TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

GENOOMIKA INSTITUUT

Töövoog fenotüüpide defineerimiseks Eesti Geenivaramu andmestikus

Magistritöö

30 EAP

Eret Orumaa

Juhendajad kaasprofessor Triin Laisk

teadur Kristi Läll

TARTU 2025

INFOLEHT

Töövoog fenotüüpide defineerimiseks Eesti Geenivaramu andmestikus

Komplekshaiguseid saab uurida ülegenoomsete assotsiatsiooniuuringutega, mille eesmärgiks on leida seoseid fenotüübi ja inimestel leiduvate geneetiliste variantide vahel. Komplekshaiguste uurimisel on tähtis, et erinevate uuringute tulemused oleksid omavahel võrreldavad ja et fenotüübid oleksid uuringutes defineeritud sarnaselt. Selleks, et võimaldada kureeritud fenotüüpide ühtset kasutamist, on loodud komplekshaiguste definitsioonide raamatukogud, millest üks tuntuim on FinnGen fenotüüpide definitsioonide raamatukogu. Assotsiatsiooniuuringute võimsust parandab suurema valimi kasutamine ning suurem valim on võimalik saada erinevate biopankade andmeid harmoniseerides. Magistritöö käigus loodi automatiseeritud töövoog juhtude ja kontrollide määramiseks Eesti Geenivaramu geenidoonorite andmestikus vastavalt erinevatele fenotüübi definitsioonidele.

Märksõnad: GWAS, FinnGen, fenotüpiseerimise tarkvara, harmoniseerimine

CERCS kood: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika; B790 Kliiniline geneetika

Workflow for defining phenotypes in the Estonian Biobank dataset

Complex diseases can be studied through genome-wide association studies, which aim to find associations between the phenotype and genetic variants. When studying complex diseases, it is important that the results of different studies are comparable and that phenotypes are defined similarly across studies. To host curated phenotypes, phenotype libraries have been created, with one of the better known ones being the FinnGen phenotype library. The power of association studies is increased by using larger sample sizes, and that can be achieved by harmonizing data from different biobanks. In this thesis, an automated workflow was created for identifying individuals in the Estonian Biobank dataset for case and control groups based on phenotype definitions.

Keywords: GWAS, FinnGen, phenotyping software, harmonization

CERCS code: B110 Bioinformatics, medical informatics, biomathematics, biometrics; B790 Clinical genetics

SISUKORD

INFOLEHT....2

SISUKORD....3

KASUTATUD LÜHENDID....5

SISSEJUHATUS....6

1. KIRJANDUSE ÜLEVAADE....7

1.1 Komplekshaigused....7

1.2 Ülegenoomsed assotsiatsiooniuuringud....7

1.3 Biopankade olulisus....10

1.4 Fenotüüpide defineerimine....12

1.5 FinnGen....13

1.6 Fenotüüpide harmoniseerimine....15

2. EKSPERIMENTAALOSA....18

2.1 Töö eesmärgid....18

2.2 Materjal ja metoodika....18

2.2.1 Valminud töövoo ülevaade....18

2.2.2 Andmefailide lugemine....20

2.2.2.1 Geenidoonorite andmete lugemine....21

2.2.2.2 Fenotüüpide definitsioonide lugemine....24

2.2.3 Juhtude ja kontrollide leidmine....29

2.2.4 Abifailid....41

2.3 Tulemused ja arutelu....43

2.3.1 Maatriksi moodustamine....43

2.3.2 Programmi testimine päris andmetega....44

2.3.3 Tehtud töö mõju analüüsides....49

2.3.4 Programmi edasiarenduse võimalused....50

KOKKUVÕTE....52

SUMMARY....53

KASUTATUD KIRJANDUS....55

KASUTATUD VEEBIAADRESSID.....62

LISAD..... 64

Lisa 1: Biopankade võrdlus.....64

Lisa 2: FinnGen fenotüüpide definitsioonide fail..... 67

Lisa 3: FinnGen fenotüüpide definitsioonide failis leitud vead ja erisused.....68

LIHTLITSENTS..... 71

KASUTATUD LÜHENDID

ATC	Anatoomilis-terapeutiline keemiline kood ravimite klassifitseerimiseks (ingl *Anatomical Therapeutic Chemical Classification System*).
COD	Surma põhjus (ingl *Cause of Death*). FinnGen fenotüüpide raamatukogus kasutatav lühend.
CTV3	Standardiseeritud terminoloogiasüsteem (ingl *Clinical Terms Version 3*).
GWAS	Ülegenoomne assotsiatsiooniuuring (ingl *genome-wide association study*).
HD	Haiglast väljakirjutamine (ingl *Hospital Discharge*). FinnGen fenotüüpide raamatukogus kasutatav lühend.
KELA	Soome sotsiaalkindlustusamet (soome *Kansaneläkelaitos*). KELA ravimi kood on number, millega Soomes tähistatakse ravimit või ravimirühma, mille eest patsient saab KELA kaudu ravimihüvitist.
LD	Aheldustasakaalutus (ingl *linkage disequilibrium*).
NCSP	NOMESCO kirurgiliste protseduuride klassifikatsioon (ingl *NOMESCO Classification of Surgical Procedures*).
NOMESCO	Põhjamaade Meditsiinistatistika Komitee (ingl *Nordic Medico-Statistical Committee*).
OPCS	Meditsiiniliste protseduuride ja operatsioonide klassifikatsioonisüsteem (ingl *Office of Population, Censuses and Surveys: Classification of Interventions and Procedures*).
Phecode	Klassifikatsioonisüsteem fenotüüpidele (ingl *phenotypic code*).
RHK	Rahvusvaheline Haiguste Klassifikatsioon (ingl *International Classification of Diseases*).
RHK-O-3	Rahvusvaheline Haiguste Klassifikatsioon kasvajatele (ingl *International Classification of Diseases for Oncology, 3rd Edition*).
SNOMED CT	Rahvusvaheline kodeeritud kliiniline meditsiiniterminoloogia (ingl *Systematized Nomenclature of Medicine – Clinical Terms*).
SNP	Üksiku nukleotiidi polümorfism (ingl *single nucleotide polymorphism*).

