TARTU ÜLIKOOL
Arvutiteaduste instituut
Informaatika õppekava

Sharon Süvari

Ravitrajektooride automaatne identifitseerimine ja puhastamine

Bakalaureusetöö (9 EAP)

Juhendaja Raivo Kolde, PhD

Tartu 2024

Ravitrajektooride automaatne identifitseerimine ja puhastamine

Lühikokkuvõte: Selle bakalaureusetöö eesmärk on luua lahendus millega andmemudeli OMOP CDM kujul raviandmetest leida uuritavale valimile spetsiifilisi ravitrajektoore. Eesmärgi saavutamiseks kasutatakse R programmeerimiskeelele loodud pakette Trajectories ja CohortContrast. Käesolev töö jaotub teoreetiliseks ja praktiliseks osaks, esimeses tutvustatakse ühtset raviandmete mudelit, kasutatavaid andmeid ja pakette, ning teises osas autori loodud lahendust ja sellega saadud tulemusi.

Võtmesõnad: OMOP CDM, ravitrajektoorid, ravisündmused

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

Automatic identifying and cleaning of treatment trajectories

Abstract: The goal of this Bachelor's thesis is to create a method to find treatment trajectories that are specific to a certain sample that is generated from a health data database following the structure of OMOP CDM. To achieve this goal, R packages Trajectories and CohortContrast will be used. This work is divided into two parts, the first one will describe a uniform health data model, used data and packages, and the second one will cover the created method and gathered results.

Keywords: OMOP CDM, treatment trajectories, treatment events

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

Sisukord

Sissejuhatus

Üle maailma kasutatakse raviandmete registreerimiseks elektroonilisi lahendusi ning selline hulk andmeid oleks väärtuslikuks ressursiks teadustöödele, mille eesmärgiks on leida tõestusi päris maailma andmete põhjal parimate ravimeetodite tuvastamiseks. Ryan ja Hripcsak kirjutavad, et just selle eesmärgiga ongi teaduskogukond *Observational Health Data Sciences and Informatics* (OHDSI) loonud partnerluse *Observational Medical Outcomes Partnership* (OMOP) [1]. Kuna andmeid sooviti koguda paljudest erinevatest andmebaasidest millel kõigil on oma struktuur, loodi OMOP *Common Data Model* (CDM), ühtne andmemudel millega standardiseerida kogutud raviandmed ühtsele kujule et nende andmete põhjal oleks lihtsam teha analüüse ning luua taaskasutatavaid koode.

OHDSI Eesti võrgustikku juhib Tartu Ülikooli terviseinformaatika teadusgrupp ning nende üheks sihiks on leida ravitrajektoore ehk huvialuses valimis sagedasti esinevaid ravisündmuste ajalisi järgnevusi [2, 3]. Selle jaoks on nad loonud R programmeerimiskeelele paketi Trajectories, millega ongi võimalik tuvastada OMOP CDM kujul andmete põhjal ravitrajektoore [4]. Kuigi Trajectories paketiga on võimalik leida ravisündmuste trajektoore, annab see palju ebaspetsiifilisi tulemusi ehk selle paketiga pole võimalik tuvastada just vaatlusalusele valimile iseloomulikke sündmuste järjestusi.

Selle uurimistöö eesmärk on parandanda Trajectories paketi tööd nii, et tulemustes esinevad just huvipakkuvale valimile statistilise olulisusega määratavad eriomased ravitrajektoorid. Selle jaoks kasutatakse teist terviseinformaatika teadusgrupi R programmeerimiskeelele loodud paketti CohortContrast [5]. CohortContrasti programmiga on võimalik tuvastada OMOP CDM kujul andmetes ravisündmuseid mida esineb uuritavas valimis statistiliselt rohkem kui ülejäänud patsientidel. Käesolevas uurimuses pannakse kahe kirjeldatud paketi töö kokku nii, et tuvastatud ravitrajektoorid on iseloomulikud just uuritavale valimile. Lisaks on eesmärk seeläbi tõhustada Trajectories programmi tööd, vähendades analüüsitavaid andmeid.

Käesolev lõputöö on jaotatud kolmeks peatükiks. Esimeses peatükis tutvustatakse lähemalt OMOP CDMi struktuuri ja töös kasutatavat andmestikku. Teises peatükis kireldatakse lähemalt Trajectories ja CohortContrasti programmide tööd. Kolmas peatükk on pühendatud

autori loodud lahendusele Trajectories paketi töö parendamiseks ja sellega saadud tulemusi, seejuures keskendutakse rohkem ühele näidisuuringule ning antakse ülevaade nelja uuringu tulemustest. Tööle on juurde lisatud kolme uurimuse tulemuste joonised ja tabelid.

1. Andmed

Selles peatükis antakse ülevaade töös kasutatavast ülemaailmsest standardsest raviandmete andmemudelist ja selle andmemudeli kujule viidud Eesti rahvastiku juhuvalimi andmestikust. Lisaks tutvustatakse kohordi olemust, selle vajadust raviandmete uurimiseks ja milliseid kohorte on täpsemalt selles töös kasutatud.

1.1 OHDSI ja OMOP CDM

OHDSI (*Observational Health Data Sciences and Informatics*) on teadusele suunatud kogukond, mis loodi eesmärgiga edendada inimeste tervist, leides selleks parimaid viise raviandmete analüüsimiseks et luua teaduspõhiseid soovitusi raviotsuste tegemiseks [1]. Blacketeri sõnul on andmete põhjal järelduste tegemiseks oluline vaatlusaluste patsientide suur hulk, kuid seejuures on tihti takistuseks erinevad andmete kogumise meetodid mis võivad uurimistöödes informatsiooni tõlgendamisel sisse tuua vigu ning suurendab tehtava töö mahtu [6]. Selleks et ühtlustada raviandmete kogumist üle maailma, on OHDSI loonud OMOP (*Observational Medical Outcomes Partnership*) CDM-i (*Common Data Model*), ühtse andmemudeli mis võimaldab täielikult standardiseerida kogutud raviandmeid et nende põhjal oleks võimalik teha statistiliselt olulisi, võrreldavaid ja reprodutseeritavaid järeldusi.

OMOP CDMis on raviandmed ühtlustatud Standardiseeritud Sõnastikena (*Standardized Vocabularies*), mis kõik on teisendatud ühisele formaadile ning mille allikaks on peamiselt erinevad üldkasutatavad standardid, näiteks WHO ICD (*World Health Organization International Classification of Disease*) [7]. Raviandmeid esitatakse Standardiseeritud Sõnastikes ühtsete reeglite ja tunnustega kontseptsioonidena, mis muu hulgas sisaldavad identifikaatorit, inglisekeelset nime ja allikaks olnud sõnastiku nimetust. Kirjeldatud standard on mõeldud olema laiaulatuslik ehk mis tahes andmebaasi igale raviandmele leidub OMOP CDMi sõnastikes standardkontseptsioon milleks see teisendada.

Raviandmete peal teadustöö tegemiseks on oluline defineerida uurimisalused ravisündmused ja valim mille kohta soovitakse informatsiooni leida. Selle jaoks on võimalik OMOP CDM kujul andmebaasides kasutada kohorte. Kristin Kostka on kohorti OHDSI raames kirjeldanud kui inimeste hulka, kes vastavad vähemalt ühele kriteeriumile mingi ajavahemiku jooksul

ning nendeks kriteeriumiteks võivad olla ravisündmused nagu konditsioonid, ravimid, protseduurid, mõõtmisetulemused, vaatlusandmed ja haigla või meditsiinipersonali külastused [8]. Huvialuste kohortide defineerimiseks on OHDSI loonud avaliku veebitarkvara ATLAS. ATLAS-es kohortide kirjeldamisel on oluline määrata esialgne sündmus, mille alusel patsient kohorti lisatakse, ning väljumiskriteerium, mille puhul kohordis olevad patsiendid sealt välja arvatakse [9, 10]. Lisaks on seal võimalik määrata erinevaid lisakriteeriumeid mille põhjal uuritavat valimit kitsendatakse, näiteks vanus, sugu või muud haigused.

1.2 Andmed

Antud uurimuses kasutatakse andmestikku milles on 10% Eesti rahvastiku juhuvalimi ehk 150 824 inimese raviandmed perioodil 2012 kuni 2019 viidud üle OMOP CDM kujule [11]. Juhuvalim on võetud MAITT andmestikust ning nende raviandmed pärinevad kolmest riiklikust elektroonilisest andmebaasist, mida kõik tervishoiuteenuse osutajad peavad kasutama: Terviseportaali epikriisid, digiretseptid ja Tervisekassa raviarved. Seejuures ei ole ükski neist surmaandmete põhiline andmebaas seega kajastub surmadest ainult 67% ehk need mis olid olemas loetletud andmebaasides.

