Uudne lähenemine – OMOP-andmemudelil põhinevad terviseuuringud

Sulev Reisberg[1, 2], Kerli Mooses[1], Raivo Kolde[1], Lenne-Triin Kõrgvee[3, 4], Jaak Vilo[1, 2]

Eesti Arst 2024; 103(9):420–429

Saabunud toimetusse: 10.04.2024
Avaldamiseks vastu võetud: 21.06.2024
Avaldatud internetis: 23.09.2024

[1] Tartu Ülikooli arvutiteaduse instituut,
[2] STACC,
[3] Tartu Ülikooli bio- ja siirdemeditsiini instituut,
[4] Tartu Ülikooli Kliinikumi vähikeskus

Kirjavahetusautor:
Sulev Reisberg
sulev.reisberg@ut.ee

Võtmesõnad:
OMOP, OHDSI, päriselu terviseandmed, terviseuuring

Meditsiin, tervishoid ja neid valdkondi reguleerivad poliitikad lähtuvad tõenduspõhisusest. Tõenduse loomiseks kasutatakse prospektiivsete juhuslikustatud kliiniliste uuringute kõrval järjest enam kliinilise ravitegevuse käigus tekkivaid päriselu terviseandmeid. See eeldab standardseid sõnastikke ja hästi struktureeritud andmeid. Üheks niisuguseks standardiks on tõusmas OMOP (*Observational Medical Outcomes Partnership*) andmemudel. OMOP-andmebaaside struktuurne ja semantiline ühetaolisus lihtsustab standardsete analüüside tegemist, mille abil saab uurimisküsimuste vastused kiiresti ja odavamalt. Rahvusvahelises koostöös OMOP-andmebaaside kasutamine tagab andmete kaitse, kuna teisele osapoolele väljastatakse üksnes isikustamata agregeeritud analüüsi tulemusi. Nüüdseks on 12% maailma rahvastiku terviseandmed viidud OMOP-kujule. Ka Eestil on võimalus seda kasutades olla päriselu andmete kasutamise arengu esirinnas ning suurendada riigisisese ja rahvusvahelise teaduskoostöö võimalusi.

Meditsiin, tervishoid ja neid valdkondi reguleerivad poliitikad lähtuvad tõenduspõhisusest ehk teaduslike ja statistikal põhinevate meetoditega saadud ning kinnitatud tõenditest (1). Tõenduse kogumiseks on mitmeid viise. Klassikaliselt on uute ravimite ja ravimeetodite ohutuse ning kasulikkuse kohta tõenduspõhise info kogumise kuldstandard olnud prospektiivne pimendatud juhuslikustatud kontrolluuring (2). Need uuringud on enamasti kallid ja aeganõudvad ega pruugi kohati väga jäigalt valitud valimi tõttu olla esinduslikud. Seetõttu on otsitud klassikalistele meetoditele alternatiive.

Viimastel aastakümnetel on erinevatesse andmekogudesse tavapäraste tervishoiuteenuste osutamise käigus tekkinud järjest enam elektroonilisi terviseandmeid ehk nn päriselu terviseandmed (ingl *real world data*). Sellised andmed on näiteks analüüside ja uuringute vastused, epikriisid, saatekirjade vastused, raviarved, retseptidega seotud info, aga ka geeniinfo, aktiivsusmonitoridest pärit info jms (3). Päriselu terviseandmed sisaldavad väärtuslikku informatsiooni, mida saab edukalt rakendada ravimite, raviprotseduuride, -protsesside, -juhiste ja -teekondade hindamiseks ning edasiarendamiseks. Varem on välja toodud, et päriselu terviseandmetel läbi viidud uuringud aitavad täiendada juhuslikustatud kontrolluuringutes leitut ning seeläbi panustada täielikuma tõenduse saamisesse (4).

Eestis kogutakse erinevad terviseandmed tervishoiuteenuste pakkujate juures (haiglad, perearstid) ning edastatakse need vastavalt kehtestatud korrale erinevatesse kesksetesse terviseandmekogudesse. Seda on põhjalikult käsitletud Eesti tervisesüsteemi ülevaateraportis (5). Näiteks analüüside vastused, epikriisid ja saatekirjade vastused edastatakse tervise infosüsteemi, raviarved Tervisekassa andmekogusse, retseptidega seonduv info retseptikeskusse, röntgeniülesvõtted pildipanka, kasvajatega seotud info vähiregistrisse jne. Need kesksed terviseandmebaasid täidavad erinevaid eesmärke. Ühelt poolt on nad vajalikud igapäevaseks asutusteüleseks kliiniliseks tööks. Teisalt saab erinevates kesksetes terviseandmebaasides talletatavat infot koondada ning rakendada tõenduse saamiseks nii ravi kui ka tervishoiusüsteemi korraldust puudutavate otsuste tegemisel (6). Kui päriselu terviseandmete esmane eesmärk on olla arstile patsiendi ravimisel infoallikaks, siis käesoleva artikli keskmes

on nende andmete uuel eesmärgil ehk nn teisene kasutus.

PÄRISELU TERVISEANDMETE KASUTAMINE TEADUSUURINGUTES – EELISED JA KITSASKOHAD

Päriselu terviseandmeid kasutatakse järjest enam tervishoiukorralduslike otsuste tegemisel, aga ka kliinilised ohutus- ja efektiivsusuuringud (sh ravimiuuringud) baseeruvad juba sageli kliinilise tegevuse käigus tekkinud andmetel. Võrreldes juhuslikustatud kliiniliste uuringutega on päriselu andmetel põhinevatel uuringutel mitmeid eeliseid (4). Esiteks kasutatakse ära juba olemasolevaid andmeid ja see hoiab oluliselt kokku andmete kogumiseks kuluvat ressurssi. Teiseks kajastavad need paremini tegeliku elu protsesse ja trajektoore, kaasuvaid haigusi, samaaegseid raviskeeme erineva soo, vanuse ja terviseseisundiga isikutel. Kolmandaks on päriselu andmetel võimalik vaadata seoseid, mille kohta juhuslikustatud uuringuid läbi viia ei ole võimalik – näiteks harvikhaigused või eetilistest kaalutlustest lähtuvalt raskesti teostatavad uuringud nagu ravimite või vaktsiinide mõju rasedusele (7). Neljandaks on päriselu andmed enamasti pärit pikemast ajaperioodist kui on tavaliselt kliinilise uuringu kestus. Nende abil on võimalik uurida väga erinevaid küsimusi, näiteks haigestumust, haiguste levimust, riskitegureid, ravitulemusi, trende, tervishoiuressursside kasutust, kulusid, ravimite ohutust ja ravimustreid (4). Seetõttu nähakse päriselus kogunevate terviseandmete teiseses kasutamises väga suurt potentsiaali tõenduse kogumiseks ning tõenduspõhiste otsuste toetamiseks.