SISSEJUHATUS

Komplekshaigused on haiguste klass, mille kujunemisel mängivad rolli nii geneetilised tegurid kui keskkonnategurid. Komplekshaiguste geneetilist tausta saab uurida ülegenoomsete assotsiatsiooniuuringutega, millega otsitakse seoseid fenotüübi ja inimestel leiduvate geneetiliste variatsioonide vahel. Binaarsete tunnuste uurimisel võrreldakse kahte gruppi inimesi – juhud ja kontrollid. Inimesed jaotatakse gruppidesse vastavalt nende fenotüübile ning otsitakse statistiliselt olulisi erinevusi kahe grupi geneetilises varieeruvuses.

Uuringute võimsust ehk geneetiliste seoste leidmise tõenäosust aitab parandada suurema valimi kasutamine. Selle jaoks saab meta-analüüside jaoks kombineerida erinevate biopankade andmeid. Biopangad on bioloogilise materjali ja terviseandmete kollektsioonid, mida saab kasutada teadusuuringuteks. Erinevatest biopankadest andmete kombineerimine eeldab, et fenotüübid on ühtselt defineerineeritud ja andmed omavahel harmoniseeritud. Fenotüüpide defineerimisel kasutatakse nõudeid andmete päritolule, diagnooside, protseduuride ja ravimite koode ning muid terviseandmeid. Fenotüüpide kvaliteetne defineerimine on vajalik, et valida uuringuks õiged isikute grupid juhtudeks ja kontrollideks ning sellega tagada uuringu korratavus ja usaldusväärsus.

Kureeritud fenotüüpide tekitamiseks ja seega fenotüüpide ühtseks ja täpseks defineerimiseks on loodud komplekshaiguste definitsioonide raamatukogud. Üks tuntud raamatukogu on Soome biopanga projekti FinnGen kliiniliste fenotüüpide definitsioonide raamatukogu. Selles on loodud fenotüüpide definitsioonid, mida pidevalt uuendatakse ja täiendatakse ning tööriist, millega saab sirvida fenotüüpide definitsioone ja töövoogu juhtude ja kontrollide leidmiseks.

Magistritöö eesmärgiks oli luua automatiseeritud töövoog, mille abil saab Eesti Geenivaramu geenidoonorite andmestikus määrata juhud ja kontrollid ülegenoomsete assotsiatsiooniuuringute läbiviimiseks vastavalt etteantud fenotüübi definitsioonidele.

Magistritöö valmis Tartu Ülikooli Genoomika instituudis, genoomse epidemioloogia töögrupis.

1. KIRJANDUSE ÜLEVAADE

1.1 Komplekshaigused

Kahe eri inimese genoomid erinevad teineteisest umbes 0,1% ning see erinevus tuleneb inimeste DNA ligikaudu 3 miljoni aluspaari variatsioonidest (*Diversity in Genomic Research*, s.a.). Need variatsioonid mängivad rolli nii meie välimuses, nagu näiteks silmavärv, kui ka erinevate haiguste avaldumises. Suurem osa erinevustest on tingitud üksikutest muutustest DNA aluspaaris ehk üksiknukleotiidi polümorfismidest (SNP). Esineb ka teisi polümorfisme, näiteks koopiaarvu muutused, deletsioonid, insertsioonid, duplikatsioonid, kuid harvemini kui ühenukleotiidilisi muutusi. Haiguseid, mis on tingitud kindlast muutusest (enamasti ühes) geenis, nimetatakse monogeenseteks haigusteks ning need on populatsioonis harvaesinevad. Komplekshaigused seevastu on populatsioonis sagedased ja nende kujunemisel mängivad rolli nii geneetilised variatsioonid kui keskkonnatingimused ja elustiil (Spataro et al., 2017). Levinud haigus/levinud variatsioon (*common disease/common variant*) hüpoteesi kohaselt põhjustavad komplekshaiguseid populatsioonis levinud geneetilised variatsioonid, millel on väike efektisuurus ehk väike mõju tunnuse varieeruvusele. Nii monogeensete haiguste kui komplekshaiguste päritavuse uurimiseks saab kasutada kaksikute ja perekondade uuringuid, komplekshaiguste haigusseoseliste variantide leidmiseks sobivad aga ülegenoomsed assotsiatsiooniuuringud (Friedman et al., 2021).

1.2 Ülegenoomsed assotsiatsiooniuuringud

Ülegenoomsed assotsiatsiooniuuringud ehk GWAS-id (*genome wide association study*) on uuringud, kus otsitakse seoseid inimestel leiduvate geneetiliste variatsioonide ja fenotüüpide vahel. GWAS-ide peamine eesmärk on leida üles genotüübi ja fenotüübi vahelised seosed ning tänu sellele paremini mõista haiguse bioloogiat, et selle tulemusel efektiivsemalt haiguseid ravida või ennetada (Visscher et al., 2017).