Nendele andmetele võimaldati uurimistöö jaoks ligipääs läbi SAPU mis on võrguühenduseta ning piiratud ligipääsuga isoleeritud arvutuskeskkond [12]. Selle sees on võimalik kasutada Eesti rahvastiku valimi terviseandmeid olemasolevates programmides nagu DBeaver ja RStudio, lisaks on turvalisuse huvides andmete keskkonda viimiseks spetsiaalne andmevärav ning andmete välja toomist peab eraldi taotlema.

1.3 Kasutatud kohordid

Kohordid on OMOP CDM kujul andmete töötlemiseks üks põhilisi tööriistu ja ka selle uurimuse jaoks defineeriti ja kasutati nelja kohorti, mille andmed on kokkuvõtlikult ära toodud tabelis 1, esimeses reas on kohortide üldnimetused ja esimeses veerus neid defineerivad parameetrid.

Tabel 1. Käesolevas uurimistöös kasutatavad kohordid

	Depressioon	**Podagra**	**Rinnavähk**	**Kopsupõletik**
Esialgne sündmus	Depressiooni diagnoos, raviandmete olemasolu vähemalt 730 päeva enne ja 365 päeva peale diagnoosi	Podagra diagnoos, raviandmete olemasolu vähemalt 730 päeva enne ja 60 päeva peale diagnoosi	Rinnavähi diagnoos, raviandmete olemasolu vähemalt 365 päeva enne diagnoosi	Kopsupõletiku diagnoos, raviandmete olemasolu vähemalt 730 päeva enne ja 60 päeva peale diagnoosi
Muud kriteeriumid	Esmane juhtum	Esmane juhtum, vanuse-vahemik 18 - 70	Esmane juhtum	Esmane juhtum Vanusevahemik 18 - 70
Väljumis-kriteerium	- (vaatlusaeg 365 päeva)	- (vaatlusaeg 60 päeva)	- (vaatlusaeg 365 päeva)	- (vaatlusaeg 60 päeva)
Ravisünd-muste koodid	435220, 4034842, 4141292, 4154309, 43531624, 42872722, 4149321, 434911, 432285, 433991, 4282316, 4098302, 4077577, 4307111, 4151170, 4228802, 4336957, 4149320	440674	45571518	255848

Kokkuvõttes kirjeldab iga kohort ühte kindlat haigust ning see ilmneb eriti hästi podagra, rinnavähi ja kopsupõletiku definitsiooni puhul, kuna nendes kasutatud sündmuste koodid vastavadki kohordi nimetusele, näiteks kood 255848 on kopsupõletiku kontseptsiooni identifikaator. Erandina on depressiooni kohordil 16 sündmuse idntifikaatorit aga see on selleks et välistada erinevad ühekordsed depressiooni episoodid ja jätta sisse ainult korduvad depressiivsed häired.

Lisaks haiguse enda olemasolule on kõikidel kohortidel kriteeriumiks ka see et tegemist oleks haigestumise esmase juhtumiga vaatlusperioodi jooksul. See on oluline et välistada korduvhaigestumiste, näiteks krooniliste haiguste mitmekordset sisse lugemist. Väljumiskriteeriumid puuduvad, selle asemel on määratud vaatlusperioodiks vastavalt kas 365 või 60 päeva alates esialgse sündmuse toimumisest.

Lisaks kohortide defineerimisele on käesolevas uurimistöös kasutatavas CohortContrast programmis tulemuste leidmiseks vaja ka võrdluskohorte ning selle jaoks on paketis olemas mitu funktsiooni millest siinkohal kasutati *createControlCohortMatching*. See loob võrdluskohordi uuritava kohordi põhjal ning lisaks saab sellega määrata ka tulemuse suuruse ehk mitu kontrollpatsienti ühe sihtkohordi patsiendi kohta leitakse. Võrdluskohorti lisatakse inimesed kasutaja etteantud andmebaasist, kes ei kuulu kohorti aga vastavad sihtpatsientide vanusele ja soole.

2. Metoodika

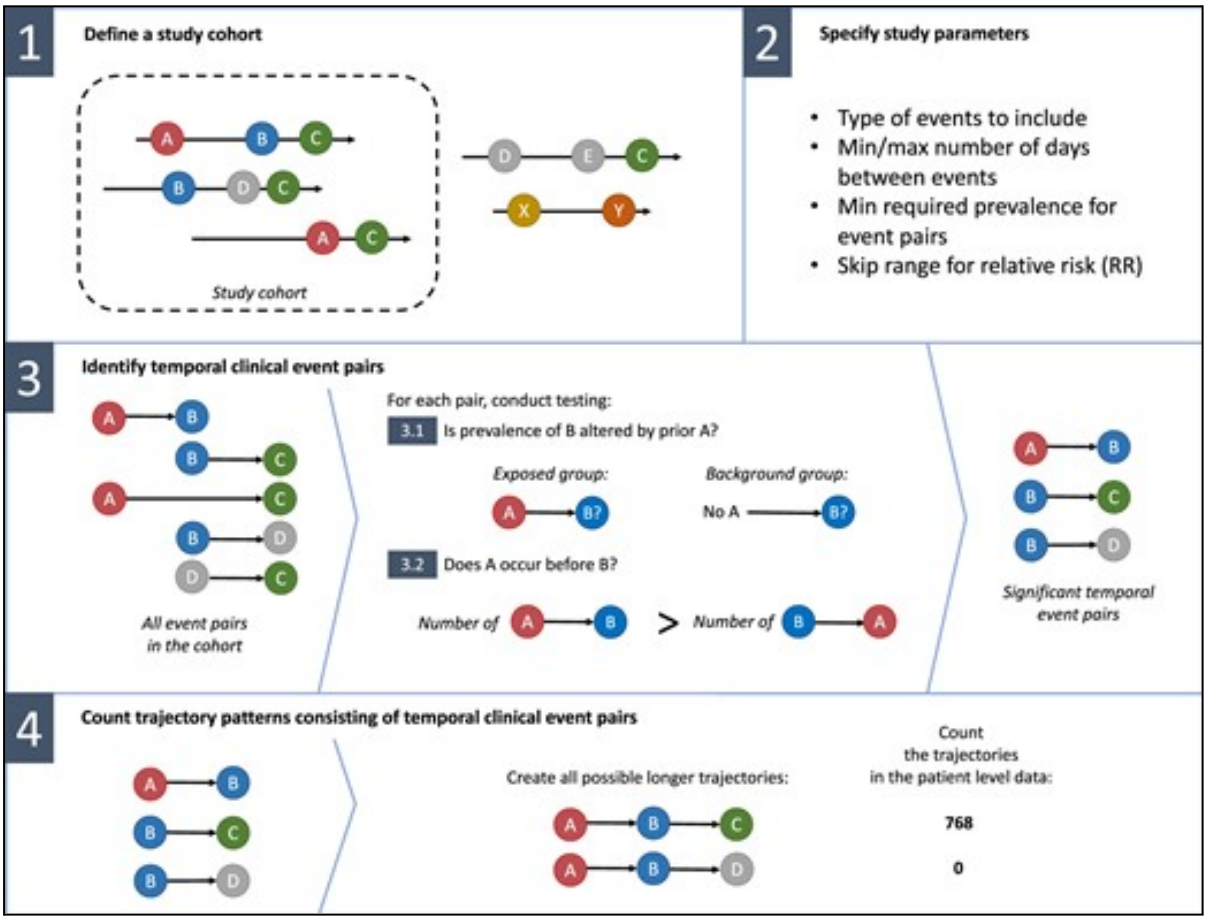
Käesolev peatükk tutvustab lähemalt käesolevas töös kasutatavaid R programmeerimiskeelele loodud pakette Trajectories ja CohortContrast ning nende tööpõhimõtteid.

2.1 R-pakett Trajectories

Üks osa käesolevast lõputööst hõlmab ravitrajektooride ehk ajalises järjestuses olevate ravisündmuste leidmist ning nende tuvastamiseks kasutatakse R-paketti Trajectories [4]. Paketi metoodika põhineb Kadri Künnapuu jt kirjutataud artiklile ja selle põhimõte on esmalt leida statistiliselt olulised sündmustepaarid et nende ühendamisel tuvastada omakorda kõikvõimalikud pikemad trajektoorid [3].

