatud Shiny raamistikku koos erinevate Shinyle mõeldud lisapakettidega. Shiny on kõige populaarsem visualiseerimise raamistik R keelele.

Shiny rakendus nõuab spetsiaalsete failide olemasolu, mida Shiny rakenduse käivitamisel otsib. Töö sisaldab 3 Shinyga integreeritud faili, mis on välja toodud tabelis 4.

Tabel 4. Shinyga seotud failid.

Fail	Kirjeldus
ui.R	Fail defineerib rakenduse kasutajaliidese.
server.R	Sisaldab vastava rakenduse loogikat. Fail omakorda toetub failidele utils.R ja filter_network.R. Fail utils.R kirjeldab endas erinevaid vajaminevaid funktsioone, mida kasutatakse server.R siseselt. Fail filter_network.R sisaldab andmete filtreemise jaoks loodud loogikat.
app.R	Faili abil käivitatakse Shiny rakendus

Shiny rakendus kasutab kahte skripti, mis omavahel suhtlevad ja vahetavad informatsiooni. Esiteks kasutajaliidese skript ui.R, mis sisaldab endas rakenduse küljendust (ingl. *layout*) ja määrab rakenduse väljanägemise. Teiseks serveriskript server.R, mida kasutatakse andmete protsessimiseks ja kasutaja sisendi muutustele reageerimiseks.

Lisaks Shinyle kasutab antud rakendus kasutajaliidese loomiseks ka teisi pakette, mis on välja toodud tabelis 5.

Tabel 5. Kasutajaliidese paketid.

Pakett	Kirjeldus
Shinydashboard	pakett on mõeldud Shiny lisapaketiks, mis lubab lihtsamini luua koondpaneele [22]
shinyWidgets	Shinywidgets pakett koondab endas hulga lisa kasutajasisendi komponente Shiny rakenduse loomise jaoks. [23]

Kuigi rakendus on mõeldud kasutamiseks eelkõige lokaalses keskkonnas, on seda võimalik vähese vaevaga teha kättesaadavaks veebikeskkonnas. Kuna tegemist on Shiny rakendusega, siis on sellel hea tugi serveril jooksutamiseks. Näiteks on võimalik rakendust juurutada (ingl. *deploy*) shinyapps.io[11] abil. Kui rakendus on veebis kättesaadav, siis on võimalik igal huvilisel seda kasutada. See muudaks analüüsi tulemuste jagamise kiireks ja kättesaadavamaks.

[11] https://www.shinyapps.io/

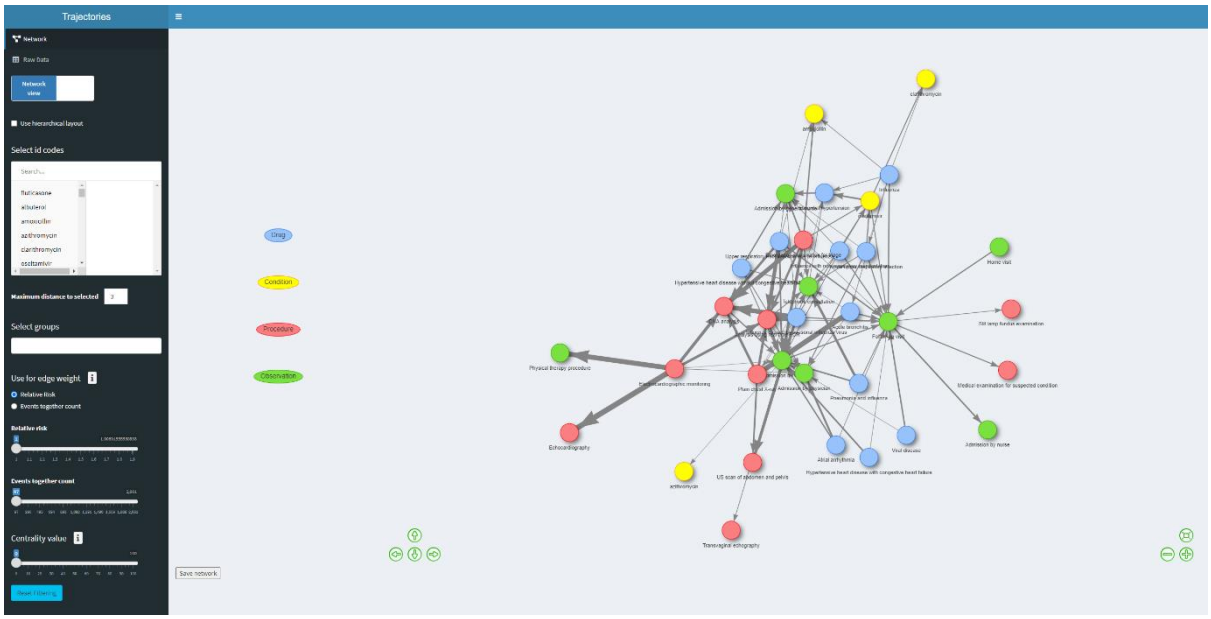
4.5.2 Kasutaja sisendi haldamine

Töös on kasutajaliidese sisendite haldamise jaoks kasutatud Shiny ja ShinyWidgetsi kasutajaliidese komponente.