GWAS-i käigus binaarsete tunnuste uurimisel võrreldakse kahte gruppi inimesi, kellest üks grupp on kontrollid ja teine grupp on juhud. Juhtudeks valitakse inimesed, kellel on uuritav fenotüüp olemas ning kontrollideks need, kellel vastavat fenotüüpi ei ole ja vajadusel rakendatakse kontrollidele ka lisatingimusi olenevalt sellest, mida uurida tahetakse. Näiteks võib olla oluline kontrollidel teatud sarnaste diagnooside välistamine. Kõige sagedamini uuritav geneetiline variatsioon ülegenoomses assotsiatsiooniuuringus on SNP (Uffelmann et al., 2021). Enamik ülegenoomseid assotsiatsiooniuuringuid teostatakse kasutades SNP

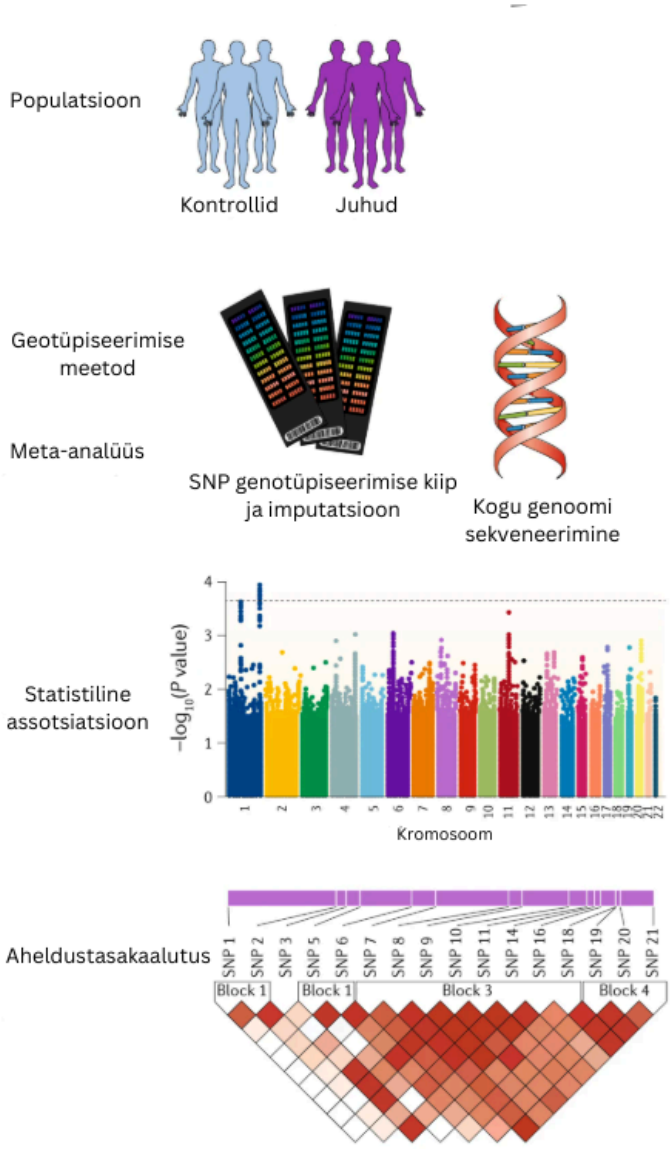
genotüpiseerimise kiipide andmeid (Tam et al., 2019). Genotüpiseerimise kiipidel on markerid, millega saab iga inimese DNA-st määrata tal esinevad SNP-id. Enamasti genotüpiseeritud andmed enne GWASi tegemist imputeeritakse, see tähendab, et andmestikust puuduvad geneetilised variandid ennustatakse olemasolevate markerite ja referentspaneeli põhjal. Referentspaneelis on miljoneid SNP-e paljude inimeste kohta ja seega kui uuringus genotüpiseeriti vähem markereid, siis on võimalik referentspaneeli põhjal puuduvaid SNP-e ennustada. Imputeerimine on vajalik selleks, et erinevates kohortides tehtud uuringute tulemusi optimaalselt kokku kombineerida ning sellega GWASi võimsust tõsta. Seejärel võrreldakse juhtude ja kontrollide geneetilisi variante ning otsitakse statistiliselt olulisi erinevusi kahe grupi SNP alleelisagedustes. Suurem osa geneetilisi variante, mida ülegenoomsete assotsiatsiooniuuringutega on uuritud, on populatsioonis levinud ehk nende harvem esineva alleeli sagedus on üle 1% (Visscher et al., 2017).

Uuringute eesmärk ei ole ainult geneetilise seose leidmine vaid ka leitud seose bioloogilise funktsiooni mõistmine (Mills & Rahal, 2019). Assotsiatsiooniuuringul tuvastatud seosed üksi ei selgita nende bioloogilist mehhanismi, aga on arendatud mitmeid jätkuanalüüse, näiteks sekveneerimisel põhinevad oomika-analüüsid, kolokalisatsiooni hindamine, bioloogiliste radade ning kudede rikastusanalüüsid jne, mis aitavad seose panna bioloogilisse konteksti (Visscher et al., 2017).

Olenemata sellest, et on leitud tuhandeid erinevate tunnustega seotud SNP-sid, jääb mingi osa nende päritavusest siiski seletamata. Seda nähtust nimetatakse leidmata päritavuseks (*missing heritability*) ning arvatakse, et see tuleneb suurest hulgast väga väikese efektisuurusega variatsioonidest, mis panustavad päritavusse, kuid mida ei ole hetkel assotsiatsiooniuuringutel võimalik tuvastada (López-Cortegano & Caballero, 2019).