Esimeseks sammuks Trajectories paketi kasutamisel on huvialuse kohordi defineerimine ja nende ravisündmuste kategooriate määramine mille kohta soovitakse infot saada, näiteks ravimid, protseduurid või konditsioonid. Statistiliselt oluliste sündmustepaaride tuvastamiseks leitakse kahe teineteisele ajaliselt järgnenud sündmuste vaheline suhteline risk *RR*, kusjuures nende vahel võib olla ka teisi sündmuseid. Suhteline risk *RR* näitab kui palju kordi suurendab esimene sündmus (*E1*) teise sündmuse (*E2*) esinemise riski uuritavas kohordis, seejuures $RR > 1$ näitab et *E1* suurendab, $RR < 1$ et vähendab ja $RR \sim 1$ et oluliselt ei mõjuta. Suhtelise riski *RR* leidmiseks võrreldakse kaht uuritavasse kohorti kuuluvatest inimestest koosnevat gruppi. Ühes, uuritavas grupis, on patsiendid kellel esineb *E1* enne *E2*-te ja teises, taustagrupis, esineb ka *E2* ent enne seda ei esine *E1*-te, kusjuures teine grupp koosneb inimestest kes on kindlate näitajate alusel sobitatud esimese grupi patsientidega. Seejärel kasutatakse Fisheri täpset testi et tuvastada kas *E2*-te esineb uuritavas grupis oluliselt rohkem kui taustagrupis. Sündmustepaaridega, mille kohta statistiliselt oluline erinevus leiti, viiakse läbi ka binoomtest et tuvastada kas ka *E1* ja *E2* vaheline suund on statistiliselt oluline. Nii Fisheri täpse testi kui ka binoomtesti korral loetakse tulemust statistiliselt oluliseks kui p-väärtus on alla 0,05. Sündmustepaarid, mis osutuvad mõlema testiga statistiliselt oluliseks, ongi nn “ehitusblokid” ning nende koos esitamisel on võimalik leida pikemaid ravitrajektoore.

Joonisel 1 on esitatud eelneva kirjelduse skeem. Esimeses ja teises punktis määratakse huvialune kohort, kolmandas punktis on näha, et statistiliselt oluliselt järjestatud sündmustepaarid leitakse kõikvõimalike järjestatud sündmustepaaride hulgast uuritavas kohordis. Lõpuks neljandas punktis pannakse leitud sündmustepaarid kokku et moodustada pikemaid ravitrajektoore. Saadud tulemused kajastatakse suunatud graafidena.



Joonis 1. Ravitrajektooride leidmine [3].

Statistiliselt oluliste sündmustepaaride leidmisel on oluline teadvustada, et samal andmestikul tehtavate testide arvu suurenemisel suureneb ka väikeste p-väärtuste saamise tõenäosus olenemata sellest kas tegelik efekt eksisteerib [13]. Selle vastu aitab p-väärtuste korrigeerimine ning Trajectories programmis kasutatakse selleks Bonferroni korrektsiooni, mille korral korrutatakse iga sündmustepaaril tehtava testi p-väärtus kõikide testide hulgaga. Saadud tulemust saab seejärel võrrelda esialgse p-väärtusega mileks siinkohal on 0,05, kui saadud tulemus on sellest väiksem siis saab sündmustepaari lugeda statistiliselt oluliseks. See omakorda tähendab seda, et mida rohkem on sündmustepaare mida testida, seda väiksem peab olema üksiku testi tulemuse p-väärtus ja suurem otsitav efekt et sündmustepaari saaks

pidada statistiliselt oluliseks. Selle tõttu võivad tulemustest välja jääda sündmustepaarid mis väiksema testide hulga korral sisse jääksid.

2.2 R-pakett CohortContrast

Lisaks eelnevalt kirjeldatule kasutatakse töös ka R-paketti CohortContrast, mille eesmärk on leida uuritavas kohordis olevaid sündmusi mida esineb eriti rohkelt just kirjeldatud kohordis olevatel patsientidel [5], näiteks diabeedi diagnoosiga inimestel esineb palju sagedamini insuliini ravimi välja kirjutamist kui ülejäänud rahvastikus. Paketi sisendiks on kasutaja defineeritud või loodud sihtkohort andmebaasis ja selle põhjal loodav kontrollkohort, nendevaheliste erinevuste leidmiseks saab kasutaja määrata parameetrieid ja aktiveerida statistilisi teste mida programm andmete põhjal läbi viib. Parameetreid on kaks, *prevalenceCutOff* määrab kui palju kordi rohkem peab sündmust sihtkohordis rohkem esinema kui kontrollkohordis, ja *presenceFilter* kui mitmel protsendil patsientidest sihtkohordis peab sündmus esinema et seda tulemustes kajastataks. Testidest on võimalik aktiveerida z-testi ja logit-testi, mõlemad tuvastavad kahe valimi vahel statistiliselt olulisi erinevusi. Selles uurimistöös kasutatakse programmi käivitamisel tekkivat CohortContrasti R objekti, milles muu hulgas on tabel kõikide kohordis esinenud sündmuste infoga ning sooritatud z-testide tulemustega.

3. Tulemused

Selles peatükis tutvustatakse autori loodud lahendust millega muuta Trajectories programmi tulemusi uuritavale kohordile spetsiifilisemaks ja tööprotsessi kiiremaks, kasutades CohortContrasti paketiga saadavaid tulemusi. Seejärel kirjeldatakse kopsupõletiku kohordi põhjal tehtud näidisuuringut ning antakse ülevaade ka kolme teise kohordi uuringutulemustest.

3.1 R-pakettide Trajectories ja CohortContrast tulemuste ühendamine

Selle lõputöö eesmärgiks on muuta Trajectories paketi tulemusi täpsemaks uuritava kohordi suhtes ja parandada töökiirust. Trajectories programmiga on võimalik leida küll statistiliselt olulisi ravitrajektoore ent tulemused ei peegelda, millised neist on iseloomulikud just uuritavale kohordile ning lisaks muudab liigne informatsioon jooniste graafid raskesti mõistetavaks ning programmi töö aeglaseks. Nende probleemide lahendamiseks ühendati pakettide Trajectories ja CohortContrasti tulemused nii et Trajectories programmis analüüsitakse vaid neid kohordile spetsiifilisi ravisündmuseid mis tuvastakse eelnevalt CohortContrast programmiga.

Kirjeldatud lahenduse teostamiseks defineeritakse esmalt huvialune kohort veebitarkvaras ATLAS, seda kirjeldavat faili kasutatakse nii CohortContrasti kui ka Trajectories paketis andmestiku põhjal kohordi loomiseks. Seejärel luuakse CohortContrast paketi jaoks ühendus andmebaasiga ning käivitatakse programm selliste parameetritega, et testitavate sündmuste hulka jäävad ainult need mida esineb sihtkohordis vähemalt kaks korda rohkem kui kontrollkohordis ja mida esineb sihtkohordis vähemalt kümnendikul inimestest. Need parameetrid on olulised et välistada üksikuid haruldasi ravisündmuseid mis kohorti sattuda võivad.

Järgmise sammuna sooritatakse kõikide tuvastatud ja kriteeriumitele vastavate sündmuste peal z-test ning saadud tulemused salvestuvad andmetabelisse. See andmetabel koos ravisündmuste koodidega ja z-testi tulemustega imporditakse andmebaasi mille jaoks luuakse ka Trajectories paketile ühendus. Selle järel käivitataksegi Trajectories pakett autori lisatud koodilõiguga. Trajectories programm loob töö käigus tabeli kus on kõik kohordist tuvastatud sündmustepaarid ning lisatud koodilõik jätab sinna omakorda ainult need paarid kus