Samas on päriselu terviseandmete kasutamisel ka puudusi (3, 4). Üheks suuremaks puuduseks võib pidada andmete ebaühtlast kvaliteeti, sest algne info on kogutud patsiendi ravi eesmärgil ega pruugi sisaldada vajalikku infot sellise detailsusega nagu kliinilistes uuringutes. Lisaks võib andmestikus olla info sellisel kujul, mis nõuab väga palju täiendavat tööd, enne kui seda on võimalik analüüsis kasutada. Näiteks on kaebused või lihtsamad mõõtmised nagu vererõhk ja kaal tihti lisatud epikriisidesse tekstiväljadel, kus nad on raviarstile küll nähtavad, kuid selliste faktide ja arvuliste väärtuste uuringus kasutamiseks on esmalt tarvis need tekstist usaldusväärselt üles leida ning seejärel teisendada analüüsitavale kujule.

Probleeme võib tekitada ka päriselu andmete killustatus, kuna need pärinevad tervishoiusüsteemi eri osapooltelt (nt haiglad, perearstid, apteegid, laborid) ja infosüsteemidest (raviarvete andmekogudest, erinevatest spetsiifilistest registritest), mis kõik on erineva toimimisloogikaga. Võimalikult tõese info kogumiseks võib olla vaja need infokillud ühendada (2, 4). Näiteks on laborianalüüside teostamise sagedust võimalik uurida Tervisekassa raviarvete alusel, kuid analüüside tulemuste jaoks peab vaatama hoopis tervise infosüsteemi. Uuringu läbiviija jaoks tähendab andmebaaside ühendamine täiendavat tööd ja ajakulu. Samas on isegi erinevate andmebaaside ühendamise korral uuringu tegemiseks kuluv aeg oluliselt lühem kui andmeid ise kogudes. Mõnedes riikides, kus puuduvad riiklikud patsiendi identifikaatorid või patsiendid liiguvad erinevate regioonide või süsteemide vahel, võib andmebaaside ühendamine osutuda keeruliseks, kui mitte võimatuks.

Lisaks võib päriselu terviseandmete kasutamist rahvusvahelistes uuringutes raskendada erinevate koodisüsteemide kasutamine nii haiguste, ravimite kui ka protseduuride dokumenteerimisel. Samuti võivad rahvusvahelist koostööd piirata õigusaktid, mis ei luba arusaadavatel põhjustel terviseandmeid teistesse riikidesse väljastada. Sugugi mitte vähem olulised ei ole andmekaitselised ja eetilised aspektid, sest erinevalt kliinilistest uuringutest ei ole patsiendid päriselu terviseandmete teiseseks kasutamiseks oma nõusolekut andnud. Kuna kõigilt patsientidelt nõusoleku võtmine polegi realistlik, on niisuguste uuringute puhul eriti oluline roll eetikakomiteedel, kes iga uuringu põhjendust ja vajalikkust (sh õigusnormide järgimist ja andmekaitselisi asjaolusid) hindavad. Samuti on oluline, et andmete edastamine uurimismeeskonnale ja edasine andmeanalüüs toimuks turvalisel viisil.

Päriselu terviseandmete teisesel kasutamisel on enne masintöötluse ja statistika tegemist tarvis alusandmed kõigepealt eri andmekogudest üles leida, väljastada, puhastada ja andmekogude ühendamisel ka ühtlustada. Enamasti on see väga suur töö ja nõuab enne analüüsimist eri spetsialistide ressurssi nii tehnilise (süsteemiadminist-

raatorid, andmehaldurid, andmeinsenerid, statistikud, keeletehnoloogid jt) kui ka sisulise poole pealt (arstid, tervishoiuspetsialistid). Oluline on märkida, et andmete puhastamiseks on üldjuhul tarvis tunda vastavaid protsesse, mille käigus algandmed on tekkinud. Seetõttu on keeruline ette kujutada, et andmete puhastamist saaks teostada ilma valdkonna ekspertide kaasamiseta või veelgi enam, osta teenusena sisse väljastpoolt Eestit. Mööda ei saa vaadata asjaolust, et kuna Eestis tuleb iga päriselu terviseandmetega tehtava teadusuuringu läbiviimiseks taotleda iga kord uut andmeväljastust, peab vaatamata sellele, et andmeandjad ja andmete probleemid on enamasti ühed ja samad, iga uurimisgrupp oma spetsiifilise uuringu alguses suuresti kordama ühtesid ning samu andmepuhastuse samme. Väikeriigis, kus mis tahes valdkonnas on spetsialiste vähe, ei saa sellist ressursikasutust pidada mõistlikuks. Võib arvata, et paljud teadusuuringud ongi jäänud Eestis tegemata kõrge tehnilise barjääri, raha ja kompetentside puuduse tõttu. Kogu maailmas ja seejuures ka Eestis on kindlasti veel palju arenguruumi ressursisäästlikumaks päriselu andmete kasutamiseks.

Eelnevalt toodud puudused ja piirangud on takistanud päriselu terviseandmete teisese kasutamise potentsiaali rakendamist, kuid õnneks on teadusmaailmal ka lahendusi. Järgnevalt on tutvustatud terviseandmete analüüsi vajadustele keskendunud ühtset andmemudelit ja selle rakendamise võimalusi. Sellise andmemudeli kasutuselevõtt toetab terviseandmete laiemat teisest kasutamist ning lahendab mitmed eelmainitud probleemid.

VAJADUS ÜHTSE ANDMEMUDELI JÄRELE

Päriselu terviseandmete uurijatele sai juba kümmekond aastat tagasi selgeks, et eri terviseandekogude andmed täiendavad üksteist ning nendes oleva info koondamine ühtsesse teadusandmebaasi on põhjalike terviseuuringute läbiviimiseks hädavajalik. Seega vajas lahendamist küsimus, milline on teadusliku andmebaasi parim struktuur. Näiteks loodi Eesti geenivaramu andmekogu rajamisel ka küsimustikupõhine terviklik andmemudel, kus enam kui tuhande välja peal sai struktuurselt andmeid koguda (8).

Kui tarkvaratehnoloogiate arenduskeskus (Tarkvara TAK, nüüd STACC) alustas terviseandmete analüüsi meetodite väljatöötamist, loodi esmane andmebaasi mudel ise. See arenes orgaaniliselt vastavalt sellele, mida rohkem sinna tervise infosüsteemi dokumentidest (sh tekstist) eraldatud fakte lisati. Õige pea jõuti tänu riigisisestele ja rahvusvahelistele ühisprojektidele (*European Medical Informatics Framework*, EMIF) järelduseni, et tark on suunduda teadusmaailmas vastselt levima hakanud andmemudelite poole. Tänasel päeval terviseandmete teisese kasutuse jaoks üks laialdasemalt kasutatavamaid terviseandmete mudeleid on andmeanalüüsi vajadustele keskendunud OMOP CDM (*Observational Medical Outcomes Partnership common data model*).