Ülegenoomsete assotsiatsiooniuuringute tulemuste tõlgendamise teeb keerulisemaks aheldustasakaalutus. Aheldustasakaalutus (*linkage disequilibrium*, LD) on nähtus, kus geneetilised variandid, mis on kromosoomi peal üksteisele lähedal, võivad jääda ka meiootilise rekombinatsiooni tulemusel üksteisele lähedale ja päranduda koos. Seega ei pruugi GWAS-il tuvastatud SNP olla seotud uuritava tunnusega, vaid teise SNP-ga, mis on tegelikult põhjuslikult tunnusega seotud (Cano-Gamez & Trynka, 2020). See nähtus muudab keerukaks hindamise, milline SNP on tegelikult tunnuse suhtes põhjuslik. GWAS-i tõlgendamise teeb keerukamaks ka see, et üle 90% uuringutel leitud geneetilistest variantidest asuvad genoomi mitte-kodeerivas alas ja ei mõjuta otseselt valkude kodeerimist. Arvatakse, et need variandid mõjutavad geenikspressiooni reguleerimist, kuid kuna haigusega seotud lookused sisaldavad tihti mitmeid geene, siis on raske eristada, millised geenid on mõjutatud

(Cano-Gamez & Trynka, 2020). Ülegenoomse assotsiatsioonianalüüsi töövoo ülevaade on toodud joonisel 1.



Joonis 1. Ülegenoomne assotsiatsioonianalüüs (kohandatud Tam et al., 2019). Esimese sammuna valitakse populatsioon, keda uurida (binaarse tunnuse puhul juhtude ja kontrollide grupid). Seejärel teostatakse genotüpiseerimine (kasutades SNP genotüpiseerimise kiipe ja imputeerimist või kogu genoomi sekveneerimist). Uuringu statistilise võimsuse parandamiseks viiakse tihti läbi meta-analüüs. Seejärel leitakse genoomi regioonid, mis on seotud uuritava haigusega. Statistiliselt kõige tugevamini seotud variatsioonid ei pruugi olla põhjuslikud, vaid on aheldustasakaalutuses põhjuslike variantidega, järgnevad analüüsid põhjuslike variatsioonide leidmiseks.

Lisaks binaarsete tunnuste uurimisele, kus kasutatakse lihtsat lähenemist ehk inimesed määratakse juhtudeks selle järgi, kas haigus esineb või mitte, on olemas mudelid ka aja jooksul tekkivate sündmuste analüüsimiseks (*time-to-event models*) (Pedersen et al., 2023).

Need elukestusanalüüsi mudelid arvestavad lisaks haiguse olemasolule ka aega, millal haigus ilmnes. Sellised mudelid võivad aidata paremini tuvastada geneetilisi variante, mis on seotud haiguse avaldumiseni kulunud ajaga (näiteks võib nooremana avaldunud haigus viidata suuremale geneetilisele mõjule) ja uurida haiguse geneetilist tausta (Dey et al., 2022). Lisaks on leitud, et need mudelid suudavad paremini kohanduda kallutatud andmetega – kui valitakse uuringu võimsuse suurendamiseks tahtlikult rohkem haigusjuhte, kui juhuslikus valimis leiduks, võib tavaline GWAS anda ebatäpseid tulemusi, kuid elukestusanalüüsi mudelid kasutavad lisainformatsiooni (vanust, sugu, sünniaastat, populatsiooni haiguslevimust jne), et kohandada hinnanguid, kui tõenäoline on, et konkreetne isik peaks olema uuringusse kaasatud, muutes sellega tulemused usaldusväärsemaks (Pedersen et al., 2023). Mudelite kasutamist piirab aga see, et haiguse avaldumise vanuseline informatsioon ei ole alati kättesaadav ning mudelite kasutamine suurtel andmekogudel nõuab palju arvutusressursse.

Suurema valimi kasutamine ülegenoomsel assotsiatsiooniuuringul parandab uuringu statistilist võimsust ja geneetiliste riskilookuste leidmise tõenäosust (Bouttle et al., 2024). Suurema valimi saamiseks ja seega seoste paremaks leidmiseks tehakse meta-analüüse, kus erinevatest biopankadest saadud analüüsid kombineeritakse kokku.