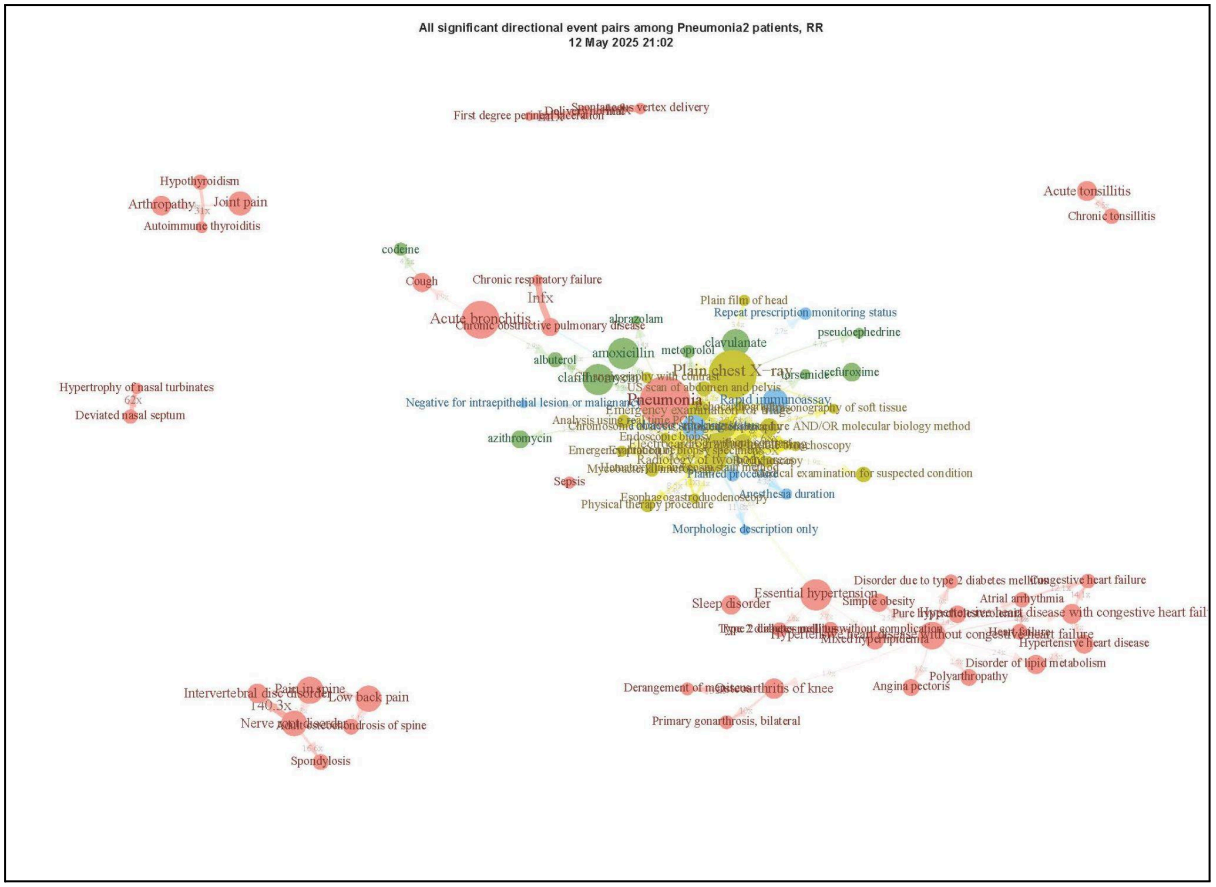
mõlemad sündmused leiduvad varem andmebaasi imporditud CohortContrasti tabelis ning on z-testi järgi statistiliselt olulised. Teisisõnu jäävad sündmustepaaride tabelisse alles vaid kohordile eriomased sündmused. Selle järel jätkub Trajectories paketi tavapärane töö, võrreldakse leitud sündmustepaaridega patsientide gruppi vastava kontrollgrupiga ja leitakse sündmuste ajalise järgnevuse olulisus. Tulemused esitatakse graafidena.

3.2 Näidisuuring kopsupõletiku kohordi põhjal

Näidisuuring teostati varem tutvustatud kopsupõletiku kohordi põhjal. Autori loodud lahenduse tulemuste võrdlemiseks käivitati esmalt ainult Trajectories programm OMOP CDM kujul MAITT andmestiku juhuvalimi peal. Programm tuvastas andmebaasist 5542 patsienti, kes defineeritud kohorti kuuluvad, ning 894 erinevat sündmustepaari, mille hulgast leiti 165 erinevat statistiliselt olulist järjestatud sündmustepaari. Kogu protsess võttis aega 20 minutit ja 20 sekundit.

Joonisel 2 on näha statistiliselt olulisi ravitrajektoore mis kirjeldatud meetodil tuvastati, iga ring tähistab erinevat ravisündmust ning ringi suurus näitab suhtelist inimeste hulka valimis kellel see ravisündmus esines. Ringide vahel olevad nooled näitavad trajektoori sündmuste esinemise järjekorda ning number noolel eelnevast sündmusest tingitud järgneva sündmuse esinemise riski suurenemist kordades. Lisaks on järgnevate värvidea tähistatud erinevad sündmuste grupid:

- Konditsioon - punane
- Vaatlusandmed - sinine
- Protseduur - kollane
- Ravim - roheline



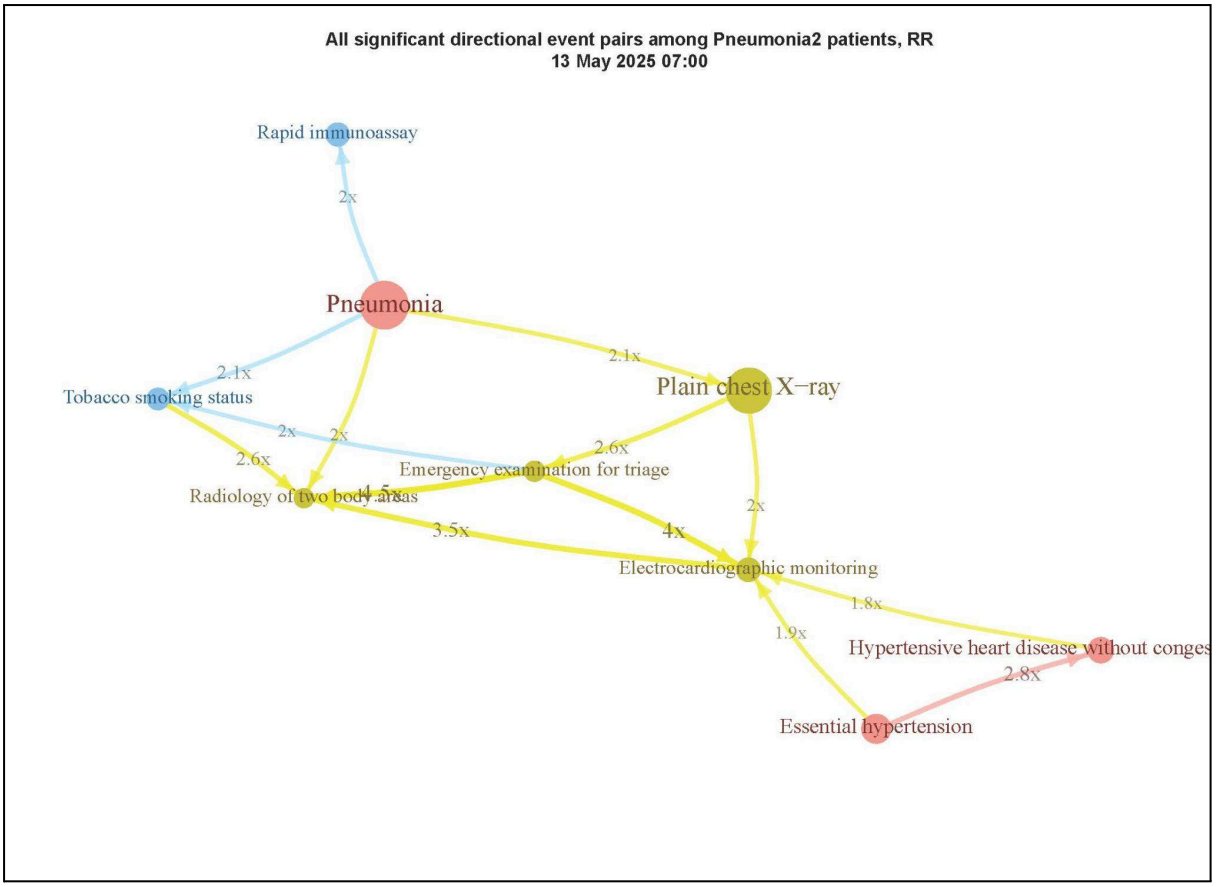
Joonis 2. Paketiga Trajectories saadud tulemus kopsupõletikuga patsientide kohordi põhjal.

