

UURINGU EETILISE HINDAMISE TAOTLUS EESTI BIOEETIKA JA INIMUURINGUTE NÕUKOGULE

1. Uuringu nimetus (inglisekeelsete taotluse puhul tuleb uuringu nimetus ära tuua ka eesti keeles)

EST-Health-30 - Eesti terviseandmete väärindamine

2. Uuringu põhieesmärk kuni 450 tähemärki (0,25 lk) (inglisekeelsete taotluse puhul tuleb uuringu põhieesmärk ära tuua ka eesti keeles)

Eesmärk on Eesti terviseandmete väärindamine läbi andmeteaduse meetodite, et pakkuda kvaliteetset tõendust tõhusate patsiendikesksete tervishoiu- ja ennetusteenuste osutamiseks.

Uuring viiakse läbi kahes etapis:

1. etapp on alusandmestiku loomine ja andmekvaliteedi tõstmise meetodite arendamine, mille käigus lingitakse tervise infosüsteemi, Tervisekassa andmekogu, retseptikeskuse, vähiregistri ja surma põhjuste registri andmed ja seeläbi koostatakse teadusuuringute läbiviimiseks alusandmestik (täpne andmekoosseis on toodud lisas 1). Alusandmestiku loomise oluliseks osaks on andmekvaliteeti ja andmehõivet parandavate meetodite sh tehisintellekti mudelite arendamine. Samuti on alusandmestiku loomine vajalik 2. etapi ellu viimiseks ja eesmärkide täitmiseks.

1. etapi eesmärgid on:

1.1 Hinnata valimi põhjal Eesti terviseandmete kvaliteeti ja hõivet võrreldes erinevatest andmeallikatest pärinevat informatsiooni.

1.2 Andmekvaliteedi ja andmehõive tõstmise meetodite sh tehisintellekti-, tekstikaeve- ja imputatsioonimeetodite arendamine.

1.3 Tervisesündmuste esituse ühtlustamine rahvusvahelistele standarditele üle aja ja andmeallikate, sh kasutades tekstikaeve, klasterdamise ja automaattõlke meetodeid.

1.4 Arendatavad meetodid on hiljem rakendatavad ka uutes uuringutes, suurematel andmehulkadel ning riiklikes infosüsteemides andmekvaliteedi ja -hõive parandamiseks.

1.5 Standardiseeritud ja ühtlustatud andmestiku loomine ja kirjeldamine teadusuuringute läbiviimiseks. Andmestikku kirjeldatakse läbi erinevate numbriliste näitajate ja teostatavusanalüüside senisest suuremal detailsusastmel.

2. etapp on teadus- ja rakendusuuringute teostamine uurimismetoodikate arendamiseks Eesti terviseandmetel. Lähtume põhimõttest, et loodavad analüüsimeetodid on rakendatavad sõltumata uuringu all olevast haigusest või patsiendigrupist. Tulenevalt käimasolevatest teadus- ja rakendusuuringu projektidest (vt. pt "5. Uuringu finantseerimine") keskendutakse taotletava teadusuuringu raames ravijärgimuse, ravitrajektooride ja -teekondade ning personaliseeritud ennetusmeetoditega seotud analüüsimeetodite arendamisele. Täiendavate uurimisküsimuste tekkimisel esitatakse Eesti bioetika ja inimuuringute nõukogule ja EST-Health-30 uuringu andmeandjatele hindamiseks uus taotlus.

2. etapi eesmärgid on:

2.1 Arendada haigustrajektooride ja raviteekondade analüüsi metoodikaid kasutades sh tehisintellekti meetodeid, et kirjeldada praeguseid ravipraktikaid, võrrelda neid raviteekondade ja -juhenditega ning modelleerida tervisetulemeid ja ravi majanduslikku mõju.

2.2 Analüüsida erinevate faktorite (nt sugu, vanus, erinevad haigused) mõju ravijärgimusele, hinnata ravijärgimuse mõju erinevatele tervisetulemitele ja luua personaalseid ennustusmudeleid, mis võimaldavad maandada ravi mittejärgimise riske.

2.3 Arendada personaliseeritud ennetusmudeleid, mis võimaldavad vaatlusandmete põhjal tuvastada praeguste haiguse ennetuste ja ravipraktikate kitsaskohti, planeerida senisest paremini haiguste ennetustegevusi ja hinnata pakutud ennetusteenuste majanduslikku mõju

Taotletav teadusuuring loob eeldused, et päriselu terviseandmeid saaks Eestis laiemalt kasutada ning aitab parandada tervishoiuteenuste kvaliteeti, tervishoiusüsteemi efektiivsust ning toetada tervishoiupoliitiliste tõenduspõhiste otsuste tegemist, toetades seeläbi rahvastiku tervise arengukava 2020-2030 elluviimist.

3. Vastutava(d) uurija(d) ning tema (nende) kontaktandmed

Vastutavad uurijad on samad teadusuuringu 1. ja 2. etapis

Eesnimi: Jaak

Perekonnanimi: Vilo

Ametikoht: andmeteaduse õppetooli juhataja, bioinformaatika professor

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

Telefon: +372 504 9365

e-post: vilo@ut.ee

Eesnimi: Raivo

Perekonnanimi: Kolde

Ametikoht: terviseinformaatika kaasprofessor

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

Telefon: +372 506 7961

e-post: raivo.kolde@ut.ee

Eesnimi: Sven

Perekonnanimi: Laur

Ametikoht: turvalise andmekaeve kaasprofessor

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

Telefon: +372 5309 0584

e-post: sven.laur@ut.ee

Eesnimi: Sulev

Perekonnanimi: Reisberg

Ametikoht: terviseinformaatika teadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

Telefon: +372 524 8123

e-post: sulev.reisberg@ut.ee

4. Uuringu läbiviijad (lisada juurde vajalik arv ridu)

1. ja 2. etapi tegevuste elluviimisega seotud uuringu läbiviijad:

1. Eesnimi: Marek

Perekonnanimi: Oja

Ametikoht: terviseinformaatika teadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

2. Eesnimi: Markus

Perekonnanimi: Haug

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

3. Eesnimi: Harry-Anton

Perekonnanimi: Talvik

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

4. Eesnimi: Hendrik

Perekonnanimi: Šuvalov

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

5. Eesnimi: Maria

Perekonnanimi: Malk

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

6. Eesnimi: Ami

Perekonnanimi: Sild

Ametikoht: terviseandmete insener-analüütik

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

7. Eesnimi: Sirli

Perekonnanimi: Tamm

Ametikoht: programmeerija

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

8. Eesnimi: Kaire

Perekonnanimi: Koljal

Ametikoht: programmeerija

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

9. Eesnimi: Sander

Perekonnanimi: Kütisaar

Ametikoht: süsteemiadministraator

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut, teadusarvutuste keskus

2. etapis lisanduvad veel järgmised teadusuuringu läbiviijad

10. Eesnimi: Kerli

Perekonnanimi: Mooses

Ametikoht: terviseinformaatika teadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

11. Eesnimi: Taavi

Perekonnanimi: Tillmann

Ametikoht: rahvatervishoiu kaasprofessor

Organisatsioon: Tartu Ülikool, peremeditsiini ja rahvatervishoiu instituut

12. Eesnimi: Kunnar

Perekonnanimi: Kukk

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

13. Eesnimi: Õie Renata

Perekonnanimi: Siimon

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

14. Eesnimi: Maarja

Perekonnanimi: Pajusalu

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

15. Eesnimi: Anton

Perekonnanimi: Vykhovanets

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

16. Eesnimi: Kermo

Perekonnanimi: Saarse

Ametikoht: terviseinformaatika nooremteadur

Organisatsioon: Tartu Ülikool, arvutiteaduse instituut

<p>17. Eesnimi: Helene Perekonnanimi: Loorents Ametikoht: programmeerija Organisatsioon: Tartu Ülikool, arvutiteaduse instituut</p> <p>18. Eesnimi: Nikita Perekonnanimi: Umov Ametikoht: nooremteadur Organisatsioon: Tartu Ülikool, peremeditsiini ja rahvatervishoiu instituut</p> <p>19. Eesnimi: Laura Perekonnanimi: Lõo Ametikoht: nooremteadur Organisatsioon: Tartu Ülikool, peremeditsiini ja rahvatervishoiu instituut</p> <p>20. Eesnimi: Neeme Perekonnanimi: Ilves Ametikoht: spetsialist Organisatsioon: Tartu Ülikool, arvutiteaduse instituut</p>	
5. Uuringu finantseerimine	
Finantseerimise allikad	<ul style="list-style-type: none"> SA Eesti Teadusagentuuri rahastatud Personaalse uurimistoetuse rühmagrant "Kliiniliselt oluliste radade tuvastamine terviseandmetest" (PRG1844, jaan 2023 – dets 2027). Vastutav täitja: Jaak Vilo. Eelarve: 1 350 000 EUR. SA Eesti Teadusagentuuri rahastatud grant SLTAT24173 (TK213U6) "Eesti Tehisintellekti Tippkeskus" (EXAI) (terviseinformaatika alamprojekt) (01.01.2024–31.12.2030); Vastutav täitja: Jaak Vilo; Eelarve: 525 222 EUR. SA Eesti Teadusagentuuri rahastatud temaatilise teadus- ja arendusprogrammi projekt "Terviseandmete teisese kasutamise võimekuse kasvatamine" (TEM-TA72, jaan 2024 - dets 2028). Vastutav täitja: Raivo Kolde. Eelarve: 1 350 000 EUR. Teadusarvutuste ja andmete säilitamiseks kasutatakse taristut Eesti Teadusarvutuste Infrastruktuur (ETAIS), mida toetab Haridus- ja Teadusministeerium otse Eesti Teadusarvutuste Infrastruktuuri ja Eesti Hariduse ja Teaduse Andmesidevõrgu kaudu¹. ETAIS tegevustoetuse maht on 305 000 EUR aastas. <p>Viited:</p> <ol style="list-style-type: none"> Riikliku tähtsusega teadustaristu toetamine (07.10.2023), <i>Riigi Teataja I</i>. Kasutatud 19.08.2024, https://www.riigiteataja.ee/akt/104102023006