OMOP-ANDMEMUDEL

OMOP-andmemudel on suunatud päriselu terviseandmetel põhinevate uuringute läbiviimisele. Andmemudeli keskne eesmärk on võimaldada iga isiku kohta võimalikult paljude faktide ja sündmuste salvestamist koos sündmuse toimumise ajaga. Isikupõhine lähenemine on turvatud pseudonüümi kasutamisega, s.t oluline on ühendada fakte sama inimese kohta, kuid pole oluline, kes see isik on. Sündmuste toimumise aeg on vajalik erinevate perioodide ja sündmuste järjekorra salvestamiseks. Andmebaasi struktuur on loodud spetsiaalselt paljusid erinevaid teaduseesmärke silmas pidades. Võrreldes haiglate keerulisemate infosüsteemidega, pole tabeleid ega veerge väga palju – kokku üksnes 394 erinevat andmeväljа. Tabelite struktuur ja andmeväljad on hoolikalt läbi mõeldud, et enamikule tervisevaldkonna küsimustele saaks vastata võimalikult lihtsalt. OMOP-andmemudel koosneb rangelt kindlaksmääratud struktuurist (tabelid, veerud) ja suurest hulgast sõnastikest, mida andmete hoidmiseks on lubatud kasutada (vt joonis 1).

Fikseeritud sõnastikud võimaldavad tagada ühetaolise andmete analüüsitavuse. Sõnastikena kasutatakse maailmas juba olemasolevaid standardseid sõnastikke, näiteks diagnooside märkimiseks *SNOMED Clinical Terms* ja ravimite jaoks *RxNorm* (9, 10). Lisaks toetab mudel otse mitmeid teisi nimestikke nagu rahvusvaheline haiguste klassifikatsioon (RHK) ja toimeainete klassifikatsioon (ATC), pakkudes sisemisi automaatseid teisendusi *SNOMED*i ja *RxNorm*i peale. See võimaldab kasutada enamikku levinud sõnastikke andmete

sisestamisel ning ühtseid terminoloogiaid nende analüüsimisel. OMOP kasutamiseks peab algse andmebaasi valdaja teisendama info ise OMOP-kujule, sest üksnes andmebaasi omanik teab kõiki selle nüansse. Kuna kõik terviseandmekogud on unikaalsed, ei ole võimalik luua ühte universaalset teisendusprogrammi, mis sobib eri riikide erinevatele andmekogudele, küll aga saab pakkuda abivahendeid ja soovitusi teatud parimate praktikate jaoks. Andmeinseneerias on selle andmepuhastustöövoo etapi jaoks nimetus ETL (*Export, Transform, Load*), millega võimaldatakse operatiivbaasist võtta välja ja teisendada sobivale uuele kujule just vajalik andmekoosseis.

Andmebaasi teisendamine OMOP-kujule võib olla ka etapiviisiline ja ajas täienev, sest OMOP-mudelis on ainult mõned põhitabelid, mille täitmine on kohustuslik. On oluline, et oma OMOP-andmebaasi kuju saab luua erineva suurusega – kas haiglapõhiselt, raviarvete baasil või ka terve riigi andmeid koondades. Erinevus on vaid mastaabis ja detailsuses, millist infot vastavatest alusandmetest välja saab võtta. Juba kord loodud ETL-protsessi on võimalik rakendada erinevate uuringute jaoks, mis kasutavad samadest terviseandmekogudest pärit andmeid. Sellegipoolest tuleb arvestada, et nii nagu muutuvad ajas meditsiinipraktikad, kasutatavad koodistikud ja nimetused, tuleb ka ETL-protsessi pidevalt ajakohastada.

OMOP-ANDMEMUDELI VÕIMALUSED

Tõeline kasu ühtsest andmemudelist ilmneb peamiselt pärast andmete OMOP-kujule viimist. Nimelt on siis võimalik kasutada kümneid erinevaid vabavaralisi analüüsitööriistu ja tarkvarapakette, mis on loodud spetsiaalselt OMOP-kujul terviseandmetele. Tüüpiliselt on analüüsid kas keerulisi kohorte statistiliselt kirjeldavad ja võrdlevad (nn *population level estimation*) või masinõppe mudelitega sündmusi ennustavad (*patient level prediction*), nende analüüside kvaliteeti saab siis samuti OMOP-mudeli peal valideerida. Olemas on eraldi tarkvarapaketid lihtsamate analüüside tegemiseks (nt olukorra kirjeldus ja trendid) ja ka esimesed tööriistad keerukamate uuringute automaatseks läbiviimiseks (ravijuhendite täitmise hindamine, erinevate ravitrajektooride kirjeldamine, ennustusmudelite loomine, kulutõhususe analüüs jt). Oluline on märkida, et välja töötatud lähenemisi saab rakendada erinevates meditsiinivaldkondades – ühe haiguse või tingimuse jaoks välja töötatud analüüsitarkvara on suhteliselt lihtsalt kohandatav ka teistele haigustele.

Ühtne andmemudel võimaldab teadlastel ise välja töötada uusi analüüsimeetodeid ja tarkvarapakette, mida rakendada teiste osapoolte (sh ka välismaistel) OMOP-andmekogudel. OMOP-kujul terviseandmete kasutamine on avanud Eesti teadlastele

Andmed:

Tabel PERSON

PERSON ID	GENDER CONCEPT ID	YEAR OF BIRTH	MONTH OF BIRTH	BIRTH DATETIME
1	8532	1982	3	1982-03-01

Tabel CONDITION_OCCURRENCE

CONDITION OCCURRENCE ID	PERSON ID	CONDITION CONCEPT ID	CONDITION START DATE
1	1	194696	2010-01-06
2	1	432347	2014-10-05

Tabel DRUG_EXPOSURE

DRUG EXPOSURE ID	PERSON ID	GENDER CONCEPT ID	DRUG EXPOSURE START DATE	DRUG EXPOSURE END DATE
1	1	1127433	2010-01-06	2010-02-05

Sõnastikud:

Tabel CONCEPT

CONCEPT ID	CONCEPT NAME	DOMAIN	VOCABULARY ID
8532	Female	Gender	Gender
194696	Dysmenorrhea	Condition	SNOMED
432347	Chronic disease of tonsils AND/OR acenoids	Condition	SNOMED
1127433	acetaminophen 325 MG Oral Tablet	Drug	RxNorm

Joonis 1. „The Book of OHDSI" põhjal loodud lihtsustatud näide ühe patsiendi andmete esituskujust OMOP-andmemudelis.