Joonisel on võimalik tuvastada mitu klastrit, millest paljud ei ole otseselt kopsupõletiku diagnoosiga seotud, näiteks ülal vasakul olevad kilpnäärme- ja liigeshaigustega seotud konditsioonid ning all vasakul seljavaluga seotud tõved. Sellised trajektoorid tekivad tulemustesse kuna olenemata kohordist, kelleks siinkohal on kopsupõletikuga inimesed, võivad ilmneda ka muud laialt levinud ravitrajektoorid mis ei ole aga spetsiifilised uuritavale kohordile. Seejuures on ainult kõige suurema klastri keskel on näha uuritava kopsupõletiku (*Pneumonia*) esinemist ent "müra" ehk ebaoluliste trajektooride tõttu on päriselt huvitavad sündmused joonisel kokku pressitud ning praktiliselt pole võimalik sealt informatsiooni välja lugeda.

Autori loodud lahenduse jaoks leiti esmalt CohortContrast paketiga kopsupõletiku kohordile spetsiifilised sündmused ning enne Trajectories programmi käivitamist lisati sellele koodilõik mis CohortContrasti tulemuste järgi oma andmeid filtreerib. Analüüsitavate sündmustepaaride arv langes 110 peale ja nendest olulisteks osutus vaid 14. Kogu protsess võttis aega 7 minutit ja 25 sekundit, seejuures CohortContrasti programmil kulus 5 minutit ja

9 sekundit ning Trajectories paketil 2 minutit ja 16 sekundit. Järelikult vähenes oluliste sündmustepaaride arv 12 korda ning tulemuste leidmise aeg ligikaudu 2,74 korda võrreldes Trajectories paketi üksi kasutamisega.

Joonisel 3 on näha kirjeldatud uuringu tulemusi. Võrreldes eelmise joonisega on sellel joonisel palju vähem "müra" ning kopsupõletikuga seotud olulised ravitrajektoorid eristuvad paremini. Joonisel on näiteks näha et kopsupõletiku (*Pneumonia*) diagnoos suurendab 2,1 korda tõenäosust rinna rõntgenpildi (*Plain chest X-ray*) ja 2 korda imuunanalüüsi (*Rapid immunoassay*) tegemist ning 2,1 korda suitsetamisstaatuse (*Tobacco smoking status*) tuvastamise toimumist. Võrreldes eelmise joonisega puuduvad nüüd eraldiseisvad klastrid mis ei ole kopsupõletikuga seotud nagu varemmainitud kilpnäärme, liigeste ja seljaga seotud vaevused. Lisaks on kopsupõletiku enda klaster palju hõredam, sealhulgas puuduvad trajektooridest täielikult ravimid. Lisaks on näha all paremal kõrgvererõhktõvega (*Essential hypertension* ja *Hypertensive heart disease without congestion*) seotud trajektoore mis ükski ei ole otseselt seotud kopsupõletiku sündmusega. Seda võib selgitada asjaoluga et kopsupõletikku jäävad pigem vanemad inimesed kellel lisaks esinevad tihtipeale ka vererõhuhäired.



Joonis 3. Pakettide Trajectories ja CohortContrast koos kasutamise tulemus kopsupõletikuga patsientide kohordi põhjal.

Lisaks sellele et käesolev lahendus muutis tulemusi täpsemaks ja protsessi kiiremaks võrreldes ainult Trajectories programmi kasutamisega, on siinkohal juures ka üks ravitrajektoor mida algsetes tulemustes ei olnud, nimelt trajektoor varemmainitud sündmusest "kõrgvererõhktõbi ilma südamepuudulikkuseta (*Hypertensive heart disease without congestion*)" sündmuseni "elektrokardiograafia monitoorimine (*Electrocardiographic monitoring*)". Selle sündmustepaari lisandumine tulemustesse on tingitud Bonferroni korrektsioonist mida Trajectories programmis kasutatakse. Kuna CohortContrasti tulemuste järgi sündmuste filtreerimine on vähendanud analüüsitavate sündmustepaaride ehk tehtavate testide arvu 12 korda, siis sellest tingituna võib iga testi p-väärtus olla kõrgem kui eelmises analüüsis et seda loetaks statistiliselt oluliseks. Teisisõnu tänu sündmustepaaride vähenemisele on tõusnud statistiliste testide tundlikkus. Sellest võib järeldada et autori esitatud lahenduse abil on võimalik tuvastada kohordile spetsiifilisi ravitrajektoore mis Trajectories programmi üksi kasutades jäid tulemustest välja kuna testide rohkuse tõttu ei osutunud nad siis statistiliselt oluliseks.

3.3 Ülevaade nelja kohordi tulemusest

Siinkohal esitatakse ülevaade podagra, rinnavähi, depressiooni ja kopsupõletiku kohortide põhjal tehtud analüüside tulemustest. Tabelis 2 on välja toodud tulemused mis saadi kui kohorte analüüsiti vaid Trajectories paketiga, esimeses reas programmi tööle kulunud aeg, teises kõik analüüsitud sündmustepaarid ja kolmandas on analüüsidega tuvastatud statistiliselt olulised sündmustepaarid vastavas kohordis.

Tabel 2. Nelja erineva kohordi tulemused Trajectories paketiga analüüsides.

	Podagra	**Rinnavähk**	**Depressioon**	**Kopsupõletik**
Aeg (min:sek)	07:47	06:07	32:05	20:20
Analüüsitud sündmuste-paarid	1058	642	2757	894
Tuvastatud statistiliselt olulised sündmuste-paarid	80	159	394	165

Tabelis 3 on esitatud vastavad tulemused mis saadi kui kasutati autori lahendust ehk Trajectories paketiga leitud sündmustepaaridest jäeti analüüsideks alles vaid need kus paari mõlemad sündmused on CohortContrasti paketi tulemuste järgi kohordile eriomased. Kulunud aja sisse on arvestatud nii CohortContrasti kui ka Trajectories programmi tööaeg. Lisaks on viimases reas välja toodud lisandunud sünmdustepaarid mis Bonferroni korrektsioonist tingituna ei tulnud välja kui kasutati analüüsimiseks ainult Trajectories paketti.

Tabel 3. Nelja erineva kohordi tulemused CohortContrasti ja Trajectories paketi analüüside tulemuste ühendamisel.

	Podagra	**Rinnavähk**	**Depressioon**	**Kopsupõletik**
Aeg (min:sek)	04:20	08:19	05:37	07:25
Analüüsitud sündmuste-paarid	16	393	159	110
Tuvastatud statistiliselt olulised sündmuste-paarid	2	105	35	14
Lisandunud sündmuste-paaride hulk	0	14	5	1

Selleks et visualiseerida tulemuste vahelisi erinevusi, on tabelis 4 välja toodud kahe eelneva tabeli vastavate tulemuste suhted, täpsemalt mitu korda vähenes autori lahendusega kulunud aja hulk ning analüüsitud ja tuvastatud sündmustepaaride arv võrreldes Trajectories paketi üksi kasutamisega. Lisaks on viimases veerus esitatud iga muutuse keskmine. Tulemused on ümardatud kahe komakohani.

Tabel 4. Algsete tulemuste ja autori lahenduse tulemuste suhe.

	Podagra	**Rinnavähk**	**Depressioon**	**Kopsu-põletik**	**Keskmine**
Aeg	1,80	0,74	5,71	2,74	2,75
Analüüsitud sündmuste-paarid	66,13	1,64	17,34	8,13	23,31
Tuvastatud statistiliselt olulised sündmuste-paarid	40	1,51	11,26	11,79	16,14

Tabelis 4 on näha et iga kohordi puhul vähendas autori lahendus nii analüüsitud sündmustepaaride kui ka lõpuks tuvastatud oluliste sündmustepaaride hulka. Küll aga suurenes rinnavähi kohordi korral programmide tööle kulunud aeg erinevalt teistest kohortidest ning samuti on mõlemate sündmustepaaride hulga muutus tunduvalt väiksem. Selle üks võimalik selgitus on see et rinnavähk on pikaajaline ja neist kolmest kõige spetsiifilisem diagnoos eelkõige raviprotseduuride ja nende rohkuse tõttu ning seega sellesse kohorti kuuluvate inimeste ravitrajektoorid eristuvad ülejäänud andmestikust piisavalt et CohortContrastiga filtreerimine ei too nii suuri tulemusi kui levinumad ja väheste spetsiifiliste raviprotseduuridega podagra, depressioon ja kopsupõletik.

Peale esitatud tabelite on käesolevale uurimistööle lisatud ka podagra, rinnavähi ja depressiooni kohortide analüüside tulemused ja seda nii Trajectories paketi üksi kui ka autori lahenduse kasutamisel. Samuti on tabelitena välja toodud rinnavähi ja depressiooni kohortide sündmustepaarid, mis Trajectories paketi kasutamisel välja jäid ent koos CohortContrasti tulemuste ühendamisega statistiliselt oluliste sündmustepaaridena välja tulid.