Uuringu üldmaksumus (summa)	Alusandmestik on vajalik projektide PRG1844, TK213U6 ja TEM-TA72 eesmärkide saavutamiseks. Eesseeisval perioodil (3a) on nende projektide tegevused planeeritud teadustööks loodaval andmestikul. Seepärast hindame projekti kogumaksumuseks ~1,8 mln EUR.
Uuritavale kompensatsiooni maksmine (jah, ei, põhjendus ja summa)	Ei, loodav andmestik ja terviseuuringud põhinevad päriselu andmete teisesel kasutusel, uuritavatega ühendust ei võeta.
Uuritavate kindlustus (jah, ei, kindlustaja ja poliis)	Ei
6. Uuringu läbiviimise aeg (algus ja lõpp kuu ja aasta täpsusega)	
Uuring viiakse läbi oktoober 2024 - detsember 2027. Lisaks arhiveeritakse andmed üheks aastaks peale uuringu lõppu, et vajadusel võimaldada arhiveerimisperioodil retsenseerimisel olevate teadusartiklite analüüsi täiendamine. Andmed hävitatakse hiljemalt 31. detsembriks 2028.	
7. Teave sama uuringu projekti varasema või samaaegse hindamise kohta (sh teistes riikides)	
EST-Health-30 projekti on varasemalt menetletud, kuid käesolevat uuringutaotlust on täpsustatud eesmärgi, andmetööstustoimingute, uuringu perioodi jt aspektides lähtudes andmeandjatega toimunud aruteludest.	
8. Lühülevaade siiani samal teemal tehtud uuringutest (kuni 900 tähemärki, 0,5 lk)	
<p>Avatud teaduskoostöö organisatsiooni OHDSI loodud ja edasiarendatavate OMOP (<i>Observational Medical Outcomes Partnership</i>) andmemudelil põhinevate rahvusvaheliste terviseuuringute arv on viimastel aastatel hüppeliselt kasvanud. Neid on avaldatud väga erinevatel teemadel - näiteks on uuritud nii rasvumist, põlveliigese vahetust, reumatoidartriiti kui ka koroonapatsiente^{1,2,3,4}. Ka käesoleva uuringu meeskonnal on pikaaegne kogemus sarnaste OMOP andmemudelil põhinevate terviseuuringute läbiviimisel. Näiteks EHDEN projektis osaleb meeskond aktiivse partnerina OMOP mudeli ja analüüsimetodite arendamisel ning nende populariseerijana Euroopas^{5,6}. Projektis "Masinõppe ja AI toega teenused" lõime käesoleva taotluse raames loodava andmestikuga sarnase, kuid väiksemas mahus (valimi suurus 10%, ei sisaldanud andmeid vähiregistrist, surma põhjuste registrist) kvaliteetse andmestiku 2012.-2019. aasta terviseandmetel^{7,8,9,10}. Lisaks varasemalt loodud andmestikku ei uuendatud. Sellest alates oleme taaskasutanud ja edasi arendanud Eesti terviseandmete jaoks spetsiifilisi automaatseid andmepuhastuse tehnikaid, mis on võimaldanud meil endal läbi viia uuringuid ja ka osaleda mitmesugustes rahvusvahelistes projektides. Heaks näiteks, on uuring, kus hindasime inimese papilloomiviiruse (HPV) tüüpide levimust Eestis tuginedes tekstikaevega tervisedokumentidest eraldatud HPV testi tulemustele⁸. Meie senise töö ja Eesti terviseandmete kvaliteeti näitab ka see, et DARWIN projektis valis Euroopa Raviamet meid üheks kvaliteetseks andmepartneriks, mille tulemusel kasutatakse Eesti OMOP kujule viidud Eesti Geenivaramu terviseandmestikku ühena teiste seas ravimite reguleerimisega seotud küsimuste uurimiseks¹¹. OPTIMA projektis analüüsimise Eesti terviseandmeid organkasvajatega patsientidele optimaalse ravi leidmiseks. Seniste uuringute tulemusi on avaldatud nii Eesti kui rahvusvahelistes teadusajakirjades, postritel kui ettekannetel, samuti uurimisgrupi töös osalenud üliõpilaste lõputöodes.</p> <p>Viited:</p> <ol style="list-style-type: none"> https://doi.org/10.1101%2F2020.04.22.20074336 https://doi.org/10.1038/s41366-021-00893-4 https://doi.org/10.1136/bmj.n1038 https://doi.org/10.1101/2020.12.14.20240994 http://dx.doi.org/10.1093/jamiaopen/ooac021 https://doi.org/10.1016/j.semarthrit.2022.152050 https://www.etag.ee/wp-content/uploads/2022/05/RITA_MAITT_LOPPARUANNE_FINAL.pdf 	

8. <http://dx.doi.org/10.1001/jamanetworkopen.2022.54075>
9. <http://dx.doi.org/10.1038/s41598-023-38691-9>
10. <https://eestiartst.ee/kroonilise-neeruhaiguse-levimus-eesti-e-tervise-andmete-alusel/>
11. <https://cs.ut.ee/et/sisu/tartu-ulikool-osaleb-andmepartnerina-andmeanaluusi-ja-tavaandmete-uurimisvorgustikus-darwin>

9. Planeeritava uuringu põhjendus ning uurimisküsimused ja/või hüpoteesid (kuni 1800 tähemärki, 1 lk)

Meditisiinis ja rahvatervises on otsuste tegemiseks vaja laiapõhjalistel teadusuuringutel põhinevat tõendust. Uuringute tarvis on reeglina vaja esinduslikke andmeid, mis on võimalikult värsked ja samal ajal piisavalt pika aegreaga. Näiteks on värskemad andmed (sh imikute ja vastsündinute kohta) vajalikud vastamaks küsimustele, mis puudutavad tervise teenuse osutamise hetkeseisu, uute ravimeetodite kasutust ja efektiivsust ning tervisesüsteemi vastust erinevatele hiljutistele sündmustele. Usaldusväärsemate ennustusmodelite loomiseks, haiguse arengu, ravi efektiivsuse, erinevate ravitrajektooride jms hindamiseks on vajalik vaadelda pikemaajajärgseid ajaperioode.

Ajakohastel andmetel tugineva tõenduse leidmiseks kasutatakse järjest enam päriselu terviseandmete teisesel kasutusel põhinevaid uuringuid. Tänu Eesti unikaalsele tervishoiuandmete korraldusele on võimalik ühendada patsiendi tervist puudutavad olulised komponendid haiguste (Tervise infosüsteem, vähiregister), ravimite (retseptikeskus), tarbitud teenuste ja kindlustuskaitse (Tervisekassa andmekogu) ning surma põhjuste kohta. Nende andmetel OMOP andmemudelil põhinevate terviseuuringute läbiviimisel on Eesti tervishoiuvaldkonnale ja ühiskonnale laiemalt mitmekordne kasu. Nii aitab igapäevase tervishoiuvaldkonna toimimise raames salvestatud andmete taaskasutamine ühelt poolt hoida kokku andmete kogumisele kuluvat aega ja raha ning teisalt annab kõige täpsema ülevaate tegelikest protsessidest ja trajektooridest erineva tervise seisundi, soo ja vanusega isikutel. Oluline on siinjuures, et terviseandmete teisese kasutamise korral ei suurene patsientide uuringu koormus kuna kasutatakse varasemalt kogutud retrospektiivseid andmeid. Terviseandmete teisene kasutamine võimaldab ilma lisakoormust põhjustamata analüüsida ja osutada efektiivsemalt tervishoiuteenuseid patsiendi gruppidel, kellel traditsiooniliste teadusuuringute tegemine ei ole eetilistel kaalutlustel võimalik (näiteks ravimite mõju lastele või rasedatele¹). Kaasates analüüsi infot patsiendi kogu raviajaloo kohta pikema aja vältel, mitte ainult uuritava haigusega seotud tegevuste kohta, võib-olla võtmetähtsusega oluliste seoste avastamisel. Lisaks on taotletava teadusuuringu käigus 30% valimi peal välja töötatud meetodid andmekvaliteedi parandamiseks rakendatavad ka teistes teadusuuringutes ning riiklikes andmebaasides kogu rahvastiku andmetel. Taotletav teadusuuring loob eeldused, et päriselu terviseandmeid saaks Eestis laiemalt kasutada ning aitab parandada tervishoiuteenuste kvaliteeti, tervishoiusüsteemi efektiivsust ning toetada tervishoiupoliitiliste tõenduspõhiste otsuste tegemist, toetades seeläbi rahvastiku tervise arengukava 2020-2030 elluviimist. Kvaliteetsem tervishoiusüsteem omakorda toetab tervelt elatud eluaastate suurenemist.

Taotletav teadusuuring koosneb kahest etapist:

1. etapp: alusandmestiku loomine ja andmekvaliteedi tõstmise meetodite arendamine

Alusandmestiku loomisel pannakse kokku ja ühtlustatakse tervise infosüsteemist, Tervisekassa andmekogust, retseptikeskusest, vähiregistrist ja surma põhjuste registrist pärit andmed. Erinevad andmebaasid kajastavad erinevaid aspekte sama patsiendi raviteekonnast. Seetõttu on võimalik saada tervikpilt raviprotsessist vaid kõikide vajalike andmete ühendamise teel. Ülekatted andmeallikate vahel võimaldavad ka süstemaatiliselt hinnata algandmete kvaliteeti ja -hõivet, mis paljuski sõltuvad nende kogumise eesmärgist ning tehnilisest võimekusest ja digitaliseerituse astmest andmete kogumise hetkel. Näiteks saame alusandmestiku loomisel hinnata kui paljude patsientide kohta on olemas vajalik informatsioon, kui suur osa tehtud mõõtmistest on andmetes kajastatud, kas ja kuidas on puuduvad mõõtmised kallutatud või kui usaldusväärased on diagnoosid. Kuna valim on piisavalt suur ja esinduslik, siis saadud tulemused üldistuvad kogu Eesti terviseandmetele.