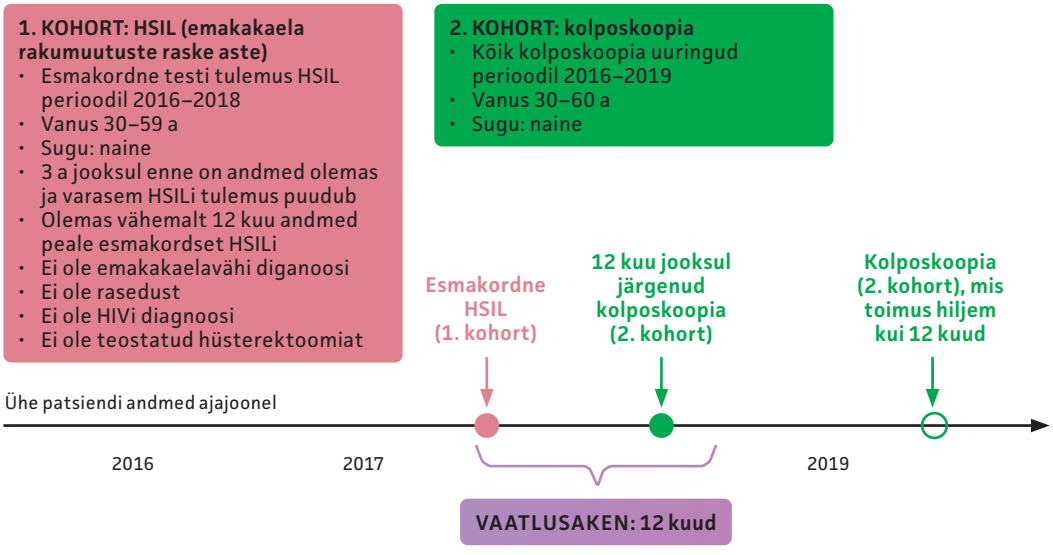
mitmeid uusi võimalusi rahvusvaheliseks teaduskoostööks. Näiteks Tartu Ülikooli terviseinformaatika töögrupp uurib ja arendab haigustrajektooride ning kulutõhususe analüüsimise meetodeid, mida on oma andmetel kasutanud ka Hollandi, Serbia, Hispaania ja Ameerika Ühendriikide partnerid (11, 12).

OMOP-põhised rahvusvahelised uuringud on üles ehitatud hajusvõrgu põhimõttel, kus omavahel vahetatakse ainult agregeeritud analüüsitulemusi ja ükski osapool ei väljasta teistele osapooltele üksikute patsientide andmeid. Universaalselt kasutatakse minimaalselt viie indiviidi nõuet (k-anonüümsus) iga raportisse lisatava arvu kohta. Seega on tagatud nii andmete turvalisus kui ka isikute privaatsuse kaitse. Standardsete analüüside jaoks (nt ravimiohutus, haiguste levimus) loodud vabavaraliste tarkvarapakettide abil tuleb konkreetse ravimi või haiguse uurimiseks paketis seadistada vaid ettenähtud parameetrid. See on viinud mitmete standardsete rahvusvaheliste uuringute korraldamise seninägematult kiireks, kus kõige aeglasemaks sammuks on uuringuprotokolli enda koostamine ning iga osaleva andmekogu ja riigi jaoks vajaliku eetikakomitee ning andmeväljastuse lubade menetlusprotsess. Üheks sellise standardse rahvusvahelise uuringu näiteks on Euroopa Ravimiameti koordineeritav võrgustik DARWIN (*Data Analysis and Real World Interrogation Network*), kus osaleb oma OMOP-andmestikuga ka Eesti geenivaramu ja kus viiakse aastas läbi *ca* 10 uuringut. Seda arvu on lähiaastatel plaanis mitmekordistada 50–60 uuringuni aastas (13, 14). Ilma OMOP-andmemudeli, selle peale kvaliteetselt ette valmistatud andmestike ja standardsete tarkvarapakettideta võtaks iga sellise uuringu läbiviimine aastaid.

Praegu arendab ja koordineerib OMOP-andmemudelit ja sellel põhinevaid analüüsimeetodeid avatud teaduskogukond Observational Health Data Sciences and Informatics (OHDSI, hääldatakse 'odüssei') (15). See kogukond on kasvanud maailmas tuhandete inimesteni ja OMOP-kujule on viidud umbes 12% maailma rahvastiku andmed (16, 17). OMOP publikatsioonide arv on olnud kasvutrendis alates 2016. aastast ning ülemaailmselt on avaldatud üle 500 teadusartikli (17, 18). Eesti teadlased on avaldanud artikleid astma (19), vähi (20), COVID-19 (21), reumatoidartriidi (22) ja sõeluuringute (23, 24) vallas.

OMOP-KOHORDID

Üks oluline uuendus, mida OMOP-andmemudel on terviseuuringute valdkonda toonud, on komplekssete kohortide loomise lihtsustamine ja nende korduvkasutamine. Kui tavapärastes uuringutes moodustab kohordi mingi kindel hulk patsiente, siis OMOP-andmestikel vaadeldakse kohorti pigem kui patsientide elus toimuvat selgelt defineeritud sündmust või nende kombinat-



Joonis 2. Kahe OMOP-kujule viidud kohordi lähtekirjeldus ja vaatlusakna rakendamise põhimõtteskeem emakakaela vähieelsete muutuste uuringus. Pilt on kohandatud Moosese jt artiklist (24).

siooni (25). Inimene võib sattuda kohorti ka mitu korda (näiteks kui uuritakse ülemiste hingamisteede külmetushaigusi, mis võivad korduda igal aastal) ja kuuluda mitmesse erinevasse kohorti. Näiteks kui uuritakse ravimite kõrvalmõjusid, võib iga kõrvalmõju moodustada eraldi kohordi. Seetõttu taandub suur hulk OMOP-andmemudeli uuringutest tegelikult kohortide omavahelise järjekorra või ülekatete uurimisele (näiteks ravimi X kõrvalmõju Y uurimiseks vaadatakse, kes kohordi X patsientidest sattus ettemääratud ajaakna jooksul ka kohorti Y). Kohortide definitsioonid ehk kaasavate ja välistavate tingimuste kirjeldamine on tehtud aga paindlikuks ja lihtsaks ning seda toetab graafiline kasutajaliides. Iga kohordi definitsioon salvestatakse masinloetaval kujul konfiguratsioonifailide või R-keele skriptidena ja täpselt samu definitsioone saab tulevikus kasutada teiste andmekogude juures või teistes uuringutes.

Joonisel 2 on toodud kahe kohordi kaasavate ja välistavate tingimuste kirjelduse näide. Neid definitsioone kasutati järgmise uurimisküsimuse vastamiseks: kui palju 30–59aastastest naistest, kelle Pap-testi tulemus on HSIL (*High-grade Squamous Intra-epithelial Lesion*), saavad kolposkoopia 12 kuu jooksul? Joonisel oleval ajateljel on kujutatud kohortide rakendamise põhimõtet. Iga kohorti kuuluva isiku kohta on teada huvipakkuva sündmuse toimumise kuupäev (1. kohort: HSILi kuupäev; 2. kohort: kolposkoopia kuupäev). Nende kohortide omavahelisel järjestamisel ja ülekatte uurimisel on võimalik vastata püstitatud uurimisküsimusele. Kui naine kuulus HSIL-kohorti, aga tema andmed puudusid kolposkoopiakohordis, siis järelikult temale kolposkoopia uuringut ei tehtud.