Kokkuvõte

Selle lõputöö eesmärk oli parandada OMOP CDM kujul raviandmete töötlemiseks loodud paketi Trajectories tulemusi nii, et sellega oleks võimalik tuvastada just uuritavale kohordile spetsiifilisi ravitrajektoore ja vähendada ka programmi tööle kuluvat aega. Selle jaoks loodi lahendus paketi CohortContrast abiga, jättes Trajectories programmiga analüüsitavate sündmustepaaride hulka vaid need kus mõlemad sündmused on CohortContrasti programmiga tuvastatud kui just uuritavale kohordile iseloomulikud sündmused.

Loodud lahenduse tulemusel vähenes kõigi nelja uuritud kohordi analüüsitavate ja tuvastatud sündmustepaaride hulk ning tööle kulunud aeg vähenes kolme kohordi puhul. Keskmiselt vähenes programmi tööaeg 2,75, analüüsitud sündmustepaaride arv 23,31 ja tuvastatud statistiliselt oluliste sündmustepaaride hulk 16,14 korda. Tänu sellele tulid välja uuritavatele kohortidele spetsiifilisemad ravitrajektoorid võrreldes Trajectories paketi üksi kasutamisega ning tulemusi oli joonisel kergem jälgida. Lisaks esines kolme kohordi korral olukord, kus tulemustesse jäid sündmustepaarid mida algse analüüsiga ei tuvastatud kuna Bonferroni korrektsioonist tingituna olid statistilised testid tundlikumad kui tehtavaid teste oli vähem.

Töö autor pakub edasiseks tööks välja loodud lahendusest paketi loomise et seda oleks võimalik mugavalt kasutada ning ka tulemustest loodavate jooniste täiustamist et programmiga tuvastatud ravitrajektoorid tuleksid selgemalt nähtavale. Tulemuste põhjal soovitab autor veel uurida, et kuidas mõjutab kohordi olemus loodud lahendusega saadud tulemusi võrreldes sellega kui kohorti analüüsitakse vaid Trajectories paketiga.

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

Viidatud kirjandus

[1] Hripcsak, G., Ryan, P. The OHDSI Community. *The Book of OHDSI, 2021*. https://ohdsi.github.io/TheBookOfOhdsi/OhdsiCommunity.html (08.12.2024).

[2] Terviseinformaatika töörühm https://health-informatics.cs.ut.ee/terviseinformaatika-tooruhm/ (02.05.2025).

[3] Künnapuu, K., Ioannou, S., Ligi, K., Kolde, R., Laur, S., Vilo, J., Rijnbeek, P. R, Reisberg, S. Trajectories: a framework for detecting temporal clinical event sequences from health data standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *JAMIA Open,* 2022, Vol 5. https://academic.oup.com/jamiaopen/article/5/1/ooac021/6549728 (08.12.2024).

[4] Trajectories. https://github.com/EHDEN/Trajectories/ (01.05.2025).

[5] CohortContrast. https://github.com/HealthInformaticsUT/CohortContrast (05.01.2025).

[6] Blacketer, C. The Common Data Model. *The Book of OHDSI,* 2021. https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html (08.12.2024).

[7] Reich, C., Ostropolets, A. Standardized Vocabularies. *The Book of OHDSI,* 2021. https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html (10.05.2025).

[8] Kostka, K. Defining Cohorts. *The Book of OHDSI,* 2021. https://ohdsi.github.io/TheBookOfOhdsi/Cohorts.html (05.01.2025).

[9] Schuemie, M., DeFalco, F. OHDSI Analytics Tools. *The Bookf of OHDSI,* 2021. https://ohdsi.github.io/TheBookOfOhdsi/OhdsiAnalyticsTools.html#atlas (01.05.2025).

[10] Atlas. https://atlas-demo.ohdsi.org/#/home (10.05.2025).

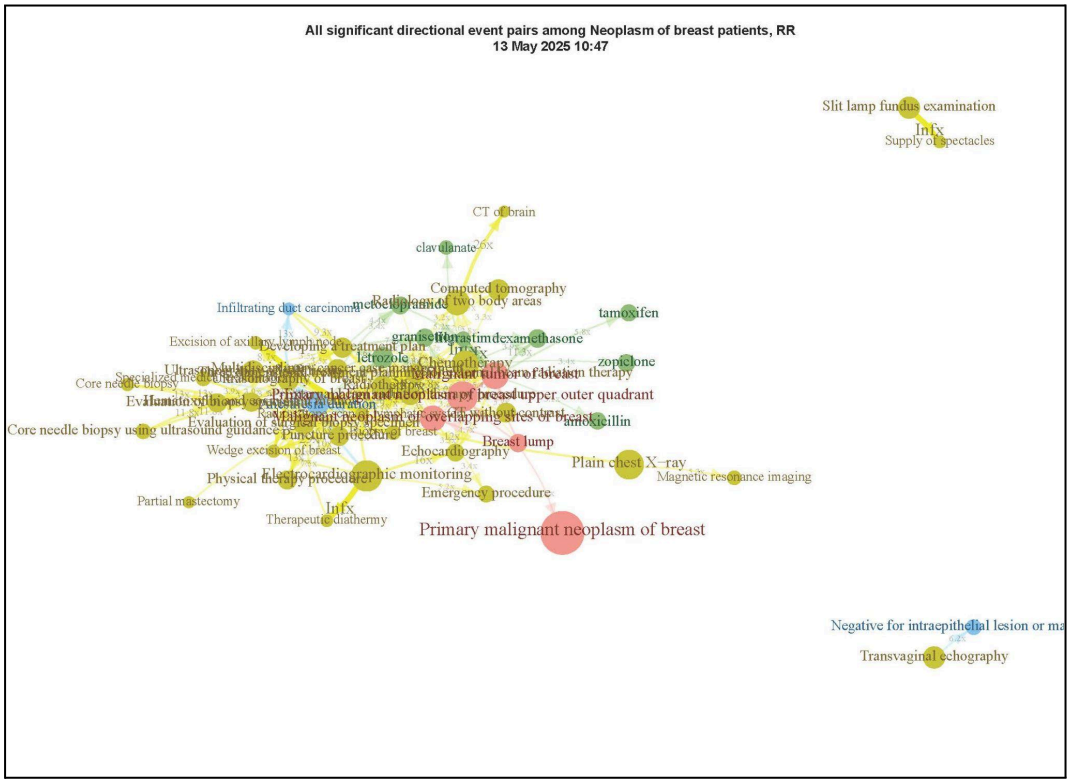
[11] Oja, M., Tamm, S., Mooses, K., Pajusalu, M., Talvik, H.-A., Ott, A., Laht, M., Malk, M., Lõo, M., Holm, J., Haug, M., Šuvalov, H., Särg, D., Vilo, J., Laur, S., Kolde, R., Reisberg, S. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. *Jamia Open,* 2023, Vol 6. https://academic.oup.com/jamiaopen/article/6/4/ooad100/7459333#428818786 (07.05.2025).