Kasutaja saab valida kahe alamlehe vahel. Esimesel alamlehel asub andmete visualiseerimise keskkond ja teisel lehel on võimalik toorel kujul lugeda kaasaantud sisendandmeid.

Esimesel ehk võrgustike alamlehel on võimalik valida kahe erineva võrgustiku vahel. Valitud võrgustikku saab vahetada kasutades kaheolekulist lülitit (ingl. *toggle switch*). Vastavalt olekule genereeritakse suunatud graaf või Sankey võrgustik.

Iga graafi filtreerimise osa jaoks on kasutatud eraldi kasutajaliidese komponenti, mis on põimitud muutujaga, mis on väärtustatud kasutaja sisendi järgi. Põimitud muutujate väärtuste järgi toimib filtreerimine. Peale filtreerimiskomponentidele uute väärtuste andmist on kasutajal soovi korral võimalik graafi filtreerimisväärtusi algväärtustada. Loodud kasutajaliidest koos algväärtustega on võimalik näha alljärgneval joonisel (Joonis 9).



Joonis 9. Kasutajaliides.

4.6 Töö integreerimine Trajectories paketiga

Visualiseeriv rakendus on osa Trajectories raamistiku töö voost. Peale algoritmi pool loodud analüüsi toimub analüüsi uurimine. Selles osas kutsutakse rakendus automaatselt välja. Tänu sellele on kasutajal peale analüüsi võimalik tulemusi koheselt sirvida. Sel juhul baseerub visualiseerimine viimase analüüsi tulemuste failil, mis on loodud rakenduse sisendfailiks.

Rakendus kutsutakse välja joonisel 10 oleva käsuga. Käsk koosneb kolmest osast:

1. *Trajectories* on paketi nimi, kuhu rakendus kuulub,
2. *visualize_data_pairs* on funktsioon, mis käivitab rakenduse.
3. *data* on kaasa antud sisendfail. Tegemist on algoritmi analüüsi poolt loodud failiga.

```
Trajectories::visualize_data_pairs(data)
```

Joonis 10. Rakendust käivitav käsk.

Funktsioon *visualize_data_pairs* kontrollib, kas kasutajal on vajaminevad lisapaketid olemas ning vajadusel installib need. Seejärel kontrollitakse kas sisendfail on sobilikul kujul. Kui sisendandmed on valideeritud, siis käivitatakse Shiny rakendus.

Rakenduse lähtekood asub alamkaustas `./inst/shiny`. Rakendust käivitav käsk asub failis `./R/VisualisingTrajectories.R`.

5. Tulemuste analüüs

5.1 Lahenduse vastavus nõuetele

Käesoleva töö raames valminud lahendus täidab kokkulepituid nõudeid ja üldist eesmärki. Töö sisaldab kahte nõutud graafivaadet. Implementeeritud on suunatud graaf ja Sankey võrgustik. Suunatud graaf on visualiseeritud vastavalt nõuetele. Tippe on võimalik eristada gruppide järgi ning graafi servade paksus sõltub määratud kaalust. Samuti on nõuetele vastavalt implementeeritud Sankey vaade. Mõlemat loodud graafi on võimalik salvestada.

Rakendus toetab graafifiltreerimist erinevate omaduste järgi:

1. sündmus;
2. sündmuse grupp;
3. serva kaalude väärtus- on võimalik määrata minimaalne kaal suhtelise riski ja sündmuste koos esinemiste arvu järgi.
4. tsentraalsuse väärtus.

Töö sisaldab tabelivaadet, mis annab ülevaate sisendandmetest tabeli kujul. Tabelivaade sisaldab otsinguvõimalust.

Töö täidab mittefunktsionaalseid nõudeid. Koondpaneeli kujul rakenduse kasutajaliides on arusaadav ning samuti kasutatud paigutusalgoritmid töötavad ootuspärastelt.