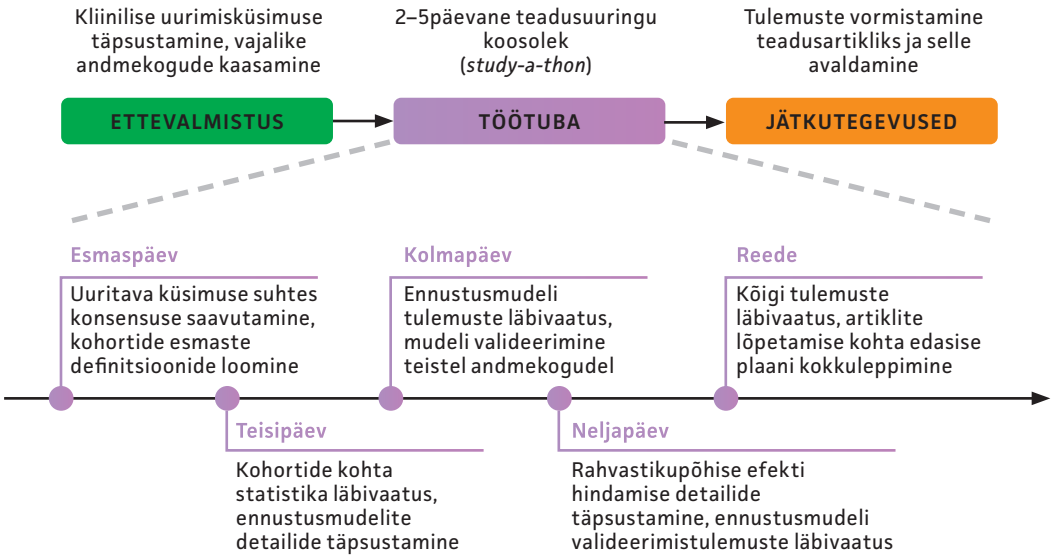
ANDMETEST ARTIKLINI TÖÖTUBADE VORMIS

Kohortide moodustamine on pärast andmete esmast puhastamist ja OMOP-kujule teisendamist tavaliselt üks uuringute aeganõudvamaid osi. Seni on kohortide moodustamine enamasti käinud nii, et uuringuprotokollis defineeritakse kohordid arstide poolt kaasavate ja välistavate tingimustena, mille järel andmebaasi spetsialist paneb kokku vastava päringu ja selgitab välja, kui palju igasse kohorti kuuluvaid isikuid on. Sõltuvalt andmebaasi keerukusest ja vajalike tunnuste iseärasustest võib kohordi defineerimine keerulises haiglainfosüsteemis võtta nädalaid, isegi kuid, ning igas erinevas andmestikus võib vajalik andmepäring olla täiesti erinev. Sõltuvalt saadud kohordi suurusest, soolisest ja vanuselisest jaotusest või muudest tingimustest võib järgneda ka kohordi definitsioonide ja päringute täpsustamine, mis pikendab kogu tsüklit oluliselt, muutes uuringute läbiviimise ilma standardse andmemudelita ajamahukaks.

Uuenduslik lähenemine, mis pärineb OHDSI kogukonnast ning kiirendab kohortide loomist ja seeläbi terve uuringu läbiviimist, on uuringute läbiviimine töötoa vormis (nn *study-a-thon*) (26). Töötubades osalevad nii vastava meditsiinilise valdkonna arstid, kelle ülesanne on muu hulgas defineerida õiged kohordid, epidemioloogid, meditsiinistatistikud, vastava valdkonna teadlased kui ka uuringus osalevate OMOP-andmekogude spetsialistid, kelle ülesandeks on OMOP-tööriistu kasutades kohe kontrollida, milline on loodud kohordi definitsiooni tulemus kasutatavas andmebaasis. Tänu kohesele tagasisidele saavad arstid viivitamata definitsioone täpsustada. Tulemusena saadakse kohortide definitsioonid paika väga kiiresti, enamasti päeva-paariga. Eriti kasulik on selline ühine defineerimine riikide võrdluses, kus sisuliselt sama info võib tulenevalt tervishoiusüsteemi toimimise eripäradest või kodeerimisega seotud põhjustest olla erinevalt kirjeldatud, kuid kohortide ühise defineerimise käigus saab kõiki neid eripärasid arvesse võtta.

Töötubades osalevad tavaliselt ka standardsete uuringupakettide spetsialistid, kes teavad, kuidas vajaliku uuringu korrektseks teostamiseks parameetreid kõige paremal viisil seadistada. Nii võib juba teisel või kolmandal töötoa päeval saada kätte esimesed uuringu tulemused ja kulutada ülejäänud päevad nende läbivaatamisele, vajadusel kohortide täpsustamisele, tulemuste tõlgendamisele ja kõige olulisemate sõnumite väljaselgitamisele. Edasi jätkub juba saadud tulemuste artikliks vormistamine, mida võib teha ka hilisema koostöö etapis vastavat kirjutamist eestvedava initsiatiivgrupi juhtimisel (vt joonis 3). Seega saab tänu OMOP kasutamisele tehnilised takistused minimeerida ja keskenduda uuringu tegelikule teaduslikule eesmärgile.

Nüüdseks on OHDSI kogukonnas toimunud juba mitmeid rahvusvahelisi töötube. Neist esimene uuris 2018. aastal



Joonis 3. Reumatoidartriiti uurinud töötoa ülesehitus. Pilt on kohandatud Hughesi artiklist (26).

põlveliigese endoproteesimist (27), teine 2020. aastal reumatoidartriidi raviskeeme (22). Mõlemad kestsid viis päeva (26), mille jooksul viibisid koos arstid, statistikud, tarkvara arendajad ja andmekogude spetsialistid. Töötubade väljundiks olid teadusartiklite esmased versioonid. Käesoleva artikli autoritel on plaan korraldada ka Eestis lähiajal esimene OHDSI töötuba, selleks ootame siinse artikli kohta ka esmast tagasisidet, ideid ja sooviavaldusi, milliseid uuringuid eelistada.