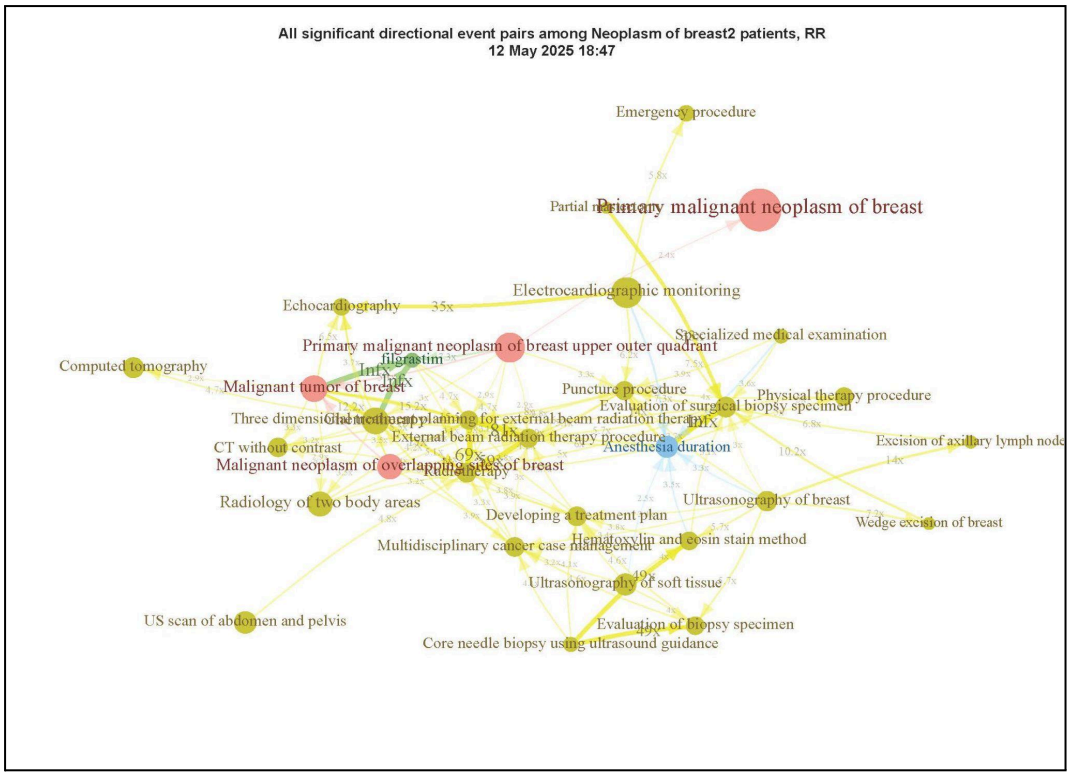
[12] Mis on SAPU? https://sapu.cs.ut.ee/index.php (01.05.2025).

[13] Armstrong, R., A. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics.* 2014, Vol 34, lk 502-508. https://onlinelibrary.wiley.com/doi/10.1111/opo.12131 (14.05.2025).

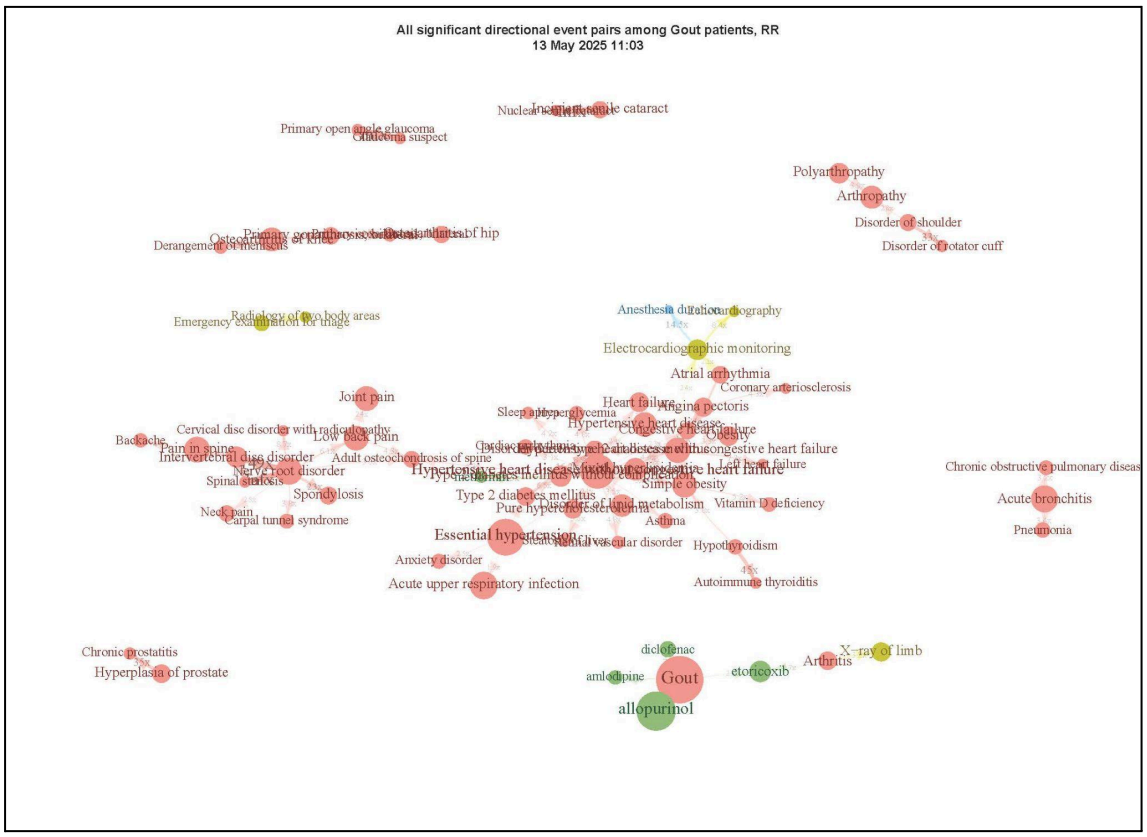
Lisad

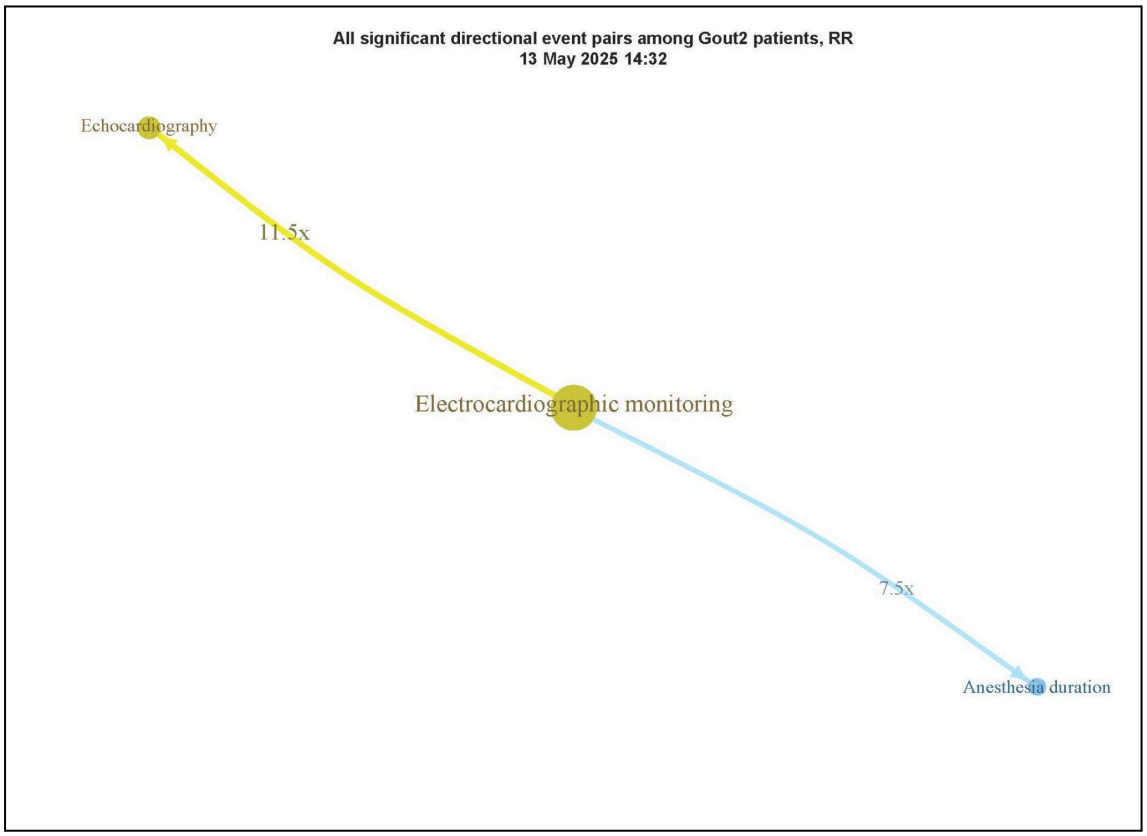
1. Rinnavähi kohordi analüüsi tulemused ainult Trajectories paketiga ja autori loodud lahendusega