Tuginedes oma varasemale teadustööle analoogsete OMOP andmebaaside loomisel teame, et on mitmeid viise kuidas andmehõivet ja -kvaliteeti läbi erinevate andmepuhastamise protseduuride tõsta. Näiteks enne andmete analüüsimist teostame täiendava anonümiseerimise vabatekstilistes andmeväljades, et eemaldada võimalikud nimed, telefoninumbriid, isikukoodid, aadressid jms. Sellega vähendame isikute tuvastamise riski. Teadusuuringu käigus on plaanis juba olemasolevat anonümiseerijat veelgi tõhustada. Lisaks on üheks väljakutseks tervise infosüsteemi andmete kasutamisel vabatekstilistes andmeväljades oleva olulise terviseinfo eraldamine, struktureerimine ja standardimine. Vabatekstilistes väljades andmeanalüüsiks oluline informatsioon võib esineda lihtsamini eraldataval kujul, nagu näiteks vererõhu, kehakaalu ja pikkuse mõõtmised, kui ka keerukamal kujul, nagu patsiendi kaebused või harjumuste kirjeldused, kus on oluline mõista ka teksti sisu. Selleks, et veelgi rohkem ja efektiivsemalt leida üles erinevate uurimisküsimustega seotud infot vaba tekstist on oluline arendada vastavaid meetodeid, sh tehisisintellekti ja tekstikaeeve meetodeid, keerukamate olukordade jaoks, mis vajaliku informatsiooni eraldaks ning viiks selle struktureeritud ja standardiseeritud kujule. Töötades järjepidevalt samal andmestikul, on võimalik erinevate uurimisküsimuste lahendamiseks andmekvaliteeti märgatavalt parandada. Töö käigus välja töötatud meetodid loovad eeldused järgnevatel uuringutel paremaks läbiviimiseks ning on

hiljem rakendatavad ka tervise infosüsteemis olevate andmete parandamiseks. Loodud metoodikate rakendamine toetab ka Eesti teadus- ja arendustegevuse, innovatsiooni ning ettevõtluse (TAIE) arengukava 2021-2035, kus üheks juhtmõtteks on liikumine teksti- ja dokumendipõhiselt terviseandme mudelist üle andmepõhisele mudelile. Taotletavas teadusuuringus loodavaid andmepuhastustehnikaid on võimalik kasutada ka uue põlvkonna tervise infosüsteemi loomisel ja arendamisel.

Andmete esitus algandmeallikates on läbi aja muutunud ja arenenud. Näiteks Tervisekassa tervishoiuteenuste loetelu uuendatakse iga kolme kuu järel ning pikema aegrea puhul ei pruugi erinevatest laboritest pärinevad mõõtmised olla samamoodi kodeeritud. Et uuringud saaks läbi viia võimalikult pikas ajaaknas, tuleb andmete kodeering viia ühtsele standardile. Selle protsessi käigus tuleb luua üleminekutabelid, kus algallikates esinevad koodid on esitatud standardiselt. Nende loomisel kombineeritakse tekstikaeve, klasterdamise ja automaattõlke vahendeid manuaalse inspeksiooni ning kliinilise ekspertiisiga. Tulemuseks saadud üleminekutabelid võimaldavad automatiseerida andmete standardkujule viimist nii antud andmestiku uuendamisel kui ka täiesti uutes uuringutes. Samas on andmehõive ja -kvaliteedi parandamine iteratiivne protsess, kuna erinevate haigustega seotud uuringutes osalemisel tulevad ilmsiks uued kitsaskohad. Seetõttu on andmekvaliteedi tõstmise juures oluline andmeid kasutada võimalikult paljudes ja erinevates uuringutes.

Linkimise, puhastamise ning eelnevalt toodud lähenemiste rakendamisel viime erinevatest allikatest pärit andmed ühtsele OMOP andmemudelile, kus iga kirje on rahvusvaheliselt defineeritud standardkujul. Andmete OMOP kujule viimine teeb võimalikuks nende analüüsi juba suure hulga olemasolevate analüüsitööriistadega, mis muu hulgas võimaldavad keerukate kohortide loomist ning jagamist, patsiendigruppide kirjeldamist, ennustusmodelite treenimist ja palju muud. Nii saab uuringuid läbi viia ka hajutatult ehk täiendavate uurimisküsimuste lahendamisel (millele küsitakse eraldi EBINi ja andmeandjate nõusolek) ei anta uutele isikutele ligipääsu andmetele vaid algne uuringumeeskond jooksub läbi uuringu koodi ning edastab vaid agregeeritud tulemused. 2. etapi uuringu läbiviijad saavad andmetele ligipääsu vajaduspõhiselt jälgides minimaalsusprintsipi läbi Eesti Teadusarvutuste infrastruktuuri hallatava turvalise töötlemiskeskonna, kus kõikide kasutajate kõik tegevused logitakse.

Eesti riiklike terviseandmebaaside omanikud on teinud avalikult kättesaadavaks erinevat riikliku tervisestatistika, mis annab mõningase ülevaate valitud tervisesündmuste toimumistest ja tervishoiuteenuste tarbimisest. Näitek TAI tervisestatistika ja terviseuuringute andmebaasis (statistika.tai.ee) on esmashaigestumuse info valitud RHK-10 peatükkide ja alampeatükkide kohta avaldatud kuni 2016 aastani ning hilisema perioodi avalik statistika puudub. Uuringute paremaks planeerimiseks ja teostatavuse hindamiseks on vaja detailsemat infot, kus on kombineeritud mitme andmepunkti ja andmeallikat ning võetud arvesse sündmuste toimumise omavahelisi ajalisi seoseid. Näiteks südameveresoontekonna haiguste ennetamiseks on oluline teada üle 40 aasta vanuste patsientide hulka, kellel on kolesterooli tase üle normi, kuid statiinipõhiseid ravimeid pole välja kirjutatud. Selleks on vaja andmeid labori analüüsides (Tervise Infosüsteem) ja retsepikeskusest (Tervisekassa). Adresseerimaks seda puudujääki, planeerime põhjalikult kirjeldada loodud andmestikku läbi erinevate numbriliste näitajate ja teostatavusanalüüsi detailsusastmel säilitades k ≥ 5 anonüümsus. Selline ülevaade esinduslikust populatsioonipõhisest andmekogumist on väärtuslikuks vahendiks tervishoiupoliitika kujundajatele ja -teenuse osutajatele ning uute uuringute planeerimisel ja teostatavuse hindamisel.

1. etapi eesmärgid kokkuvõtvalt on:

- 1.1. Hinnata valimi põhjal Eesti terviseandmete kvaliteeti ja hõivet võrreldes erinevatest andmeallikatest pärinevat informatsiooni.
- 1.2. Andmekvaliteedi ja andmehõive tõstmise meetodite sh tehisintellekti-, tekstikaeve- ja imputatsioonimeetodite arendamine.
- 1.3. Tervisesündmuste esituse ühtlustamine rahvusvahelistele standarditele üle aja ja andmeallikate, sh kasutades tekstikaeve, klasterdamise ja automaattõlke meetodeid
- 1.4. Arendatavad meetodid on hiljem rakendatavad ka uutes uuringutes, suurematel andmehulkadel ning riiklikes infosüsteemides andmekvaliteedi ja -hõive parandamiseks.
- 1.5. Standardiseeritud ja ühtlustatud andmestiku loomine ja kirjeldamine teadusuuringute läbiviimiseks. Andmestikku kirjeldatakse läbi erinevate numbriliste näitajate ja teostatavusanalüüsides senisest suuremal detailsusastmel.

2. etapp: teadus- ja rakendusuuringu teostamine uurimistetoodikate arendamiseks Eesti terviseandmetel

Enamasti kasutatakse teadusuuringutes *ad hoc* meetodikat, mis on optimeeritud konkreetsetele andmetele ja uurimisküsimustele ning mis ei üldistu uutele küsimustele ega ole kasutatavad sama uuringu kordamiseks teisel andmestikul. Selline lähenemine, kus samasid protseduure leiutatakse iga uuringu grupi poolt uuesti, on aega jt ressursse raiskav ning tingitud andmete vähesest standardiseeringust. Tänu OMOP andmemudeli laiemale levikule on võimalik luua terviseandmetele analüüsitööriistu, mis töötavad universaalselt sõltumata uuringu all olevast haigusest või patsiendi grupist. Juba OMOP platvormile arendatud arvutuslik taristu demonstreerib, et selline lähenemine kiirendab analüüsiprotsessi ning muudab seda läbipaistvamaks ja kvaliteetsemaks. Siiski on palju uurimisküsimusi, kus olemasolevast meetodikast ei piisa. Näiteks puudub nii ravitrajektooride kui ka -teekondade kirjeldamise ja modelleerimise jaoks üldistatud meetodika. Samas on Tervisekassa üheks eesmärgiks kiirendada ja laiendada raviteekondade kasutuselevõttu Eesti tervishoiusüsteemis ja seeläbi muuta pakutavaid tervishoiuteenuseid inimkeskemaks ja integreeritumaks ning toetada efektiivset ressursikasutust². Üldistatud meetodika puudumine tähendab, et iga uue raviteekonna kirjeldamisega tegelev töörühm peab töötama välja olukorra kirjeldamise meetodika. Esinduslikul andmebaasil rakendatava standardse meetodika välja töötamine kiirendab ja optimeerib raviteekondade valmimise protsessi ning toetab Tervisekassa eesmärkide täitmist.

Taotletava teadusuuringu 2. etapi üks eesmärk on arendada standardseid analüüsimeetodeid, mis on rakendatavad sõltumata uuringu all olevast haigusest või patsiendi grupist. Meetodite arendusel on oluline, et loodud meetodid on efektiivsed ja rakendatavad erinevate kliiniliste küsimuste puhul. Selleks testime erinevaid analüüsimeetodeid nii laialt üle erinevate haiguste ja patsiendigruppide kui ka läheme sügavuti konkreetsetesse kliinilistesse domeenidesse. Kõikide uuringute puhul kaasame ka kliinilised eksperdid, kes aitavad tulemusi konteksti panna ning tagavad uuringutulemuste rakendatavuse. Tulenevalt käimasolevatest teadusprojektidest (vt pt 5 Uuringute rahastamine) keskendutakse taotletava teadusuuringu raames ravijärgimuse, ravitrajektooride ja -teekondade ning personaliseeritud ennetusmeetoditega seotud analüüsimeetodite arendamisele.