5.2 Edasiarenduste võimalused

Rakenduse arendamisel peeti tähtsaks edasiarendamise võimalusi. Koodi kirjutamisel pandi rõhku headele tavadele ja koodi selgusele. Kuigi töö täidab oma eesmärki ja nõudeid, on erinevaid võimalusi loodud rakenduse edasiarenduseks.

Suunatud graafi abil on võimalik edasi anda rohkem informatsiooni kui praeguseks versiooniks on implementeeritud. Suunatud graafis on informatsiooni edastamiseks kasutatud serva paksust, kuid graafi tipud on kõik ühe suurusega. Tipu suurusega on võimalik samuti informatsiooni edasi anda. Suunatud graafis oleks kasulik kui tipu suurus oleks vastava tipu sündmuse esinemiste arvuga seotud.

Loodud rakendus on mõeldud jooksmaks lokaalselt. Soovi korral on võimalik paari muudatusega tööd jooksutada veebikeskkonnas. Töö on loodud Shiny raamistikuga, millel on palju erinevaid viise jooksutamaks rakendust serveris.

Antud rakendus on osa Trajektories paketist ning selle edasise lahendusega saab kursis olla jälgides Githubi repositooriumi https://github.com/EHDEN/Trajectories.

6. Kokkuvõte

Trajectories R paketi eesmärk on olla raamistik statistiliselt tähtsate sündmuste trajektooride tuvastamiseks ja nende visualiseerimiseks OMOP formaadis andmestikust. Raamistik valmib EHDEN projekti raames. Pakett loob analüüsi elektroonilistest meditsiiniandmetest ning tehtud analüüsi tulemusel toimub visualisatsioon. Analüüsi tulemuste visualiseerimine oli käesoleva töö eesmärgiks.

Töö raames valmis Shiny rakendus Trajectories paketile, mis aitab visualiseerida Trajectoriese analüüsi poolt loodud andmeid graafide abil. Analüüsi poolt loodud andmeid visualiseeritakse võrgustikena ning neid on võimalik hõlpsasti filtreerida.

Töös loodud rakendus sisaldab kahte erinevat graafivaadet: suunatud graafi ja Sankey võrgustikku. Loodud graafid aitavad andmestikust välja tuua huvipakkuvaid trajektoore ja sündmuste seoseid. Tekkinud graafe on võimalik salvestada PDF formaadis.

Suunatud graafi puhul on kõik servad suunatud. Tänu sellele tekivad suunatud sündmuste jadad, millest on võimalik välja lugeda huvi pakkuvaid trajektoore. Graafi on võimalik paigutada kas hierarhiliselt või tavavaates. Hierarhiline kuju lubab haiguse kulgu paremini jälgida, samas tavavaates jääb suuremate graafide paigutus selgem.

Sankey vaade aitab visualiseerida voo suurusi tippude vahel ning annab seostest lineaarsema ülevaate. See võimaldab näha seoseid, mis suunatud graafis võisid jääda märkamata. Graaf töötab hästi spetsiifilisemate otsingute puhul, sest suure elementide hulga korral jääb graaf ebaselge.

Peale graafide on võimalik sisendandmeid vaadata tabeli kujul. Selle abil on võimalik andmeid manuaalselt sirvida. See on kasulik kindla paari täpse info kättesaamiseks või andmete kontrollimiseks. Sirvimise hõlbustamiseks on võimalik kasutada otsingu võimalust.

Et leida graafist huvipakkuvaid trajektoore, on loodud erinevad filtreerimisvõimalused. Graafe on võimalik filtreerida sündmuse nime või grupi järgi. Sündmuse järgi filtreerides jäävad graafile ainult need sündmused, mis asuvad valitute naabruses. Grupi järgi filtreerimine lubab eemaldada graafilt mitte huvipakkuvad domeenid. Lisaks on võimalik filtreerida erinevate tipu ja servade omaduste järgi nagu kaalude suhteline risk ja koos esinemiste arv.

Rakenduse loomiseks loodi analüüs, mis toetus ülesande püstitusele, ning lepiti kokku tööle kehtivad nõuded. Loodud töö vastab nõuetele, kuid sellegipoolest omab erinevaid edasiarendus võimalusi. Trajectories projekt on lõputöö esitamise eesmärgiks kloonitud isiklikku repositooriumisse.

Kirjandus

[1] K. Künnapuu, K. Ligi, R. Kolde, S. Laur, J. Vilo, ja S. Reisberg, „Trajectories: an R package for visualizing directional event pairs from OMOP-formatted data“. 2021.

[2] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, ja D. Sontag, „Learning a Health Knowledge Graph from Electronic Medical Records“, *Scientific Reports*, kd 7, nr 1, Art. nr 1, juuli 2017, doi: 10.1038/s41598-017-05778-z.