1.3 Biopankade olulisus

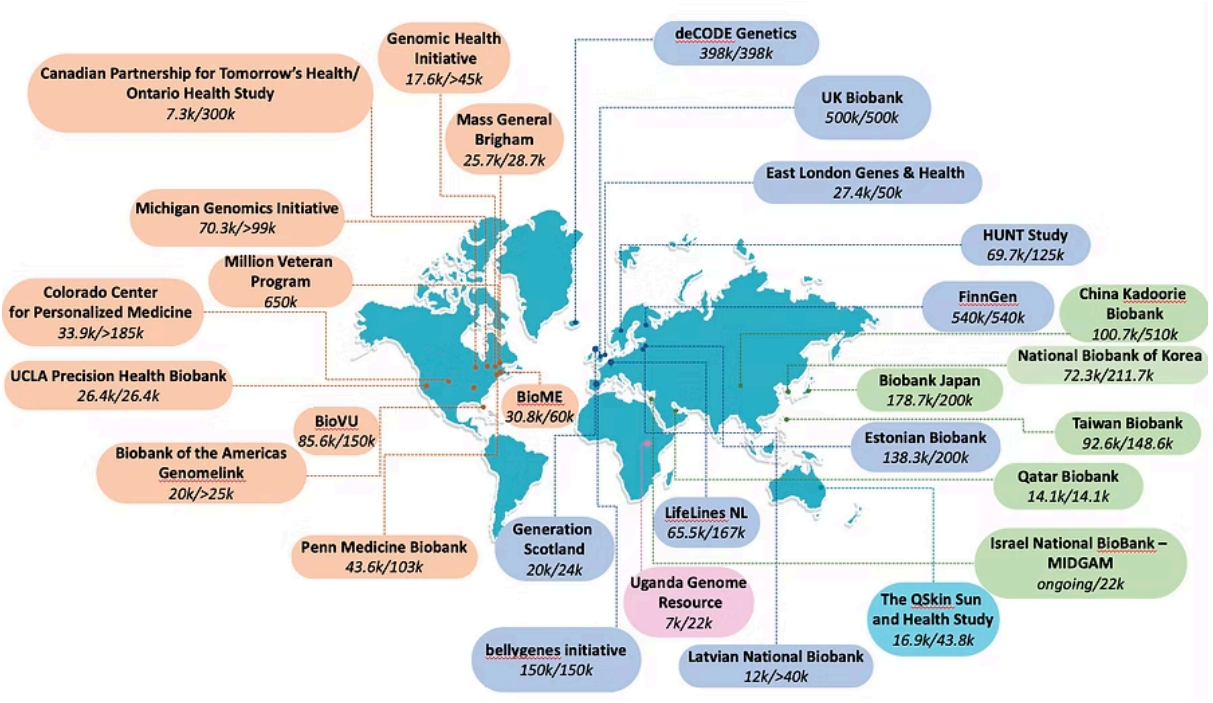
Biopankadeks nimetatakse kollektsiooni bioloogilisest materjalist ja sellega seotud terviseandmetest, mida säilitatakse (meditsiinialasteks) teadusuuringuteks (Annaratone et al., 2021). Biopankadel on oluline roll ülegenoomsete assotsiatsiooniuuringute, geneetilise epidemioloogia ning inimese genoomi struktuuri ja funktsiooni uurimisel (Lazareva et al., 2022). Biopanku on loodud erinevates riikides ning neid on väga erineva suurusega. On leitud, et riikides, kus on suurem võimekus teadustööks, võiks olla 11-30 biopanka miljoni elaniku kohta, millest 2 võiks olla suured biopangad, kus on üle 1000 proovi ning ülejäänud sellest väiksemad (O'Donoghue et al., 2022). Biopankade klassifitseerimiseks on olemas erinevaid süsteeme, aga klassikaliseim neist on jagada biopangad populatsioonipõhisteks biopankadeks ja patsiendipõhisteks biopankadeks (Colwill et al., 2024).

Suuremad ja tuntumad biopangad on Ühendkuningriigi biopank, Jaapani biopank, FinnGen, Eesti Geenivaramu, Hiina Kadoorie biopank, Tohoku Medical Megabank projekt, Taiwani biopank, *Lifelines Cohort Study*, Korea rahvuslik biopank, Karolinska biopank, HUNT biopank, Kanada *Partnership for Tomorrow's Health*, *All Of Us Research Program*, BioVU ja Penn Medicine biopank (Lazareva et al., 2022). Lisaks eelpool väljatoodud biopankadele on

ühed suuremad biopangad ka *Million Veteran Program* (Gaziano et al., 2016) ja deCODE (Gudbjartsson et al., 2015). Nimetatud suuremate biopankade võrdlus on toodud lisas 1.

Biopankadel on oluline roll näiteks ravimite arendamisel, sest need pakuvad teadlastele ligipääsu suurele hulgale bioloogilistele materjalidele ja terviseinfole, mida saab kasutada nii potentsiaalsete ravimi sihtmärkide leidmiseks, ravimi tõhususe ja ohutuse kontrollimiseks kui ka kliiniliste katsetuste efektiivsemaks läbiviimiseks tänu suurele hulgale hästi iseloomustatud bioloogilistele proovidele (Batheja et al., 2023). Biopankadest saadavad andmed on olulised haiguste uurimiseks, näiteks USAs tehtud uurimistööd vähi uurimiseks toetuvad 40% ulatuses biopankade andmetele ning biopangad on olulised tervishoiusüsteemi kõigile aspektidele (Colwill et al., 2024).

Kuna suurema valimiga on võimalik assotsiatsiooniuuringute statistilist võimsust suurendada, siis on tehtud uuringuid mitme biopanga andmeid kombineerides. Biopankade koostöö edendamiseks loodi 2019. aastal *Global Biobank Meta-analysis* initsiatiiv, millega on liitunud 29 biopanka ja projekti ning mille üheks eesmärgiks teiste seas on suurendada GWAS-ide võimsust (*GBMI Home*, 2024). Initsiatiiviga ühinenud osalejad on toodud joonisel 2.



Joonis 2. *Global Biobank Meta-analysis* initsiatiiviga liitunud biopangad ja projektid (*GBMI Home*, 2024). Erinevad värvid tähistavad biopanga geograafilist kuuluvust. Number enne kaldkriipsu näitab biopangaga/projektiga liitunud osalejate hulka ja number pärast kaldkriipsu näitab kavandatavate osalejate hulka.

Global Biobank Meta-analysis initsiatiivi raames teostati 18 biopangast andmeid kaasates seni suurim astma geneetiline assotsiatsioonianalüüs, mille tulemusel leiti mitmeid uusi seoseid ning kinnitati juba varasemalt leitud assotsiatsioone (Tsuo et al., 2022). Südamepuudulikkuse geneetilise riski määramiseks arvutati polügeensed riskiskoorid kasutades samuti initsiatiivi raames andmeid mitmest populatsioonist ning leiti üldisemad riskifaktorid, mis kehtivad laiemalt erinevatele etnilistele rühmadele (Wu et al., 2021). Primaarse avatud nurgaga glaukoomi uurimisel 15 biopanga andmeid kasutades leiti 18 uut lookust (Faro et al., 2022), insuldi uurimisel 16 biopanga andmeid kasutades leiti 2 seni kirjeldamata varianti (Surakka et al., 2022) ja venoosse trombemboolia uurimisel 9 biopanga andmeid kasutades leiti 9 potentsiaalselt uut seost (Wolford et al., 2022). Mitme biopanga andmete kaasamine on kasuks ka haruldaste haiguste uurimisel, mille puhul on patsientide arv ja seega ka uuringute statistiline võimsus väike. *Global Biobank Meta-analysis* initsiatiivi raames uuriti haruldast haigust nimega idiopaatiline kopsufibroos kaasates andmeid 13 biopangast ning leiti 7 uut assotsiatsiooni (Partanen et al., 2021).