2. etapi eesmärgid kokkuvõtvalt on:

- 2.1. Arendada haigustrajektooride ja raviteekondade analüüsi meetodikaid kasutades sh tehisintellekti meetodeid, et kirjeldada praeguseid ravipraktikaid, võrrelda neid raviteekondade ja -juhenditega ning modelleerida tervisetulemeid ja ravi majanduslikku mõju.
- 2.2. Analüüsida erinevate faktorite (nt sugu, vanus, erinevad haigused) mõju ravijärgimusele, hinnata ravijärgimuse mõju erinevatele tervisetulemitele ja luua personaalseid ennustumudeleid, mis võimaldavad maandada ravi mittejärgimise riske.
- 2.3. Arendada personaliseeritud ennetusmudeleid, mis võimaldavad vaatlusandmete põhjal tuvastada praeguste haiguse ennetuste ja ravipraktikate kitsaskohti, planeerida senisest paremini haiguste ennetustegevusi ja hinnata pakutud ennetusteenuste majanduslikku mõju

Täiendavate uurimisküsimuste tekkimisel nii EST-Health-30 kui ka kolmandatest isikutest uuringu läbiviijate poolt esitatakse igakordselt Eesti bioetika ja inimuuringute nõukogule ja andmeandjatele uus uuringutaotlus uurimisküsimuste vastavuse hindamiseks käesoleva uuringutaotluse eesmärkidega. Kui täiendav uuring kooskõlastatakse, tagavad EST-Health-30 vastutavad uurijad, et konkreetse projekti uurimismeeskond saab ligipääsu vaid EST-Health-30 andmestiku osale, mis on vajalik vastava uurimisküsimuste lahendamiseks (lähtudes võimalikult vähesete andmete kogumise põhimõttest). Selleks võimaldatakse uurimismeeskonnale ligipääs EST-Health-30 andmestikule Tartu Ülikooli hallatava turvalise töötlemiskeskonna kaudu.

Viited:

1. 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* [Internet]. 2021 May;6(Suppl 2). Available from: <http://dx.doi.org/10.1136/bmjgh-2020-003540>
2. Eesti Tervisekassa. <https://tervisekassa.ee/raviteekondade-arendamine>

10. Uurimismetoodika (kuni 1800 tähemärki, 1 lk)

Valim ja uuringuperiood

Andmetiku aluseks on 30% juhuvalim Eestis alates 2012. aastast tervishoiuteenuseid tarbinud Eesti elanikest (andmesubjektid). Nende andmesubjektide kohta saadakse pseudonüümitud andmed alates 2012 kuni 2026 aastani Tervisekassa andmekogust, retseptikeskuse andmekogust, Tervise Infosüsteemist, surma põhjuste registrist ja vähiregistrist.

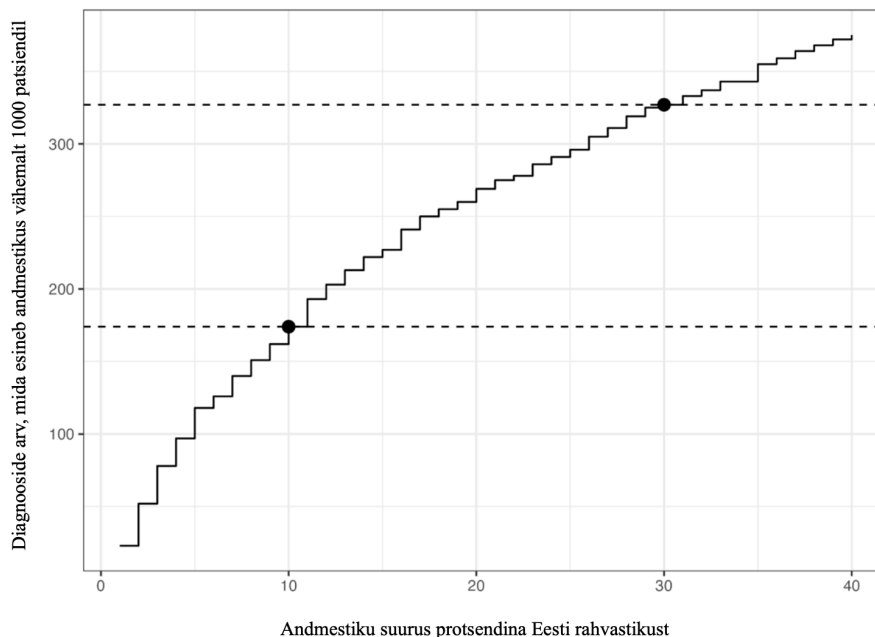
Eelnevalt peatükis 9. "Planeeritava uuringu põhjendus ning uurimisküsimused ja/või hüpoteesid" kirjeldatud eesmärkide saavutamiseks tuleb luua võimalikult pika perioodiga laiapõhjaline andmestik. Andmete vaatlemine üle pika ajaperioodi võimaldab analüüsida tervisesündmuste esituse ja kvaliteedi arengut läbi aja ning uurida terviklikke haigustrajektoore alates ennetustegevustest, esmasdiagnoosist, raviteenuste osutamisest lõpptulemini. Tuginedes uurimisühma varasemale kogemusele teame, et planeeritavateks uuringuteks vajaliku detailsusastmega terviseandmed on Eestis olemas alates 2012. aastast. Värskeimad andmed (sh imikute ja vastsündinute kohta) on vajalikud vastamaks küsimustele, mis puudutavad tervise teenuse osutamise hetkeseisu, uute ravimeetodite kasutust ja efektiivsust ning tervisesüsteemi vastust erinevatele hiljutistele sündmustele. Samuti on laiem andmestiku ajaraam vajalik, et hinnata riskimudelite 10-aastast ennustusvõimet vastavalt rahvusvaheliselt levinud praktikale.

Laiapõhjalisus võimaldab hinnata andmekvaliteeti üle paljude haiguste ja tagab, et loodud andmetöötlusmeetodid üldistuvad uutele uuringutele sõltumata vaatluse all olevast haigusest ja patsiendi grupist. Lähtuvalt 1. ja 2. etapi eesmärkidest ei saa teadusuuringut viia läbi üksikute diagnooside, vanuse ega muude parameetrite järgi kitsendatud andmestikul. Iga piirang vähendab andmestiku esinduslikkust üldpopulatsiooni suhtes ja muudab arendatavad meetodid mitteüldistuvaks, piirates loodud meetodite nii kasutamist järgnevatel uuringutes kui ka rakendatavust tervise infosüsteemi andmekvaliteedi tõstmisel.

Projekti "Tehisintellekti kasutamise võimalused meditsiinis"¹ raames kasutasime 10% juhuvalimit 2012.-2019. aasta andmetest. Projekti käigus saime kinnitust, et erinevate riiklike terviseandmebaaside ühendamine ja laiapõhjalise terviseajaloo kasutamine annab olulist lisandväärtust andmete analüüsis. Näiteks koostöös Naistearstide seltsiga hindasime emakakaela vähi ennetuspraktika vastavust ennetusjuhendile² ja erinevate papilloomviirustüvede levikut Eestis³. Siiski ilmnemise 10% juhuvalimil koostatud andmestiku kasutamisel ka olulised kitsaskohad. Esiteks, oli andmestiku ajaraam liiga lühike, et vastavalt rahvusvahelistele praktikatele hinnata riskimudelite 10-aastast ennustusvõimet. Kuna andmestikku ei uuendatud, puudus info uute haiguste (nt COVID-19), ravipraktikate ning nende esituse ja andmekvaliteedi kohta. Näiteks seetõttu ei olnud võimalik ravijuhendite arendamisel teha koostööd ravijuhendite püsisekretariaadiga (<https://tervis.ut.ee/et/ravijuhendid>) ega panustada rahvusvahelistesse Euroopa Raviameti uuringutesse. Kõige olulisemana aga oleme korduvalt näinud, et 10% juhuvalim on liiga väike mitmete haigusseisundite või detailsemalt defineeritud kohortide uurimisel. Näiteks eesnäärmevähi rahvusvahelises uuringus osaledes saime vastused vaid osadele uurimisküsimustele, sest täpsemate tervisetulemite analüüsimiseks jäi valim liiga väikeseks⁴.

Vajalik patsientide arv sõltub väga uurimisküsimusest ning ka ühe uuringu sees võib see erineda, kuid arvestades patsientide kliinilise pildi varieeruvust ja retrospektiivsete terviseandmete kvaliteeti, oleks hinnanguliselt minimaalne vajalik patsientide arv uuringute jaoks sadades kuid veel parem üle tuhande. Oleme välja arvutanud, et taotletava ajaperioodi ja andmete ulatuse korral on 10%-lise juhuvalimiga andmestikus diagnoose, mida esineb vähemalt tuhandel patsiendil 175, samas kui 30%-lises juhuvalimi korral on selliseid diagnoose ligi kaks korda rohkem - 330 diagnoosi (joonis 1). Suurema valimi korral suureneb oluliselt saadavate tulemuste statistiline usaldusväärsus ja samas ka haiguste hulk, millele saame välja töötatud meetodeid rakendada.

Selle analüüsi põhjal näeme, et teadusuuringule seatud eesmärkide saavutamiseks on minimaalne valimi suurus 30%.



Joonis 1. Vähemalt 1000 patsiendil esinevate diagnooside arvu sõltuvus juhuvalimi suurusest, mis on võetud taotletava ajaperioodi ja andmete ulatusega Eesti terviseandmetest.

Andmete väljastamine, uuendamine ja säilitamine

Esimene väljastusel toimub 2024. aasta oktoobris, kui uurimisgrupile väljastatakse räsi abil pseudonüümitud andmed 2012-2023 aastate kohta. Depseudonüümimise võtit ehk räsifunktsioonid kasutatavat räsiparooli uurimisgrupile ei väljastata. Andmeid uuendatakse regulaarselt, kuid mitte tihedamini kui kord aastas, et mitte üle koormata Eesti tervishoiuandmete eest vastutavate asutuste nagu TEHIK ja Tervisekassa andmeväljastuse võimekust. Andmete uuendamine toimub järgmiselt:

- 2024 aasta andmed väljastatakse andmed 2025. aasta I kvartalis
- 2025 aasta andmed väljastatakse andmed 2026. aasta I kvartalis
- 2026 aasta andmed väljastatakse andmed 2027. aasta I kvartalis

Täpne väljastatav andmekoosseis on toodud taotluse lisa 1. Andmekoosseisu defineerimisel lähtusime eelnevate uuringute kogemusest ning kaasasime tunnused, mis on tüüpiliselt vajalikud andmete OMOP kujule viimiseks ja erinevate terviseuuringute läbiviimiseks. Andmete OMOP kujule viimiseks ja puhastamiseks kasutatakse varasemates projektides loodud tarkvara ja töövoogusid, mida projekti raames täiendatakse vastavalt vajadusele. See tagab projekti kiire käivitamise ning iga-aastase ladusa andmete uuendamise.

Andmeid uuendatakse samamoodi nagu toimus algne andmete edastamine ehk kasutatakse sama räsifunktsiooni ja räsiparooli. Andmete uuendamise käigus isikute arv valimis suureneb, kuna andmestikku lisanduvad sündid ja isikud, kellel varasemalt isikukood puudus. Nende andmete lisamine on oluline, sest see aitab tagada, et igal uuringu aastal on valimis 30% rahvastikust. Uute sündide lisamine võimaldab hinnata imikute ja väikelastega teostatud ravipraktikate muutust ajas ning seeläbi panustada haavatavate rühmade raviteenuste kvaliteedi parandamisse kasutades neid vähekoormavaid uurimismeetodeid.