OHDSI JA OMOP EESTIS

Eestis on OMOP-lahendustega tegelenud peamiselt Tartu Ülikooli terviseinformaatika uurimisgrupp (28). Esimesed sammud tehti 2017. aastal, kui OMOP-kujule hakati viima geenivaramu doonorite küsimustike andmeid. 2018.–2019. aastal viidi OMOP-andmebaasi kujule ka esimesed tervise infosüsteemist ja Tervisekassa andmekogust pärit andmed. Need olid teadusliku töö käigus tehtud katsetused ja pärast teadustöö lõppu need OMOP-andmekogud kustutati, lähtudes eetikakomitee selle hetke loa ulatusest. Samas on uurimisgrupp rakendanud osalt samasid, kuid ka edasiarendatud ETL-protsessi teisendusreegleid järgnevate uuringute algandmetel (vt joonis 4). Kuigi uuringud ja andmekomplektid, mille raames andmeid on OMOP-kujule teisendatud, on erinevad, pärinevad andmed ikkagi algselt ühtedest ja samadest andmekogudest ehk tervise infosüsteemist, Tervisekassast ja registritest. See alusandmete stabiilsus on võimaldanud uurimisgrupil iga kord kasutada ja pidevalt täiendada üht ja sama automaatset andmepuhastuse töövoogu ehk ETL-protsessi, mis nüüdseks on kasvanud mitmesammuliseks ja erinevaid keerulisi andmeteaduse meetodeid kasutavaks tarkvaraks (6).

OMOP-andmestike põhjal on teadusgrupp iseseisvalt, koos Tartu Ülikooli meditsiinivaldkonna ja teiste Eesti teadlastega avaldanud mitmeid teadusartikleid (19–22, 29–32), samuti on tehtud hulgaliselt bakalaureuse- ja magistritöid (33). Lisaks meditsiiniliste küsimuste uurimisele oleme arendanud ka uusi analüüsimeetodeid haigustrajektooride ja ravijärgimuse uurimiseks ning visualiseerimiseks (6, 11, 12, 34, 35).

Tartu Ülikool pole aga sugugi ainus, kes OMOP-andmemudeliga Eestis tegeleb. Juba 2019. aastal läbisid vastava koolituse ja said OMOP-kujule viimiseks sertifikaadi kaks Eesti tarkvaraettevõtet – STACC ja Quretec (36). Praegu on Tartu Ülikooli Kliinikumis käimas DigiONE I3 projekt vähiandmete viimiseks OMOP-kujule.

Mitmete allüksuste ja uute asutuste ülene tegevus on tekitanud Eestis vajaduse siinseid OMOP-tegevusi paremini koordineerida ja kasvatada laiemat teadlikkust. 2023. aasta detsembris asutasid Eesti OMOP-huvilised riikliku OHDSI haru „OHDSI Estonia" (37). Selle eesmärkideks on muu hulgas vahendada oma liikmete ja rahvusvaheliste OHDSI harude vahel suhtlust, ühtlustada Eesti-sise-

Joonis 4. Praeguse seisuga teadusuuringute raames OMOP-kujule viidud või viidavad andmed Eestis.

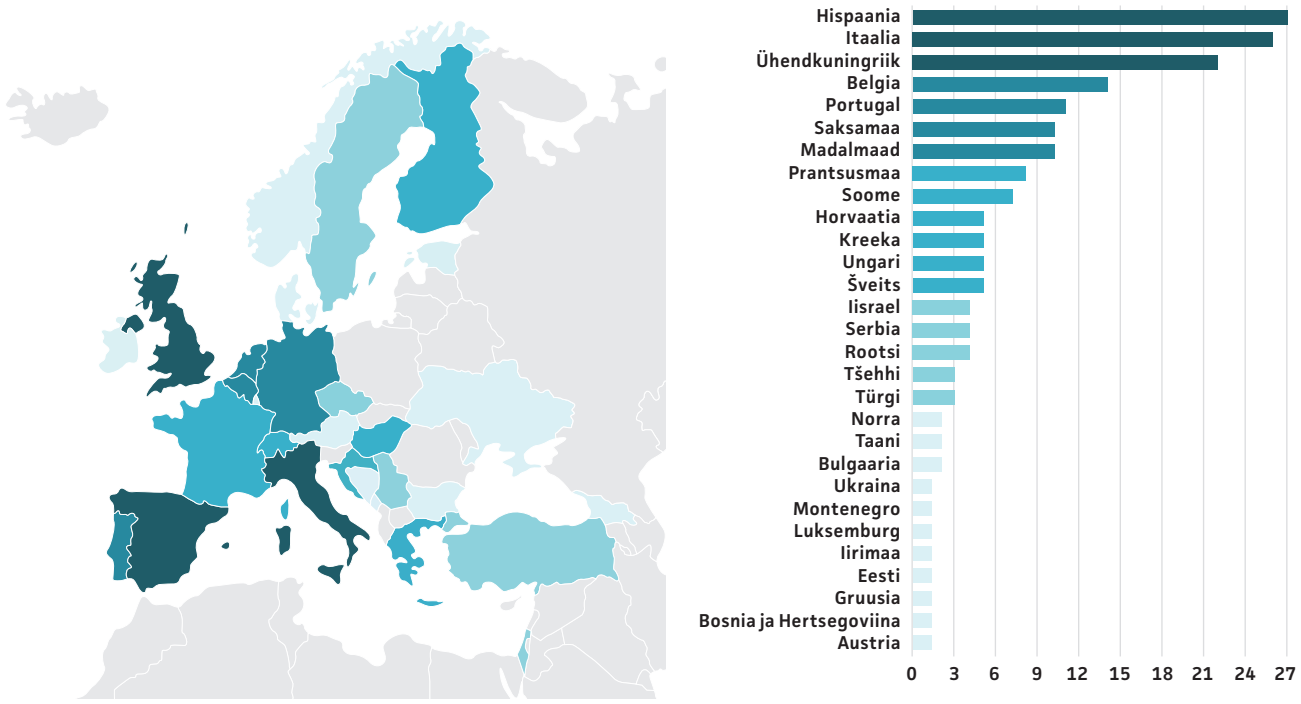
selt OHDSI ja OMOP tegevusi ja töövahendeid, teha OMOP-teemalist teavitustööd ning suurendada OMOP-andmemudeli kasutamist ja sellekohast kompetentsi Eestis. Võtmesõnaks OHDSI kogukondade puhul on vabatahtlikkus – igaühel, kes peab sellist lähenemist õigeks ja soovib panustada, on võimalus seda teha. Käsu korras kõiki andmestikke OMOP-kujule viia pole võimalik – kõigepealt peab tekkima andmekogude omanikel motivatsioon OMOP kasutamiseks. Kui see on olemas, siis tekivad juba uued võimalused ka koostööks ja uuringutes osalemiseks. OHDSI Estonia püüab seda igati toetada.

MIDA TOOB TULEVIK

Pole kahtlust, et järjest enam andmeomanikke hakkab oma andmeid OMOP-kujule viima ja regulaarselt uuendama (vt joonis 5). See avab palju uusi Eesti-siseseid ja rahvusvahelisi koostöövõimalusi ning võimaldab teha teadusuuringuid senisest oluliselt kiiremini. Tehniliste raskuste minimeerimine võiks motiveerida ka arste rohkem uuringute läbiviimises osalema.