Erinevate biopankade andmeid uuringuteks kasutades on võimalikult täpsete tulemuste saamiseks oluline, et fenotüübid oleksid õigesti defineeritud ning et andmeid oleks võimalik omavahel harmoniseerida.

1.4 Fenotüüpide defineerimine

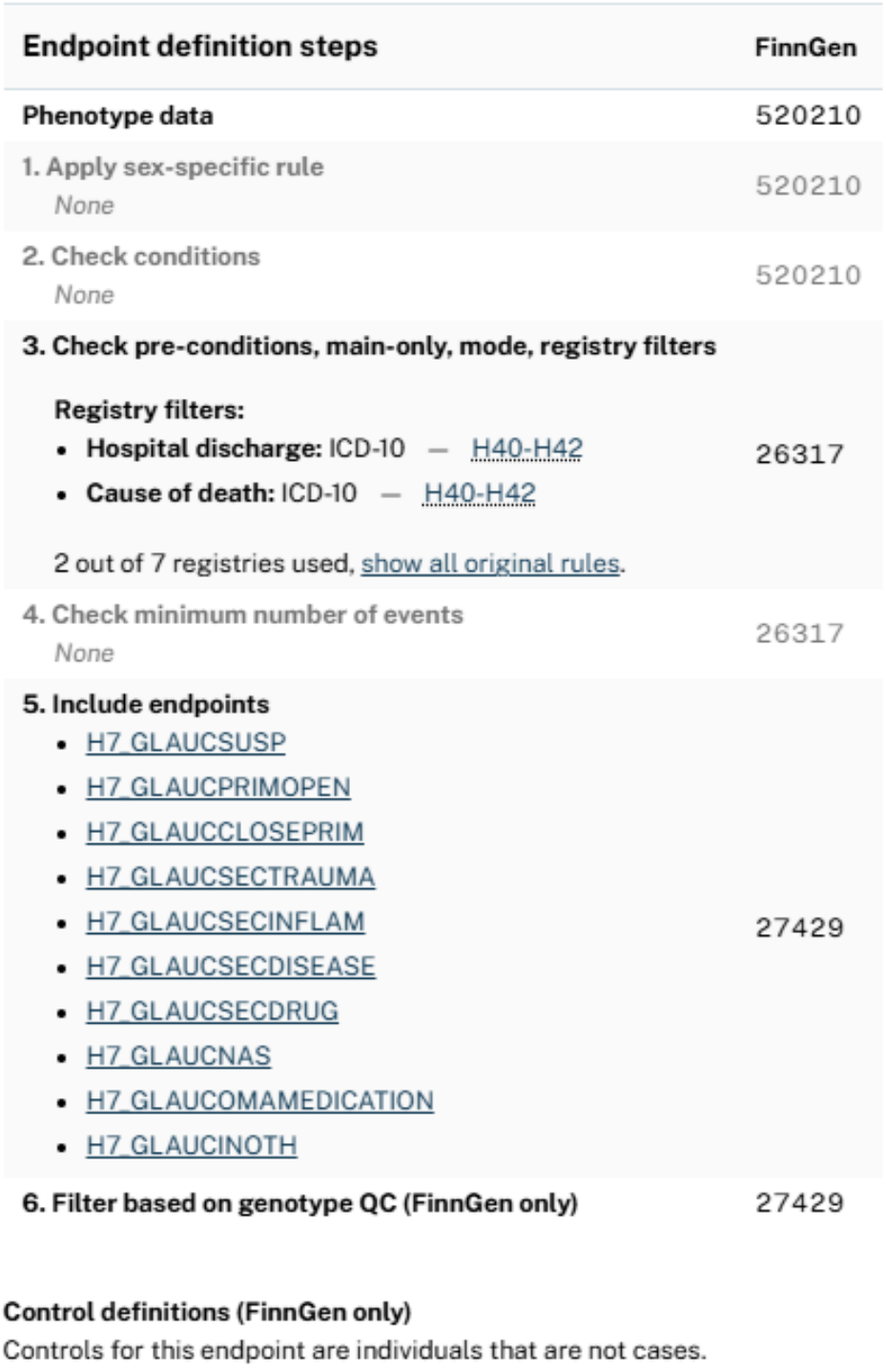
Fenotüübi mõiste biopankade kontekstis tähistab erinevate patsientide gruppide poolt jagatavaid tunnuseid, mis võivad olla nii haiguse või seisundi esinemine kui ka tarvitatavad ravimid, elustiili faktorid või kaasnevad haigused ning nende erinevate tunnuste kombinatsioonid (Chapman et al., 2021). Fenotüüpide defineerimine võimaldab leida seoseid erinevate tervisenäitajate ja geneetiliste markerite vahel. Hiljuti tehtud uuringus defineeriti 122 rasedusaegset fenotüüpi, mis sisaldasid nii vere biomarkereid kui rasedusaegseid tüsistusi ning tuvastati olulisi rasedusaegseid geneetilisi markereid ja ka nende mõju ema pikaajalisele tervisele (Liu et al., 2024).