Peale projekti lõppemist (dets 2027) andmed arhiveeritakse üheks aastaks, et vajadusel võimaldada arhiveerimisperioodil retsenseerimisel olevate teadusartiklite analüüsi täiendamine. Andmed hävitatakse hiljemalt 31. detsembriks 2028 kasutades hävitamise ajahetkel parimat kasutusolevat praktikat vastavalt rakendatavale infoturbestandardile (ISO/IEC 27001). Andmete hävitamine dokumenteerikse hävitisaktis.

Andmete hoidmine ja töötlemine

Peame loodava andmestiku ja läbiviidava uuringu puhul esmatähtsaks isikuandmete kaitse ja infoturbe tagamist. Peamised ohud isikuandmete töötlemisele, nende juhtimiseks rakendatavad kaitsemeetmed ja ohuhinnangud on täpsemalt kirjeldatud lisa 2 "EST-Health-30 andmekaitsealane mõjuhinnang". Andmekaitsealane mõjuhinnang vaadatakse üle ja ajakohastatakse iga olulise muudatuse puhul

andmetöötluse metoodikas ja korraliselt vähemalt kord aastas, et ajakohastada ja vajadusel täiendada ohuhinnangut ja rakendatavaid kaitsemeetmeid.

Uuringu andmetöötlus nõuab kõrget ekspertiisi ja tehnilisi lahendusi, et tagada andmete töötlemine keskses turvatud infrastruktuuris ja välistada andmetöötlus juhuslikes ebaühtlase turbetasemega seadmetes. Koostöös Eesti Teadusarvutuste infrastruktuuriga (<https://etais.ee>) töötab uurimisgrupp infoturbe riskide maandamiseks välja minimaalsusprintsipi tagava andmetele ligipääsu andmise protsessi ja paigutab loodava andmestiku turvalisse andmeanalüüsi serverisüsteemi (andmepuuri) SAPU (<https://sapu.cs.ut.ee>). Andmete hoidmisega SAPU-s väheneb oluliselt andmelekkete risk rakendades andmete eksportimisele ranged piirangud ning logides kõiki andmetega tehtavaid tegevusi kasutaja tasandil kogu andmete säilitamise ja arhiveerimise perioodil. Andmeandjate pöördumisel ja arvestades kohalduvaid isikuandmete kaitse nõudeid, võimaldab uurimisgrupp kontrollida andmetöötluse logiandmeid. Andmete edasisel töötlemisel SAPUs ja uuringu eesmärkide täitmiseks kasutatakse OMOP mudeli jaoks loodud ja publitseeritud tööriistu (Atlas server, Usagi, Data Characterisation, PopulationLevelEstimation ja PatientLevelEstimation R-paketid, uurimisrühma arendatud tööriistad jt). Kusjuures Atlas server on OMOP andmemudelile loodud kohortide defineerimise tööriist, mis võimaldab patsiendigruppe defineerida ilma andmetele ligipääsu omamata.

SAPU serverist väljastatakse andmeid ainult vastutava uurija nõusolekul ja agregeeritud kujul nii, et on tagatud $k \geq 5$ anonüümsus. SAPU serveris olevatele andmetele on ligipääs ainult käesolevas taotluses toodud uurijatel. Uurimisgrupp tagab, et andmetele ligipääs on vastavuses kehtiva seadusandlusega. Täiendavate uurimisküsimuste tekkimisel esitatakse Eesti bioetika ja inimuuringute nõukogule ja andmeandjatele igakordselt taotlus. Kooskõlastatud uurimisküsimuste lahendamiseks võimaldatakse ligipääs andmetele lähtudes minimaalsuse printsibist. Võimalusel viiakse lisandunud uuring läbi nii, et 1. etapi uurimisrühmale edastatakse uue uurimisküsimuse lahendamiseks kood. Uurimisrühm verifitseerib eelnevalt edastatud koodi turvalisuse ning seejärel rakendab koodi andmestikul. Vastutavad uurijad kontrollivad enne tulemuste tagastamist, et saadud tulemused on kooskõlas eetikakomitee antud loaga ja järgivad $k \geq 5$ anonüümsuse printsipi. Isikuandmetega seotud rikkumiste eest kaitsvate meetmete kohta vt p 15 ja taotlusele lisatud andmekaitsealast mõjuhinnangut (lisa 2).

Teadusuuringu tulemused publitseeritakse nii Eesti kui ka rahvusvahelistes teadusajakirjades (eelistatult vabalt kättesaadavate artiklitena), konverentsidel ja ettekannetel, üliõpilaste lõputöodes. Andmestiku metakirjeldus (kirjeldus andmeväljade kohta, ei sisalda reaalseid andmeid) OMOP-andmekogude registritesse, (nt <http://portal.ehden.eu>). Teadusuuringu muid tulemeid (andmestikku kirjeldus, andmehõive statistika ja -hinnangud, andmete standardiseerimise ja tekstikaeve meetodid, üleminekutabelid, teadusuuringute kokkuvõtted jms) tutvustatakse soovi korral andmeandjatele ning otsitakse koostööviise tulemite rakendamiseks andmeandjate hallatavates andmekogudes.

1. etapi uurimisküsimuste metoodika

Teadusuuringute läbiviimiseks on vaja täpselt aru saada andmehõivest ja -kvaliteedist. Näiteks kui paljudel patsientide kohta on olemas vajalik informatsioon, kui suur osa tehtud mõõtmistest on andmetes kajastatud, kas ja kuidas on puuduvad mõõtmised kallutatud või kui usaldusväärsed on diagnoosid. Siin saame ära kasutada fakti, et erinevad andmeallikad kajastavad samu tervisesündmuseid erinevad detailsusastme ja kvaliteediga. Nii saame Tervisekassa andmekogu, retseptikeskuse andmekogu, Tervise Infosüsteemi, surma põhjuste registri ja vähiregistri andmeid ühildades anda hinnangud Eesti terviseandmete andmehõive ja -kvaliteedi kohta. Näiteks Tervise infosüsteemi epikriisides on vaid haigusloo kokkuvõte, mitte aga kõik teostatud uuringud ja analüüsid. Kasutades raviarvetel kajastuvat informatsiooni uuringute ja analüüside tellimise kohta, saame hinnata laborianalüüside andmehõivet ja raporteerimise kallutatust.

Läbi aastate on samasid tervisesündmusi ja teenuseid tähistatud erineva koodiga ehk kodeeritud erinevalt. Et uuringud saaks läbi viia võimalikult pikas ajaaknas, tuleb andmete kodeerimine üle viia ühtsele standardile. Siin lähtume OMOP andmemudeli poolt paika pandud andmestandardist erinevate andmetüüpide jaoks (nt SNOMED, RxNorm, LOINC) ning loome üleminekutabelid Eestis kasutatud koodidelt rahvusvahelisele standardile üle minekuks. Üleminekutabelite loomiseks kasutame nii automaatseid tekstikaeve ja klasterdamise meetodeid kui ka manuaalset ülevaatus ja vastava valdkonna kliinilist ekspertiisi. Loodud üleminekutabelid on rakendatavad ka andmeandmeandjate andmebaaside kvaliteedi ühtlustamiseks.

Terviseinfosüsteemi dokumentides on suur osa informatsioonist poolstruktureeritud kujul või vaba tekstina, eriti vanemate dokumentide puhul. Sellest faktide ja tervisesündmuste eraldamiseks arendame tekstikaeve töövoogu, mis esmalt teksti väiksemateks tükikideks lahutab, eemaldab duplikaadid ning lisab tekstiosadele kuupäevad. Seejärel rakendame juba konkreetseid faktieralduse meetodeid, mis aitavad täiendada struktureeritud kujul puuduvaid andmeid. Näiteks kolesterooli näitu veres hakati struktureeritud

kujul esitama alles 2016. aastast ning üleminek oli haiglali erinev. Varasemalt kajastati kollesterooli näitu tekstiväljas vabatekstina. Seega saab struktureeritud andmete kvaliteeti tõsta vabatekstist leitud andmetega ning seeläbi pikendada analüüsivast aegrida. Standardsemalt esitatud faktide eraldamiseks saame me kasutada regulaaravaldisi, kuid keerukamate kontseptsioonide eraldamiseks tuleb arendada tehismetelli meetodeid nagu BERT ja keelemudelid, mis suudavad eraldada fakte, mille puhul on oluline teksti kontekst ja mõttest aru saamine.

2. etapi uurimisküsimuste metoodika

Haigustrajektoore ja raviteekondade modelleerimisel on mitmeid väljakutseid, mis tulenevad peamiselt kliiniliste juhtumite keerukusest ning patsientide multimorbiidsusest. Keeruline on identifitseerida olulisi sündmusi trajektoiril, võrrelda tegelikke trajektoore ravijuhendite põhjal oodatavatega, automatiseerida trajektoore andmetest õppimist ning modelleerida trajektoore teostamiseks majandusliku mõju analüüsi. Uurimisgrupi teadustöö otsib neile probleemidele tehnilisi lahendusi. Näiteks uurime statistilist ülesindatust oluliste sündmuste tuvastamiseks, graafiesitust haigustrajektoore kirjeldamisel ja Markovi-, masinõppe- ning keelemudeleid trajektoore simuleerimiseks ja sündmuste ennustamiseks. Kirjeldatud eesmärkideni jõudmiseks kasutame kaheastmelist lähenemisviisi. Esmalt läheneme üksikutele diagnoosidele nagu südamepuudulikus, aktiivsus- ja tähelepanuhäire, psühhos ja viljatus, mille põhjal loodud meetodite rakendatavust analüüsime koos kliiniliste ekspertidega. Seejärel testime arendatud meetodite universaalsust rakendades neid paralleelselt laial hulgal haigusseisunditel. Seega, käesoleva teadusuuringu käigus loodavad meetodid ei ole optimeeritud ühelegi konkreetsele haigusele ning nende rakendamine laial diagnooside hulgal, sh erinevate kaasuvate haigustega patsientidel, võimaldab hinnata meetodite universaalsust ja ka süstemaatiliselt kirjeldada tegelikke raviteekondi Eestis.