Veidi keerulisem on olukord siis, kui algselt on samade patsientide info laiali eri andmekogudes ja kvaliteetse teaduse tegemiseks on tarvis andmed enne standardsele kujule viimist ühendada. Mis tahes kvaliteetne terviseuuring eeldab vähemalt diagnooside ja ravimite infot, mis praegu on kahes eraldi riiklikus andmekogus – tervise infosüsteemis ning retseptikeskuses. Andmete linkimine ja ühendamine ei ole tehniliselt keerulised tänu meie unikaalsele isikukoodile, mida saab edukalt ka pseudonüümida. Sel juhul jääb raskus teadlaste kanda, kes peavad esmalt taotlema andmeväljastust eri andmekogudest ning puhastama ja viima andmed OMOP-kujule oma teadusuuringu raames. Eespool kirjeldatud asjaoludel pole see aga mõistlik, sest protsess ise on tavaliselt suhteliselt sarnane. Kuigi Tartu Ülikooli terviseinformaatika töögrupp on loonud Eesti kesksetest terviseandmebaasidest pärit terviseandmete puhastamiseks korduvkasutatavaid andmepuhastuse töövoogusid, mis kiirendavad ja lihtsustavad andmepuhastuse protsessi, oleks otstarbekas luua Eestis spetsiaalselt erinevate teadusuuringute tarbeks OMOP-kujul andmekogu, mis on piisavalt esinduslik ning ühendab samade patsientide andmeid erinevatest andmekogudest. Andmekogusid võib olla ka mitu, kuid tuleb arvestada nende aluseks olevate patsientide gruppide ja andmete potentsiaalsete erinevustega.

Tartu Ülikooli terviseinformaatika uurimisrühm on saanud sel aastal eetikakomiteede load arendada tõhusate patsiendikesksete tervishoiu- ning ennetusteenuste



Joonis 5. EHDENi (The European Health Data & Evidence Network) projekti toel OMOP-kujule viidud andmestike arv riikide kaupa Euroopas. Pilt on kohandatud EHDENi veebilehelt (38).

osutamiseks ja kvaliteetse tõenduse pakkumiseks andmeteaduse meetodeid ning kasutada selleks Eesti päriselu terviseandmeid. Selleks koostatakse Eesti terviseandmete baasil representatiivne kureeritud kvaliteetne ja pseudonüümitud 30% juhuvalimi põhine OMOP-kujul terviseandmestik (EST-Health-30), mida uuendatakse regulaarselt. Uuringu raames luuakse tark- ja riistvarapõhine turvaline taristu, arendatakse ja valideeritakse ravijärgimuse, haigustrajektooride ja personaliseeritud ennetusmeetmete analüüsi metoodikaid. Tulevikus võiks täiendavate eetikakomitee lubade saamise korral olla võimalik kasutada sama andmestikku ka muudeks uuringuteks, kuid vastavad praktikad peavad alles tekkima.

Tänu kiiresti kasvavale uute moodustatavate OMOP-andmestike loomisele võib ennustada ka OMOP-põhiste teadusuuringute kiiret kasvu. Seda ei soodusta mitte üksnes efektiivsed kvaliteetsed tööriistad, vaid ka kõikide meetodite läbipaistvus ning igaühe võimalus arendamisse panustada. Kõik OHDSI tööriistad on vabalt kasutatavad ja avatud lähtekoodiga. Enamasti on uuringu tarkvara lähtekood koos masinloetavate kohortide kirjeldustega lisatud ka teadusartiklite juurde, mis võimaldab igaühel neid taaskasutada ja tulemusi oma andmetel kergemini reprodutseerida.

Barjäär tervisevaldkonna uuringute läbiviimiseks ja meditsiinilise tõenduse saamiseks on nüüd palju madalam kui varem. Loodame, et tänu kvaliteetsele tõendusele paraneb nii patsientide ravi kui tervishoiukorralduslikud otsused ning seeläbi meie kõigi tervelt elatud aastate arv.

TÄNUSÕNAD

Uuringut toetas Eesti Teadusagentuur (projektid RITA1/02-96, PRG1844) ja Euroopa Regionaalarengu Fond (projekt EU48684).

VÕIMALIKU HUVIKONFLIKTI DEKLARATSIOON

Sulev Reisberg on OHDSI Estonia juht ja Raivo Kolde asejuht. Jaak Vilo on Quretec osanik ja juhatuse liige.

[1] University of Tartu, Institute of Computer Science, Estonia,
[2] STACC, Estonia,
[3] University of Tartu, Institute of Biomedicine and Translational Medicine, Estonia,
[4] Tartu University Hospital, Cancer Centre, Estonia

Correspondence to:
Sulev Reisberg
sulev.reisberg@ut.ee

Keywords:
OMOP, OHDSI, real-world data, health study

SUMMARY

Novel approach - health studies based on the OMOP data model

Sulev Reisberg[1,2], Kerli Mooses[1], Raivo Kolde[1], Lenne-Triin Kõrgvee[3,4], Jaak Vilo[1,2]

Medicine and health policies are based on scientific evidence. In addition to prospective randomised clinical trials, real-world data produced during clinical practice are used to create such evidence. This requires

standard dictionaries and well-structured data. One such widely used standard is the OMOP (Observational Medical Outcomes Partnership) data model. Its structural and semantic homogeneity simplifies the implementation of standard analyses, making the conduction of studies fast and cheap. The use of OMOP databases in international cooperation ensures data protection, as only anonymised aggregated results are shared with partners. Today, the health data of 12% of the world's population has been transformed to the OMOP format. Using OMOP databases, Estonia has the opportunity to be at the forefront of developments in the usage of real-world data and to increase the opportunities for domestic and international research collaboration.