Erinevates uuringutes on fenotüübid defineeritud erinevalt ning see takistab assotsiatsioonide leidmist meta-analüüsides (Bennett et al., 2011). Fenotüüpide andmed on heterogeensemad ja raskemini kasutatavamad kui genotüüpide andmed, sest need ei ole nii hästi standardiseeritud ning on tihedalt seotud keskkonnatingimustega ja ajas muutuvad (Deans et al., 2015). Kvaliteetne fenotüüpide defineerimine on oluline, et valida välja õiged isikute grupid uuringuteks ning tagada tulemuste valiidsus ja korratavus (Chapman et al., 2021).

Fenotüübi defineerimist ja valideerimist nimetatakse fenotüpiseerimiseks ning selle tulemusel defineeritud fenotüüpide jaoks on loodud fenotüüpide raamatukogud, kuhu fenotüüpide definitsioonid saab üles laadida, säilitada ja kasutajate poolt alla laadida (Chapman et al., 2021). Tuntumad ja suuremad fenotüüpide raamatukogud on UK fenotüüpide raamatukogu, mis käivitati 2021. aastal ning milles on defineeritud 1049 fenotüüpi ja FinnGen projekti käigus loodud fenotüüpide raamatukogu (Thayer et al., 2024).

1.5 FinnGen

FinnGen on Soome ülikoolide, haiglate ja biopankade koostööprojekt, mis käivitati 2017. aastal kaasates 500 000 osalejat (Kurki et al., 2023; Lazareva et al., 2022). FinnGen projekti kohordis on umbes 217 000 osalejatest mehed ja umbes 281 000 naised ning kohordi mediaanvanus on 53 aastat. Umbes 315 000 osalejat on saadud kas üldarstide või erialaarstide vastuvõttudelt, 58 000 osalejat on terved veredoonorid, 2000 osalejat on hematoloogilised vähihaiged ja 180 000 osalejat on saadud üleriigilistest populatsiooni- ja haigusepõhistest uuringutest. Projektis kogutakse osalejate andmeid erinevatest elektroonilistest tervise andmebaasidest, mis sisaldavad informatsiooni hospitaliseerimise, retseptiravimite ostu, protseduuride ja surma põhjuste kohta (Kurki et al., 2023). FinnGen projekti käigus on loodud fenotüüpide definitsioonid, mida pidevalt täiendatakse ja parandatakse, ning tööriist *Risteys*, millega on võimalik FinnGen andmeid vaadata fenotüüpide tasemel, sealhulgas näha fenotüüpide definitsioone ja erinevaid statistilisi näitajaid, näiteks isikute arvu ja soolist jaotust (*Risteys*, 2024). Näide *Risteys* tööriistaga glaukoomi fenotüübi definitsiooni järgi juhtude ja kontrollide leidmisest on toodud joonisel 3.



Joonis 3. *Risteys* tööriista ekraanitõmmis glaukoomi fenotüübile juhtude ja kontrollide leidmise sammudest (*Risteys*, 2024)**.** Glaukoomi fenotüübil ei ole defineeritud sugu, tingimusi ega eeltingimusi. Defineeritud on RHK-10 ehk rahvusvahelise haiguste klassifikatsiooni 10. versiooni koodid H40–H42 nii surmapõhjusena kui haiglast pärit koodidena. Samuti kaasatakse juhud teistest glaukoomiga seotud fenotüüpide juhtudest. Kontrollideks arvatakse kõik inimesed, kes juhtude alla ei kuulu. Igal sammul on näha, mitu inimest on juhtude seas pärast filtreerimist alles. Kui ülejäänud sammud filtreerivad inimesi juhtude seast välja ehk juhte jääb vähemaks, siis "*Include endpoints*" sammuga leitakse juhte teiste seotud fenotüüpide juhtude seast juurde.

FinnGen fenotüüpide raamatukogus on defineeritud 4648 fenotüüpi, mida pidevalt vaadatakse üle, parandatakse ja täiendatakse. Iga fenotüüp on defineeritud kindlate reeglite järgi, mis on kirjas *Exceli* failis, kus igal real on defineerimise reeglid ühe fenotüübi kohta. Ekraanitõmmis FinnGen fenotüüpide definitsioonide failist on toodud lisas 2. Uus fenotüüpide definitsioonide fail avaldatakse korra aastas. Esimene fenotüüpide definitsioonide fail avaldati 2018. aasta jaanuaris ning see sisaldas 2057 fenotüübi definitsiooni, mis kattis RHK-10 peatükid 1-14. Esimeses versioonis olid olemas "*Include*" ja "*Condition*" reeglid ja võimalus luua soospetsiifilisi fenotüüpe, samuti olid olemas kontrollidest välistamise reeglid (reeglitest on pikemalt kirjutatud peatükis 2.2.2.2). Alates esimese fenotüüpide faili avaldamisest on mitmed erinevad eksperdid pidevalt tegelenud uute fenotüüpide defineerimisega, olemasolevate fenotüüpide definitsioonide parandamisega, skriptide uuendamisega ja kvaliteedikontrolliga. Uusi fenotüüpide definitsioonide faile avaldatakse kuni aastani 2027 (*FinnGen*, 2025).

1.6 Fenotüüpide harmoniseerimine

Kuna ülegenoomsetel assotsiatsioonianalüüsidel võib suurem valim võimaldada leida rohkem riskilookuseid ja parandada uuringu statistilist võimsust, siis hea võimalus valimi suurendamiseks on kasutada mitme biopanga andmeid. Selleks, et erinevate biopankade andmeid kasutada, on need vaja eelnevalt harmoniseerida. Andmete harmoniseerimine tähendab erinevate andmete ühendamist ühtsetesse või võrreldavatesse vormingutesse (Bennett et al., 2011). Sellepärast on tähtis teada, milliseid kodeerimise süsteeme eri biopankades kasutatakse ja kuidas neid andmeid assotsiatsioonianalüüsil kasutada.