Ravijärgimuse analüüsil keskendume ravijärgimuse kirjeldamisele üle ravimite, et identifitseerida seda mõjutavaid faktoreid, kirjeldada patsiendi põhlist mõju ning ennustada tuleviku käitumist. Selleks arvutame ravijärgimuse näitajaid erinevate valimite kohaselt, uude lähenemisena kasutame lineaarseid segamudeleid laiapõhjaliste mõjutegurite leidmiseks ning rakendame masinõppe ja tehismetelli meetodeid tulevikukäitumise ennustamiseks. Süstemaatiline ravijärgimuse ja seda mõjutavate faktorite kirjeldamine, kasutades 30% Eesti elanike retseptiandmeid võimaldab teha üldistavaid järeldusi ravijärgimuse kohta, kirjeldada ajalisi trende ning identifitseerida riskifaktoreid.

Personaliseeritud ennetusteenuste arendus vajab suhteliselt sarnast tõendusmaterjali sõltumata meditsiinilisest probleemist. Südameveresoonekonna haiguste ja diabeedi näitel tuvastame praeguse ennetustegevuse kitsaskohad, analüüsides ravimikasutuse ja ravimite võrdlevat efektiivsust. Seejärel tuleb luua masinõppe mudelid, mis hindavad konkreetse haiguse riski. Samuti tuleb hinnata ennetusteenuse rakendamise majanduslikku mõju, milleks saab muu hulgas kasutada Markovi mudeleid ja seotud meetodeid. Antud projekti pikaajaline ja laiapõhjaline terviseandmestik loob suurepärase võimaluse selle töövoogi testimiseks nii, et loodavad mudelid oleksid universaalsed ja üldistatavad erinevatele haigustele ning sihtrühmadele.

Praeguses etapis, kus arendame analüüsimeetodeid, on vaja kõiki taotletud andmeid püstitatud eesmärkide saavutamiseks. Täiendavate uurimisküsimuste tekkimisel esitatakse Eesti bioetika ja inimuuringute nõukogule ja andmeandjatele hindamiseks uus taotlus ning nendele vastamiseks võimaldatakse andmetele ligipääs lähtuvalt andmete minimaalsuse printsiibist.

Viited:

1. https://www.etag.ee/wp-content/uploads/2022/05/RITA_MAITT_LOPPARUANNE_FINAL.pdf
2. Mooses K, Šavrova A, Pajusalu M, et al. Using electronic health records to evaluate the adherence to cervical cancer prevention guidelines: A cross-sectional study. Preventive Medicine, 183, 2024, doi: 10.1016/j.ypmed.2024.107982
3. 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 (4). DOI: 10.1093/jamiaopen/ooad100
4. 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. European Urology, 2023. DOI: 10.1016/j.eururo.2023.06.012

11. Uuritavate valim ja värbamise viisi kirjeldus.

Uuritavate informeerimise ja nõusoleku vormid, ankeetide, küsitluste ja testide vormid esitada taotluse lisadena.

Valimi suurus ja kontrollgruppide olemasolu	<p>Andmetiku aluseks on 30% juhuvalim Eestis alates 2012. aastast tervishoiuteenuseid tarbinud Eesti elanikest. Valim sisaldab kontrollgruppi.</p> <p>Aasta 2024 oktoobris väljastatakse andmed perioodi 01.01.2012-31.12.2023 kohta. Esimesel andmeväljastusel on valimisse kuuluvate isikute arv ligikaudu 380 000. Järgmised andmeväljastused toimuvad vastavalt: (1) 2024 aasta andmed väljastatakse andmed 2025. aasta I kvartalis; (2) 2025 aasta andmed väljastatakse andmed 2026. aasta I kvartalis; (3) 2026 aasta andmed väljastatakse andmed 2027. aasta I kvartalis. 2024-2027 toimuvatel andmeväljastustel väljastatakse uued andmed juba andmestikku kuuluvad isikute kohta. Lisanduvad andmed nende isikute kohta, kellel varasemalt isikukood puudus või kes esmakordselt vaatlusalusel perioodil tervishoiuteenuseid tarbisid ja kelle isikukood vastab räsifunktsiooniga seatud tingimustele. Täpsemalt on valiku suurust põhjendatud punktis "10. Uurimismetoodika".</p>
Kes värbab uuritavaid ja kuidas/kus/kelle poolt võetakse informeeritud nõusolek? (kui on asjakohane)	<p>Käesolevas uuringus uusi isikuandmeid ei koguta ja andmesubjektidega ühendust ei võeta. Kasutatakse juba olemasolevaid terviseandmeid teadusuuringute läbiviimise eesmärgil.</p>
Kuidas ja kelle hulgast toimub uuritavate valik? Millised on uuritavate kaasamise või väljajätmise kriteeriumid?	<p>Tervisekassa valib juhusliku räsiparooli ja edastab selle krüpteeritult teistele andmeallikatele (TEHIK, TAI). Kõik andmeallikad kasutavad seda räsiparooli, et arvutada räsidi oma andmekogudes olevatest isikukoodidest. Räsi iga sümbol on üks kuueteistkümnese sümbolilisest hulgast: a-f või 0-9. Need jaotuvad ühtlaselt. Valimisse kuuluvad isikud, kelle räsi esimene sümbol kuulub hulka {a,b,c,d,e,f,0} ja teine sümbol hulka {a,b,c,d,e,f,0,1,2,3,4}. Sellisel juhul satub valimisse $(7/16) \cdot (11/16) \cdot 100 = 30.08\%$ isikukode. Räsi arvutamise protsess on täpsemalt kirjeldatud lisas 2 "EST-Health-30 andmekaitsealane mõjuhindang".</p> <p>Räsi või tema osa kasutatakse ühtlasi isikute pseudonüümideks, sest ilma andmeväljastuses kasutatud räsiparooli teadmata ei ole neid võimalik tagasi isikukoodideks kodeerida. Tervisekassa poolt valitud räsiparooli uurimismeeskonnale ei väljastata, mille tõttu uurimismeeskond ei tea, kes täpselt valimisse kuuluvad. Meetme eesmärk on vähendada veelgi privaatsuse riive riski. Küll aga säilitavad andmeallikad räsiparooli kuni projekti perioodi lõpuni, et tagada andmete uuendamise võimalus.</p> <p>Uuritavad valitakse räsipõhise juhuvalimi alusel Eestis alates 2012. aastast tervishoiuteenuseid tarbinud Eesti elanikest. Jahuvalimi kasutamine aitab vähendada valiku nihet, kuna kõikidel isikutel on sarnane tõenäosus valimisse sattuda¹. Valimisse kaasamise kriteeriumid:</p> <ul style="list-style-type: none"> - omab Eesti isikukoodi, mis vastab räsifunktsioonile - on uuringuperioodil tarbinud tervishoiuteenuseid <p>Viited</p> <ol style="list-style-type: none"> 1. Tyrer S, Heyman B. Sampling in epidemiological research: issues, hazards and pitfalls. BJPsych Bull. 2016 Apr;40(2):57-60. doi: 10.1192/pb.bp.114.050203.
Sekkumiste liik (füüsiline, vaimne või andmed, sh eriliiki isikuandmed)	<p>Tegemist on teisese andmekasutusega, seega kasutatakse retrospektiivselt tervishoiuteenuste pakkumisel Eesti tervishoiu</p>

	infosüsteemidesse kogutud andmeid. Uuritavatelt täiendavaid andmeid ei koguta ja nendega ühendust ei võeta. Tegemist on eriliiki isikuandmetega.	
Koormus uuritavale (kontaktivõtmise viisid, visiitide arv, uuringute tüüp ja arv, kutsete saatmise kordus jms)	Uuritavatele isikutele andmestiku loomine ega sellel läbiviidavad terviseuuringud lisakoormust ei põhjusta. Uuritavatelt täiendavaid andmeid ei koguta ja nendega ühendust ei võeta.	
12. Koeproovide väljastamine kolmandatele osapooltele (RNA, DNA, plasma vms)		
Mitme geenidoonori koeproove ja mis tüüpi koeproove väljastatakse?	0, koeproove ei kasutata.	
Kui palju ühe geenidoonori kohta koeproove väljastatakse?	0, koeproove ei kasutata.	
Kuhu koeproov väljastatakse (riik, asutuse nimetus, aadress)?	Koeproove ei väljastata.	
Mida tehakse järelejäänud koeproovidega (kas ülejääk hävitatakse või saadetakse tagasi)?	Koeproove ei kasutata.	
13. Uuringu eetiliste aspektide analüüs (3600 tähemärki, kuni 2 lk). Kõik uuringud, mille objektiks on inimesed, peavad olema läbi viidud, arvestades eetilisi nõudeid, eelkõige autonoomia austamise, heategemise ja kahju vältimise ning õigluse printsiipe. https://www.coe.int/en/web/bioethics/guide-for-research-ethics-committees-members . vt ka https://www.etag.ee/wp-content/uploads/2020/01/Eetika_Tabel_EST_2020.pdf		
Uuritavate isikuandmeid töödeldakse isikuandmete kaitse üldmääruse (EL) 2016/679 artikkel 9(2)(j) ja isikuandmete kaitse seaduse (IKS) § 6 lõigete 1, 3 ja 4 alusel teadusuuringu läbiviimise eesmärgil. Isikuandmete töötlemise õigusliku aluse põhjalikum hinnang on toodud käesoleva taotluse lisas 2 "EST-Health-30 andmekaitsealane mõjuhinnang", peatükis "töötlemise õiguslikud alused". Võrdse kohtlemise printsiip – käesolev projekt ei diskrimineeri kedagi rahvuse, soo, vanuse ega muu kuuluvuse alusel. Kõikide ühiskonna- ja vanusegruppide puhul on eesmärk toetada elukvaliteedi parandamist ja tervena elatud aastate arvu suurendamist. Juhuvalimi tõttu on kõikidel inimestel võrdne tõenäosus valimisse sattuda. Heategemise printsiip – antud projektist saadav kasu on uuritavatele pigem kaudne, kuna alustatakse baasteadmiste loomisega ja otsene kliinilises praktikas kasutatav kasu haigusega tegelemiseks võib tekkida alles mitme aasta pärast. Seega saavad antud tööst kasu pigem tulevased patsiendid. Kahju vältimise printsiip - Planeeritav projekt on kooskõlas mittekahjustamise printsiibiga, kuna uuring ei koorma patsiente ega põhjusta riske nende tervisele.		
13 a Inimesed		
Abiküsimused	Ei	Jah
Kas uurimisobjektiks on inimesed?		Jah, uurimisobjektiks on inimesed, kuid analüüs toimub olemasolevatel pseudonüümitud terviseandmetel, mille puhul vastavalt isikuandmete kaitse seadusele informeeritud nõusolekut vaja ei ole.