KIRJANDUS / REFERENCES

1. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992;268:2420–5.
2. Corrao G, Cantarutti A. Building reliable evidence from real-world data: Needs, methods, cautiousness and recommendations. Pulm Pharmacol Ther 2018;53:61–7.
3. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022;22:287.
4. Dang A. Real-World Evidence: A Primer. Pharmaceut Med 2023;37:25–36.
5. Kasekamp K, Habicht T, Võrk A jt. Eesti: Tervisesüsteemi ülevaade. Tervisesüsteemid muutustes. 2023;25: i–204.
6. Oja M, Tamm S, Mooses K, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) common data model: Lessons learned. JAMIA Open 2023;6:ooad100.
7. Dodd C, Andrews N, Petousis-Harris H, Sturkenboom M, Omer SB, Black S. Methodological frontiers in vaccine safety: qualifying available evidence for rare events, use of distributed data networks to monitor vaccine safety issues, and monitoring the safety of pregnancy interventions. BMJ Glob Health 2021;6(Suppl 2).
8. Kamm L, Krushevskaja D, Talvik HA, Veldemann J, Vilgota A, Vilo J. Flexible database platform for biomedical research with multiple user interfaces and a universal query engine. In: Databases and Information Systems V. IOS Press, 2009:301–10.
9. Reich C, Ostropolets A, Ryan P, et al. OHDSI Standardized Vocabularies-a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc 2024;31:583–90.
10. Karu K. SNOMED CT – elektroonilise haigusloo loomise vahend. Eesti Arst 2011;90:466–73.
11. Künnapuu K, Ioannou S, Ligi K, et al. 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;5:ooac021.
12. Haug M, Oja M, Pajusalu M, et al. Markov modeling for cost-effectiveness using federated health data network. J Am Med Inform Assoc 2024;31:1093–101.
13. Le Page M. Teadus. AS Postimees Grupp; 2024 [cited 2024 Mar 12]. Tartu Ülikool osaleb uurimisvõrgustikus DARWIN EU tõenduspõhise terviseinfo jagamisel. Available from: https://teadus.postimees.ee/7745414/tartu-ulikool-osaleb-uurimisvorgustikus-darwin-eu-toenduspohise-terviseinfo-jagamisel.
14. Szymański P, Weidinger F, Lordereau-Richard I, et al. Real world evidence: Perspectives from a European Society of Cardiology Cardiovascular Round Table with contribution from the European Medicines Agency. Eur Heart J Qual Care Clin Outcomes 2023;9:109–18.
15. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574–8.
16. Hripcsak G, Schuemie MJ, Madigan D, Ryan PB, Suchard MA. Drawing Reproducible Conclusions from Observational Clinical Data with OHDSI. Yearb Med Inform 2021;30:283–9.
17. OHDSI. OHDSI. 2023 [cited 2024 Mar 12]. Our Journey. Where the OHDSI Community Has Been and Where We Are Going. Available from: https://www.ohdsi.org/wp-content/uploads/2023/11/OHDSI-Book2023.pdf.
18. Reinecke I, Reich C, Sedlmayr M, Bathelt F. The Usage of OHDSI OMOP – A Scoping Review. In: German Medical Data Sciences 2021: Digital Medicine: Recognize – Understand – Heal. IOS Press, 2021:95–103.
19. Markus AF, Rijnbeek PR, Kors JA, et al. Real-world treatment trajectories of adults with newly diagnosed asthma or COPD. BMJ Open Respir Res 2024;11.
20. Gandaglia G, Pellegrino F, Golozar A, et al. Clinical Characterization of Patients Diagnosed with Prostate Cancer and Undergoing Conservative Management: A PIONEER analysis based on big data. Eur Urol 2024:5:457–65.
21. Mooses K, Vesilind K, Oja M, et al. The use of prescription drugs and health care services during the 6-month post-COVID-19 period. Sci Rep 2023;13:11638.
22. Yang C, Williams RD, Swerdel JN, et al. Development and external validation of prediction models for adverse health outcomes in rheumatoid arthritis: A multinational real-world cohort analysis. Semin Arthritis Rheum 2022;56:152050.
23. Uusküla A, Oja M, Tamm S, et al. Prevaccination Prevalence of Type-Specific Human Papillomavirus Infection by Grade of Cervical Cytology in Estonia. JAMA Netw Open. 2023 Feb 1;6(2):e2254075.
24. Mooses K, Savrova A, Pajusalu M, et al. Using electronic health records to evaluate the adherence to cervical cancer prevention guidelines: a cross-sectional study. Prev Med 2024;183:107982. Doi: 10.1016/j.ypmed.2024.107982. Epub 2024 May 1.
25. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI, 2019:458.
26. Hughes N, Rijnbeek PR, van Bochove K, et al. Evaluating a novel approach to stimulate open science collaborations: a case series of "study-a-thon" events within the OHDSI and European IMI communities. JAMIA Open 2022;5:ooac100.
27. Williams RD, Reps JM, OHDSI/EHDEN Knee Arthroplasty Group, Rijnbeek PR, Ryan PB, Prieto-Alhambra D. 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. Knee Surg Sports Traumatol Arthrosc 2022;30:3068–75.
28. Research Group of Health Informatics at the University of Tartu, Insitute of Computer Science, Tartu Ülikool; 2022 [cited 2024 mar 12]. Research group of Gealth Informatics at the University of Tartu. available from: https://health-informatics.cs.ut.ee/.
29. Uusküla A, Jürgenson T, Pisarev H, et al. Long-term mortality following SARS-CoV-2 infection: A national cohort study from Estonia. Lancet Reg Health Eur 2022;18:100394.
30. Meister T, Pisarev H, Kolde R, et al. Clinical characteristics and risk factors for COVID-19 infection and disease severity: A nationwide observational study in Estonia. PLoS One 2022;17:e0270192.
31. Rosenberg M, Thetloff M, Tamm S, Kuusk K, Reisberg S, Vilo J. Kroonilise neeruhaiguse levimus Eesti e-tervise andmete alusel. Eesti Arst 2023;102:263–76.
32. Voss EA, Shoaibi A, Yin Hui Lai L, et al. Contextualising adverse events of special interest to characterise the baseline incidence rates in 24 million patients with COVID-19 across 26 databases: a multinational retrospective cohort study. EClinicalMedicine 2023;58:101932. doi: 10.1016/j.eclinm.2023.101932. Epub 2023 Apr 4.
33. Research Group of Health Informatics at the University of Tartu. Thesis list of the research group of health informatics. University of Tartu, Institute of Computer Science, Tartu Ülikool; 2022 [cited 2024 Mar 12]. Thesis. Available from: https://health-informatics.cs.ut.ee/thesis/.
34. Pajusalu M, Mooses K, Oja M, Tamm S, Haug M, Kolde R. TrajectoryViz: Interactive visualization of treatment trajectories [Internet]. bioRxiv. 2024. Available from: https://www.medrxiv.org/content/10.1101/2024.04.01.24305168v1.abstract.
35. Talvik HA, Oja M, Tamm S, et al. Repeatable process for extracting health data from HL7 CDA documents [Internet]. SSRN. 2024. Available from: https://ssrn.com/abstract=4776237.
36. Hughes N. EHDEN. 2024 [cited 2024 Mar 12]. EHDEN certifies the next 6 SMEs. Available from: https://www.ehden.eu/ehden-certifies-the-next-6-smes/.
37. Reisberg S. OHDSI. [cited 2024 Mar 12]. OHDSI Estonia. Available from: https://www.ohdsi-europe.org/index.php/national-nodes/estonia.
38. EHDEN. ehden.eu. EHDEN; 2020 [cited 2024 Mar 27]. Data Partners. Available from: https://www.ehden.eu/datapartners/.