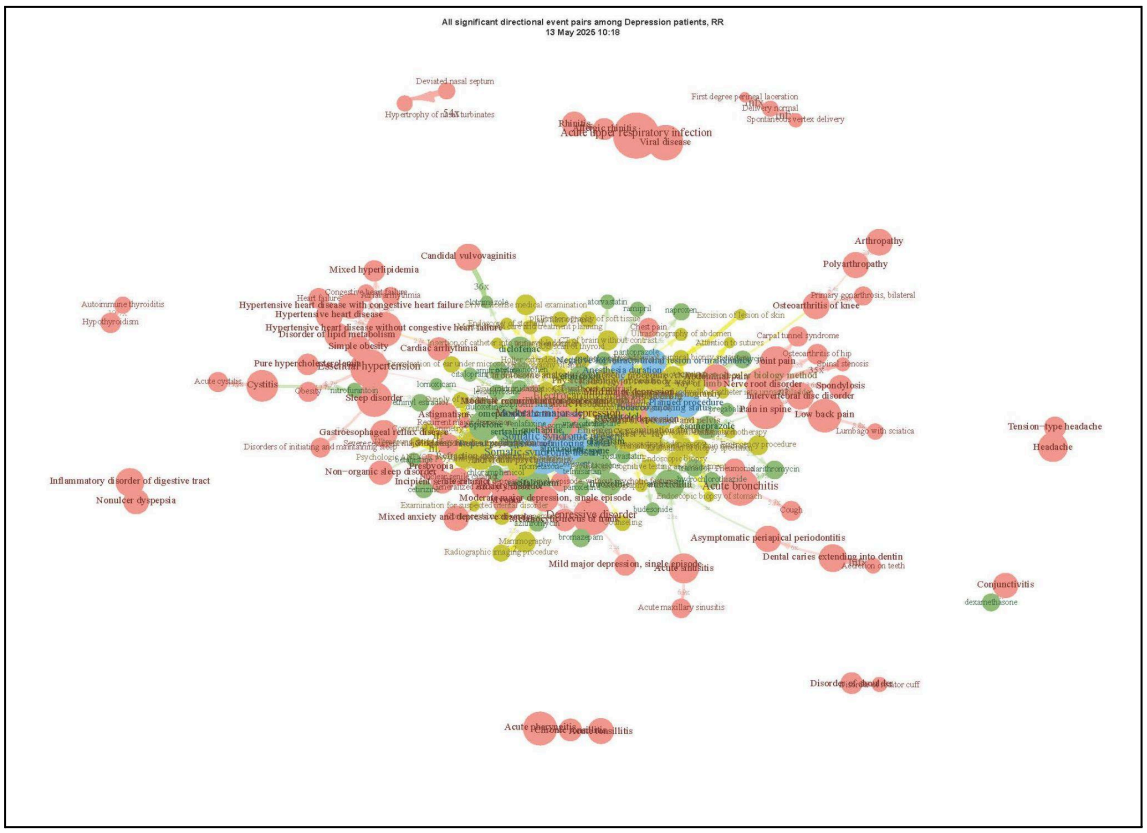


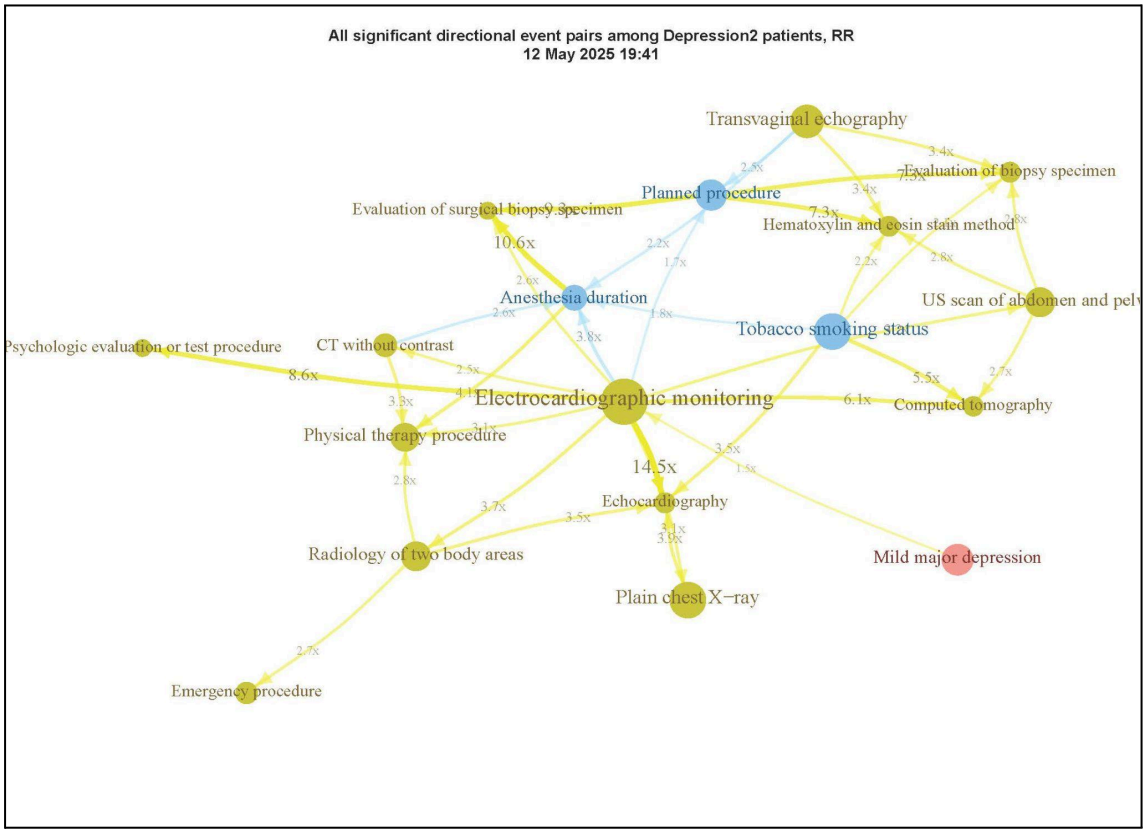
2. Podagra kohordi analüüsi tulemused ainult Trajectories paketiga ja autori loodud lahendusega





3. Depressiooni kohordi analüüsi tulemused ainult Trajectories paketiga ja autori loodud lahendusega





4. Depressiooni kohordi analüüsimisel autori lahendusega lisandunud statistiliselt olulised sündmustepaarid

E1 kood	E2 kood	E1 nimetus	E2 nimetus
4 027 510	1 176 109	Transvaginaalne ultraheli	Anesteesia kestus
43 054 909	1 176 109	Tubaka suitsetamise staatus	Anesteesia kestus
43 054 909	4 014 506	Tubaka suitsetamise staatus	Hematoksüliin-eosiinvärving
43 054 909	40 480 861	Tubaka suitsetamise staatus	Biopsiamaterjali hindamine
44 791 515	4 158 569	Kahe kehapiirkonna radioloogia	Erakorraline protseduur

5. Rinnavähi kohordi analüüsimisel autori lahendusega lisandunud statistiliselt olulised sündmustepaarid

E1 kood	E2 kood	E1 nimetus	E2 nimetus
4 112 853	4 163 903	Rinnavähk (tuumor)	Kompuutertomograafia kontrastaineta
4 112 853	4 300 757	Rinnavähk (tuumor)	Kompuutertomograafia
4 163 903	1 242 725	Kompuutertomograafia kontrastaineta	Kiiritusravi
4 187 078	4 312 604	Elektrokardiograafia monitoorimine	Lumbaarpunktsioon
4 187 850	1 304 850	Esmane vähkkasvaja (neoplasma) rinna ülemises välimises neljandikus	filgrastim (ravim)
4 187 850	4 162 253	Esmane vähkkasvaja (neoplasma) rinna ülemises välimises neljandikus	Esmane vähkkasvaja (neoplasma) rinnas
4 187 850	4 312 604	Esmane vähkkasvaja (neoplasma) rinna ülemises välimises neljandikus	Lumbaarpunktsioon
4 264 054	4 312 604	Rinna ultrasonograafia	Lumbaarpunktsioon
4 273 629	4 163 903	Keemiaravi	Kompuutertomograafia kontrastaineta
4 305 221	1 242 725	Alakeha ja vaagnapiirkonna ultraheli	Kiiritusravi
40 480 642	4 022 110	Biopsia nõelaga ultraheli abil	Raviplaani koostamine
40 480 642	44 808 909	Biopsia nõelaga ultraheli abil	Multidistsiplinaarne vähijuhtumi haldamine
44 791 515	1 242 725	Kahe kehapiirkonna radioloogia	Kiiritusravi
44 791 515	4 141 448	Kahe kehapiirkonna radioloogia	Välise kiiritusravi protseduur

7. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Sharon Süvari,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Ravitrajektooride automaatne identifitseerimine ja puhastamine", mille juhendaja on Raivo Kolde, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Sharon Süvari
15.05.2025