Kas uurimisobjektiks on haavatavad isikud või isikute grupid?		Kuna valim koostatakse üle kõigi Eesti inimeste, kuulub andmestikku ka haavatavaid isikuid. Samas mitmete haavatavate gruppide puhul, kes andmestikku satuvad (näiteks lapsed, rasedad, vaimsete häirete inimesed jt), saab teisene andmete kasutamine anda olulist infot ravi tõhustamiseks ilma täiendava uuringu koormuseta.
Kas uurimisobjektiks on isikud, kes ei saa ise anda teadlikku nõusolekut uuringus osalemiseks (sh piiratud teovõimega isikud)?		Andmestik luuakse olemasolevate pseudonüümitud terviseandmete teisesel kasutusel, mille puhul vastavalt isikuandmete kaitse seadusele informeeritud nõusolekut vaja pole.
Kas uurimisobjektiks on alaealised?		Kuna valim koostatakse üle kõigi Eesti inimeste, kuuluvad andmestikku ka alaealised. Alaealiste kaasamine on vajalik, et hinnata teadusuuringute tulemuste üldistatavust ka alaealiste erinevatele vanusegruppidele. Lisaks, alaealiste väljajätmine ei võimaldaks anda Eesti rahvastiku kohta esinduslikku pilti. Loodavad analüüsimeetodid peavad olema rakendatavad ka alaealistega seotud haiguste ja ravitrajektoorie uurimisel ning ennetusmodelite loomisel, mistõttu on oluline loodavasse andmestikku kaasata alaealiste andmed samadel põhimõtetel kui ülejäänud andmed. Juhuleide üksikisiku tasandil ei teki, kuna tegemist on statistilise analüüsiga. Uuritavatega antud uurimistöö raames ühendust ei võeta.
Kas uurimisobjektiks on patsiendid?		Jah.
Kas uurimistöös kogutakse inimestelt bioloogilisi proove? Kas inimestelt võetud bioloogilisi proove kavatakse eksportida kolmandasse riiki (https://www.aki.ee/et/teenused-pordumisvormid/andmete-edastamine-valisriiki) või importida neid teisest riigist Eestisse?	Ei, bioloogilisi proove ei koguta.	
13 b Isikuandmed ja andmestikud		
	Ei	Jah

<p>Kas uurimistöö käigus kogutakse või analüüsitakse isikuandmeid, sh eriliiki isikuandmeid?</p>		<p>1) Täpne andmete koosseis, mis on vajalik uuringu läbiviimiseks (võib esitada lisana).</p> <p>Täpne andmete koosseis koos põhjendusega on esitatud taotluse lisas 1.</p> <p>2) Kinnitada, et informeeritud nõusolek on olemas või saadakse enne uuringu algust, kui uuring põhineb nõusolekul.</p> <p>Uuringus kasutatakse olemasolevaid pseudonüümitud terviseandmeid, mille puhul vastavalt isikuandmete kaitse seadusele informeeritud nõusolekut vaja ei ole.</p> <p>3) Selgitada, miks on kõik töödeldavad andmed asjakohased ja vajalikud (lähtudes andmete minimeerimise põhimõttest).</p> <p>Täpne andmete koosseis koos põhjendusega on esitatud taotluse lisas 1. Eesti eri terviseandmekogudes sisaldub patsientide tervise kohta erinevat liiki ja eri kvaliteediga infot. Selleks, et saada tervisevaldkonna küsimuste uurimiseks patsiendi tervises seisundist kvaliteetne pilt, on vajalik vaadelda korraga andmeid mitmest andmekogust (vt ka "11. Uurimismetoodika").</p> <p>Tervise infosüsteemi epikriisides on kirjas diagnoosid, raviskeemid, saatekirjades ja saatekirja vastustes ka laborianalüüside tulemused, kuid info võib olla puudulik. Epikriisides sisaldub vaid haigusloo kokkuvõte, mitte aga kõik teostatud uuringud või väljakirjutatud ravimid, mis mitmetes uuringutes on oluline info. Teiselt poolt on epikriisid ainsad tervisedokumendid, kus sisaldub vabas vormis kirja pandud info patsiendi kaebuste, üldseisundi, allergiate, ravi kõrvalmõjude jms kohta. Tekstilistest osadest suudame eraldada tehisintellekti meetodite abil ka muud olulist infot, mis kodeeritud väljades puudub, näiteks patsiendi kaebusi, ravimite nõrgemaid kõrvalmõjusid. Tervisekassa andmekogu info tervishoiuteenuste kohta on täielikum ja sisaldab ka teenuste hindu, kuid selle detailsus on samas madalam (mitmed tervishoiuteenused märgitud sama koodiga) ja laborianalüüside kohta puuduvad analüüsitulemused. Kõige parema pildi patsiendile välja kirjutatud ravimite osas annab retseptikeskus, lisaks on seal ka info ravimi väljaostmise kohta, mis on väga oluline indikaator ravijärgimuse hindamiseks. Samas</p>
--	--	--

		<p>puudub retseptikeskuses info käsimüügiravimite kohta (seda infot võib potentsiaalselt leida epikriisidest). Paljude tervisevaldkonna uuringute puhul (nt pahaloomulised kasvaja) on oluline uurida suremust, selleks on kõige kvaliteetsem info kirjas surma põhjuste registris, mis sisaldab nii surma kuupäeva kui ka spetsialisti poolt kinnitatud surma põhjuseid.</p> <p>Andmeväljade valikul oleme lähtunud uurimistöö eesmärkidest tuginedes uurimismeeskonna senistele kogemustele ja minimaalsuse printsiibist. Laiapõhjalisus võimaldab hinnata andmekvaliteeti üle paljude haiguste, sh arvestades kaasuvaid haigusi, ja tagab, et loodud andmetöötlusmeetodid üldistuvad uutele uuringutele sõltumata vaatluse all olevast haigusest ja patsiendi gruppidest. Lähtuvalt sellest, et üheks taotletava teadusuuringu eesmärgiks on, et arendatavad analüüsimeetodid ja ennetusmeetodid üldistuksid ka uutele andmestikele ja haigustele, ei saa me piirata andmestikku ei diagnooside, vanuse ega muude parameetrite järgi. Iga piirang vähendaks andmestiku esinduslikkust üldpopulatsiooni suhtes ja muudaks arendatavad meetodid vähem üldistuvaks ning piiraks loodud meetodite kasutamist teistes uuringutes.</p> <p>Ajaperioodi määramisel oleme lähtunud varasemast kogemusest 2012.-2019. aasta andmetega ning teadustöö eesmärkidest. Pikk ajaperiood tagab, et andmetes on näha tervisesündmuste esituse ja kvaliteedi areng läbi aja ning me saame uurida terviklikke haigustrajektoore, alates ennetustegevustest, esmasdiagnoosi, raviteenuste osutamise ja lõpptulemini. Samuti võimaldab loodava andmestiku ajaraam hinnata riskimudelite 10-aastast ennustusvõimet vastavalt rahvusvahelistelt levinud praktikatele.</p> <p>4) Kas andmesubjektid on tuvastatavad? Kui jah, siis kirjeldada, kuidas on täidetud järgmised tingimused:</p> <ul style="list-style-type: none"> a. pärast tuvastamist võimaldavate andmete eemaldamist ei ole andmetöötluse eesmärgid enam saavutatavad või neid oleks ebamõistlikult raske saavutada; b. teadus- või riikliku statistika tegija hinnangul on selleks ülekaalukas avalik huvi; c. töödeldavate isikuandmete põhjal ei muudeta
--	--	---

		<p>andmesubjekti kohustuste mahtu ega kahjustata muul viisil ülemäära andmesubjekti õigusi.</p> <p>Andmesubjektid ei ole otseselt tuvastatavad. Kaudse tuvastamise risk säilib. Kaudse tuvastamisega seotud riskide maandamiseks võtame kasutusele mitmeid meetmeid, mis on täpsemalt kirjeldatud lisas 2 "EST-Health-30 andmekaitsealane mõjuhindang".</p>
<p>Kas uurimistöö hõlmab üksikisiku süsteemset jälgimist, tema andmeprofiili kogumist või töödeldakse suures ulatuses eriliiki ja/või tundlikke andmeid või kasutatakse (sekkuvaid) andmete töötlemise meetodeid varjatud viisil (nt elulemuse uuringud, jälgimine, järelevalve, audio ja video salvestamine, geo-positsioneerimine jne) või mistahes andmete töötlemise protsessi, mis võib kahjustab uuritavate õigusi ning vabadust?</p>		<p>Uurimistöö käigus töödeldakse suures mahus andmesubjektide terviseandmeid ja uuritakse nende raviteekondasid, ravijärgimust ja personaliseeritud ennetusmeetmeid. Seega eksisteerib teoreetiliselt oht uuritavate õigustele ja vabadustele, kuid uurimistöö läbiviimisel on rakendatud hulgaliselt meetmeid erinevate riskide maandamiseks ning need on kirjeldatud lisas 2 "EST-Health-30 andmekaitsealane mõjuhindang".</p>
<p>Kas uurimistöös analüüsitakse eelnevalt kogutud isikuandmeid?</p>		<p>Uurimistöö käigus analüüsitakse eelnevalt igapäevase tervishoiusüsteemi toimimise käigus kogutud terviseandmeid. Andmekogude loetelu on toodud punktis 14 ja täpne andmete koosseis koos põhjendusega on esitatud taotluse lisas 1. Analüüs toimub retrospektiivselt.</p>
<p>Kas uurimistöös analüüsitakse avalikult kättesaadavaid andmeid?</p>	Ei	
<p>Kas kavatakse edastada isikuandmeid või võimaldada neile juurdepääs kolmandast riikidest (https://www.aki.ee/et/teenused-poorordumisvormid/andmete-edastamine-valisriiki)?</p>	Ei, isikuandmeid kolmandatesse riikidesse ei edastata.	
<p>Kas uurimistöö lõppedes toimub isikuandmete hävitamine/anonüümimine?</p>		<p>Jah. Peale uurimisprojekti arhiveeritakse andmed üheks aastaks, et vajadusel võimaldada arhiveerimisperioodil retsenseerimisel olevate teadusartiklite analüüsi täiendamine. Andmed hävitatakse hiljemalt 31. detsembriks 2028. Andmete hävitamise protokoll saadetakse kõikidele vastutavatele andmetöötajatele.</p>
13 c Teised eetilised küsimused		