[3] T. Siggaard *et al.*, „Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients“, *Nature Communications*, kd 11, nr 1, Art. nr 1, okt 2020, doi: 10.1038/s41467-020-18682-4.

[4] „resources:2020_data_network [Observational Health Data Sciences and Informatics]“. https://www.ohdsi.org/web/wiki/doku.php?id=resources:2020_data_network (vaadatud apr 16, 2021).

[5] „OHDSI – Observational Health Data Sciences and Informatics“. https://ohdsi.org/ (vaadatud apr 12, 2021).

[6] „European Health Data Evidence Network“, *ehden.eu*. https://www.ehden.eu/ (vaadatud apr 12, 2021).

[7] H. Tanushi, H. Dalianis, ja G. Nilsson, „Calculating Prevalence of Comorbidity and Comorbidity Combinations with Diabetes in Hospital Care in Sweden Using a Health Care Record Database“, *undefined*, 2011, Vaadatud: apr 14, 2021. [Online]. Available at: /paper/Calculating-Prevalence-of-Comorbidity-and-with-in-a-Tanushi-Dalianis/9c254edd38722dcedc04c1a7cdad15af25edf340.

[8] „Adobe Flash Player End of Life“. https://www.adobe.com/products/flashplayer/end-of-life.html (vaadatud apr 14, 2021).

[9] A. Buldas, P. Laud, ja J. Villemson, „GRAAFID“, 2003.

[10] I. Herman, G. Melancon, ja M. S. Marshall, „Graph visualization and navigation in information visualization: A survey“, *IEEE Transactions on Visualization and Computer Graphics*, kd 6, nr 1, lk 24–43, jaan 2000, doi: 10.1109/2945.841119.

[11] „layout_nicely: Choose an appropriate graph layout algorithm automatically in igraph: Network Analysis and Visualization“. https://rdrr.io/cran/igraph/man/layout_nicely.html (vaadatud apr 26, 2021).

[12] „Reingold Layout - an overview | ScienceDirect Topics“. https://www.sciencedirect.com/topics/computer-science/reingold-layout (vaadatud apr 26, 2021).

[13] „igraph R manual pages“. https://igraph.org/r/doc/layout_with_drl.html (vaadatud apr 26, 2021).

[14] „Shiny“. https://shiny.rstudio.com/ (vaadatud apr 05, 2021).

[15] W. Chang *et al.*, *shiny: Web Application Framework for R*. 2021.

[16] R. Malan, H.-P. Company, D. Bredemeyer, ja B. Consulting, „Functional Requirements and Use Cases“, lk 8.

[17] L. Chung ja J. C. S. do Prado Leite, „On Non-Functional Requirements in Software Engineering“, *Conceptual Modeling: Foundations and Applications*, kd 5600, A. T. Borgida, V. K. Chaudhri, P. Giorgini, ja E. S. Yu, Toim Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, lk 363–379.

[18] „Introduction to visNetwork“. https://cran.r-project.org/web/packages/visNetwork/vignettes/Introduction-to-visNetwork.html (vaadatud apr 20, 2021).

[19] C. Sievert *et al.*, *plotly: Create Interactive Web Graphics via „plotly.js“*. 2021.

[20] „centrality: Calculate node and edge centrality in tidygraph: A Tidy API for Graph Manipulation“. https://rdrr.io/cran/tidygraph/man/centrality.html (vaadatud apr 20, 2021).

[21] L. C. Freeman, „A Set of Measures of Centrality Based on Betweenness“, *Sociometry*, kd 40, nr 1, lk 35–41, 1977, doi: 10.2307/3033543.
[22] W. Chang, B. B. Ribeiro, RStudio, A. S. (AdminLTE theme for Bootstrap), ja A. S. I. (Source S. P. font), *shinydashboard: Create Dashboards with „Shiny“*. 2018.
[23] V. Perrier *et al.*, *shinyWidgets: Custom Inputs Widgets for Shiny*. 2021.

Lisad

I. Githubi repositooriumid

Trajectoriese projekt on lõputöö esitamise eesmärgiks kloonitud isikliku repositooriumisse, millele pääseb ligi lingil https://github.com/Patricklomp/VisualisingHealthTrajectories.

Trajectoriese repositoorium oli töö esitamise hetkel privaatne, kuid repositoorium avalikustatakse hiljem aadressil https://github.com/EHDEN/Trajectories.

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Patrick Lomp,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose **Diagnoosi trajektooride visualiseerimine**, mille juhendajateks on **Raivo Kolde ja Maarja Pajusalu**, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Patrick Lomp

07.05.2021