Kas uurimistöö läbiviimine võib kaasa tuua eelpool kirjeldamata eetilisi riske?		Jah, kasutame ka tehisintellekti meetodeid, millega kaasnevad ülal kirjeldamata riskid. Tehisintellekti kasutamisega seonduvate riskide analüüs on toodud lisas 3.
14. Täita, kui uuring põhineb andmekogu ja/või andmeallika andmetel.		
Andmekogu ja/või andmeallika nimetus <ol style="list-style-type: none"> 1. Tervise infosüsteem (ambulatoorne, statsionaarne, kodu- ja iseseisva statsionaarse õenduse, koduõenduse ja päevaravi epikriisid; saatekirjad, saatekirjad haiglaravile, õendusabile, ambulatoorsele vastuvõtule, e-konsultatsioonile ja õendusabile; saatekirja vastused; immuniseerimise teatised) 2. Tervisekassa andmekogu (raviarved ja kindlustuskaitse andmed) 3. Retseptikeskus 4. Vähiregister 5. Surma põhjuste register 		
Isikuandmete töötlemise eesmärk <p>Eesmärk on Eesti terviseandmete väärimine läbi andmeteaduse meetodite, et pakkuda kvaliteetset tõendust tõhusate patsiendikesksete tervishoiu- ja ennetusteenuste osutamiseks.</p>		
Andmekoosseis ja periood, mille kohta andmed kogutakse (vajadusel lisana) <p>Andmekoosseis on esitatud lisana (Lisa 1. Andmekoosseis). Andmeperiood on 2012-2026.</p>		
15. Isikuandmete kaitse meetmete kirjeldus, sealhulgas andmete hoidmise, säilitamise, turvalisuse ja kustutamise kohta, sh andmete ja/või koodivõtme kustutamise kuupäev (kuni 1800 tähemärki, 1 lk).		
Kirjeldada ja põhjendada uuringu vajaduseks kogutud andmete säilitamist ja tähtaega.	Andmeandjatelt saadud algandmed paigaldatakse SAPUs spetsiaalselt kirjutuskaitstud kausta, et vältida andmete juhuslikku hävimist või kahjustumist. SAPU-st tehakse regulaarselt krüpteeritud varukoopiaid Eesti Teadusarvutuste infrastruktuuri lindirobotile, mis asub füüsiliselt teises asukohas (andmekeskuses). Andmed arhiveeritakse üks aasta peale projekti lõppu. Andmete arhiveerimine üheks aastaks peale projekti lõppu on vajalik retsenseerimisel olevate artiklite publitseerimisprotsessi lõpetamiseks. Arhiveerimisperioodi lõpus, hiljemalt 31.12.2028 andmed kustutatakse. Projekti käigus võib osutada, et ka edaspidine säilitamine, arendamine ja kasutus, sh vahepeal saadud teadustulemuste valideerimise ja täpsustamise eesmärgil on vajalik. Sellise vajaduse tekkimisel esitab uurimisgrupp vastavasisulise taotluse asjakohas(t)ele eetikakomitee(de)le. Uuringu tarbeks kogutud andmete säilitamine on täpsemalt kirjeldatud lisas 2 "EST-Health-30 andmekaitsealane mõjuhinna".	
Kirjeldada isikuandmete pseudonüümimise protsessi ja vahendeid.	Isikuandmed pseudonüümitakse andmekogude vastutavate/volitatud töötajate poolt, kasutades üksnes andmeandjatele teadaolevat räsiparooli, mida nad omavahel jagavad krüpteeritult. Seda räsiparooli kasutatakse, et arvutada räsids oma andmekogudes olevatele isikukoodidele. Pseudonüümina kasutatakse saadud räsi. Andmeallikad asendavad oma	

	<p>andmetes isikukoodid pseudonüümidega enne andmete uuringumeeskonnale väljastamist.</p> <p>Pseudonüümimise protsess ja vahendid on täpsemalt kirjeldatud lisas 2 "EST-Health-30 andmekaitsealane mõjuhinnaang".</p>
Kas kavatakse geenidonorite isikuandmeid depseudonüümida?	<p>Ei, geenidonorite andmeid ei töödelda.</p> <p>Uuritavate osas on korralduslikult tagatud, et uurimisgrupi liikmetel ei ole lubatud ega võimalik andmeid depseudonüümida (mh ei avaldata uurimisgrupi liikmetele räsiparooli).</p>
Kas toimub isikuandmete transportimine ning kirjeldada, kuidas on tagatud andmete turvalisus.	<p>Iga andmeandja väljastab valimisse kuuluvate isikute kohta oma andmebaasist pseudonüümitud andmed uuringumeeskonnale (ilma isiku otsest tuvastamist võimaldavate andmeteta). Kaudse tuvastamise riski maandamiseks rakendatakse täiendavaid andmekaitse mõjuhinnaangus kirjeldatud meetmeid, sh anonüümimise rakendust vaba teksti väljadel ja otsest tuvastamist võimaldavate andmeväljade (nt aadress) välistamist. Uuringumeeskond ühendab saadud andmekomplektid pseudonüümide alusel. Nii on tagatud, et andmeandjad töötlevad vaid nende juures hoitavaid andmeid ega näe teiste asutuste andmeid.</p> <p>Kirjeldatud viisil andmete vahetamisel ei liigu samaaegselt kunagi isiku otsest tuvastamist võimaldavad isikuandmed ja tema terviseandmed.</p> <p>Andmete edastamiseks sõlmitakse leping iga andmekogu omaniku ja Tartu Ülikooli vahel. Andmete edastus toimub krüpteeritult läbi turvalise andmevahetusserveri, mis on üles seatud Eesti Teadusarvutuste infrastruktuuri poolt või kasutades andmeallikate tavapäraseid andmete väljastamise viise vastavalt nende sisemistele protseduurireeglitele (enamasti konkreetsele juhtivuurijale parooliga ligipääsetav kataloog andmeallika serveris, andmefail on krüpteeritud).</p> <p>Analüüsiks viib uuringumeeskond need andmed eraldi virtuaalserverisse SAPU (vt järgmine punkt), mida haldab Eesti Teadusarvutuste infrastruktuur.</p> <p>Isikuandmete turvaline transportimine on täpsemalt kirjeldatud lisas 2 "EST-Health-30 andmekaitsealane mõjuhinnaang".</p>
Kirjeldada, kuidas on andmed kaitstud loata või ebaseadusliku töötlemise eest.	<p>Peame tähtsaks loodava andmestiku infoturbelist kaitset, mis nõuab kõrget ekspertiisi ja tehnilisi lahendusi, et välistada andmete hoidmine juhuslikes ebaühtlase andmekaitsega arvutites. Koostöös Eesti Teadusarvutuste infrastruktuuriga (https://etais.ee) töötab uurimisgrupp infoturbe riskide maandamiseks välja minimaalsusprintsipi tagava andmete ligipääsu andmise protsessi ja paigutab loodava andmestiku turvalisse andmeanalüüsi serverisüsteemi (andmepuuri) SAPU (https://sapu.cs.ut.ee). Andmete hoidmisega SAPU-s väheneb oluliselt andmelekkete risk rakendades andmete eksportimisele ranged piirangud ning logides kõiki andmetega tehtavaid tegevusi kasutaja tasandil kogu andmete säilitamise ja arhiveerimise perioodil. Soovi korral tagatakse andmeandjatele andjatele ligipääs logidele. Andmete edasise töötlemisel SAPUs ja uuringu eesmärkide täitmiseks kasutatakse OMOP mudeli jaoks loodud ja publitseeritud tööriistu (Atlas server, Usagi, Data Characterisation, PopulationLevelEstimation ja PatientLevelEstimation R-paketid, uurimisrühma arendatud tööriistad jt). Kusjuures Atlas server on OMOP andmemudelile loodud kohortide defineerimise tööriist, mis võimaldab patsiendigruppe defineerida ilma andmetele</p>

	<p>ligipääsu omamata. Seega on Atlas tööriist vajalik just minimaalsusprintsipi tagamiseks.</p> <p>SAPU serverist väljastatakse andmeid ainult vastutava uurija nõusolekul ja agregeeritud kujul nii, et on tagatud $k \geq 5$ anonüümsus. SAPU serveris olevatele andmetele on ligipääs ainult käesolevas taotluses toodud uurijatel. Uurimisgrupp tagab, et andmetele ligipääs on vastavuses kehtiva seadusandlusega. Täiendavate uurimisküsimuste tekkimisel esitatakse Eesti bioetika ja inimuuringute nõukogule ja andmeandjatele hindamiseks uus taotlus. Lisandunud uurimisküsimuste lahendamiseks antakse ligipääs andmetele lähtudes minimaalsuse printsiibist. Võimalusel viiakse lisandunud uuring läbi nii, et 1. etapi uurimisrühmale edastatakse uue uurimisküsimuse lahendamiseks kood, mille uurimisrühm läbi jooksub ning tagastab ainult analüüsi tulemused lähtudes $k \geq 5$ anonüümsusest. Loata või ebaseadusliku töötlemise eest kaitsvate meetmete kohta vt taotlusele lisatud "EST-Health-30 andmekaitsealane mõjuhindang" (lisa 2).</p>
<p>Kinnitan, et kõik uuringu läbiviijad on teadlikud projekti läbiviimisega kaasnevatest eetilistest ja isikuandmete kaitsega kaasnevatest nõuetest.</p>	
<p>Vastutavate uurijate allkirjad</p> <p>/Jaak Vilo digiallkiri/ /Raivo Kolde digiallkiri/ /Sven Laur digiallkiri/ /Sulev Reisberg digiallkiri/</p>	<p>Taotluse esitamise kuupäev</p> <p>20.08.2024</p>
<p>Taotluse EBIN ID (täidab hindaja)</p>	

Lisadokumentide loetelu:

1. Taotletav andmekoosseis (MS Exceli fail kolmel lehel)
2. Andmekaitsealane mõjuhindang
3. AI eetika küsimused
4. Viited vastutavate uurijate avalikele CV-dele:
 - a. Jaak Vilo: https://www.etis.ee/CV/Jaak_Vilo/est/
 - b. Raivo Kolde: https://www.etis.ee/CV/Raivo_Kolde/est/
 - c. Sven Laur: https://www.etis.ee/CV/Sven_Laur/est/

d. Sulev Reisberg: https://www.etis.ee/CV/Sulev_Reisberg/est/