Üks laialdaselt kasutusel olev klassifitseerimise süsteem on rahvusvaheline haiguste klassifikatsioon ehk RHK (*International Classification of Diseases, ICD*). RHK esimene versioon loodi 1893. aastal ning seda on pidevalt muudetud ja täiendatud (Hirsch et al., 2016). 1948. aastal võttis klassifitseerimissüsteemi uuendamise üle Maailma Terviseorganisatsioon ning loodi uus laiendatud versioon RHK-6, millele järgnesid RHK-7 1958. aastal, RHK-8 1968. aastal, RHK-9 1979. aastal, RHK-10 1993. aastal ja RHK-11 2022. aastal (*International Classification of Diseases (ICD)*, 2025).

Eestis ja ka teiste riikide biopankades on enamasti kasutusel RHK-10 (biopankades kasutatavad klassifitseerimissüsteemid on kirjeldatud järgnevates lõikudes). Selle kood koosneb ühest tähest ja kahest numbrist, millele võib täpsustusena järgneda pärast punkti veel numbreid. RHK 10. versioonis on haigused jagatud 21 peatüki vahel, kus igale peatükile on koodina määratud kindel tähtede ja numbrite vahemik (*RHK ehk rahvusvaheline haiguste*

klassifikatsioon | Tervise Arengu Instituut, 2025). RHK-11 ei ole Eestis veel kasutusse võetud, nagu ka paljudes teistes riikides, sest see vajab enne kasutusele võtmist tõlkimist ja muid ettevalmistustöid. Maailma Terviseorganisatsiooni andmetel on 2024. aasta maikuu seisuga 132 riiki RHK-11 klassifitseerimissüsteemi implementeerimise erinevates etappides (*International Classification of Diseases (ICD)*, 2025). RHK-11 koodid on teistsuguse vorminguga jäädes vahemikku 1A00.00-ZZ9Z.ZZ ning haigused on jagatud 28 peatüki vahel (Fung et al., 2020).

Erinevates biopankades on kasutusel erinevad klassifitseerimissüsteemid. Näiteks Ühendkuningriigi biopangas on kasutusel RHK-10, RHK-9, OPCS-3 ja OPCS-4 koodid. OPCS (*Office of Population, Censuses and Surveys: Classification of Interventions and Procedures*) koodidega tähistatakse operatsioone ja protseduure (*UK Biobank*, 2025).

FinnGen projektis on kasutusel RHK-10, RHK-9, RHK-8 koodid, vähkide klassifitseerimiseks RHK-O-3 koodid, ravimite tähistamiseks ATC ja KELA koodid ja protseduuride jaoks NOMESCO NCSP koodid ja Soome-spetsiifilised protseduuride koodid (*Endpoints*, 2024). RHK-O-3 ehk kasvajate rahvusvahelise haiguste klassifikatsiooni 3. versiooni koodidega klassifitseeritakse kasvajaid topograafia ja morfoloogia põhjal. Ravimite klassifitseerimiseks kasutatavad ATC koodid ehk anatoomilis-terapeutilised keemilised koodid on ravimite klassifitseerimise süsteem, mille järgi ravimid jaotatakse erinevatesse farmakoloogilistesse gruppidesse ning KELA koodid on Soome-spetsiifilised Soome sotsiaalkindlustusameti koodid ravimite klassifitseerimiseks. Protseduuride klassifitseerimiseks kasutatav NOMESCO NCSP klassifitseerimissüsteem on Põhjamaade ühtne klassifikatsioon kirurgilistele protseduuridele.

Eesti Geenivaramus on kasutusel RHK-10 koodid ja ATC-koodid (Leitsalu et al., 2015). Kirurgiliste protseduuride klassifitseerimiseks kasutatakse NOMESCO NCSP klassifitseerimissüsteemi (suuline konsultatsioon Eesti Geenivaramu andmeanalüütiku Diana Rogozina´ga, 08.04.2025).

Lisaks nimetatud koodidele on olemas veel mitmeid erinevaid kodeerimise süsteeme, mis on erinevate riikide biopankades kasutusel. Fenotüüpide koodid (*Phecodes*) on hierarhiline fenotüüpide klassifikatsiooni süsteem, mis loodi ühendades RHK-9 ja seejärel RHK-10 koodid elektroonilise terviseinfosüsteemiga (Denaxas et al., 2022). Fenotüüpide koodid on loodud, et siduda genoomi andmed fenotüübi andmetega ja lihtsustada suuremahulist andmeanalüüsi. Hetkel on defineeritud üle 1800 fenotüübi koodi (Linder et al., 2021). Ühendkuningriigi biopangas on kasutusel haiglaravil viibivatele patsientidele eelnevalt

nimetatud RHK-kodeerimise süsteem, aga perearstide koodidena oli varasemalt kasutusel standardset sõnastikku pakkuvad Read2/CTV3 süsteemid, mis on nüüdseks kodeeritud ümber *Phecodes* süsteemi (Denaxas et al., 2022).

Erinevate riikide biopankades on kasutusel lisaks veel erinevaid terminoloogiasüsteeme. Üheks ulatuslikuks süsteemiks on SNOMED CT (*Systematized Nomenclature of Medicine – Clinical Terms*). See on elektroonilistes haiguslugudes kasutamiseks mõeldud struktrueeritud kliiniline sõnastik, mis tagab standardiseeritud terminid haiguste, protseduuride, ravimite ja muude tervisega seotud mõistete jaoks (*SNOMED CT*, 2025). SNOMED CT süsteem on kasutusel alates 2018. aastast Ühendkuningriigi biopangas (*UK Biobank*, 2025).

2. EKSPERIMENTAALOSA

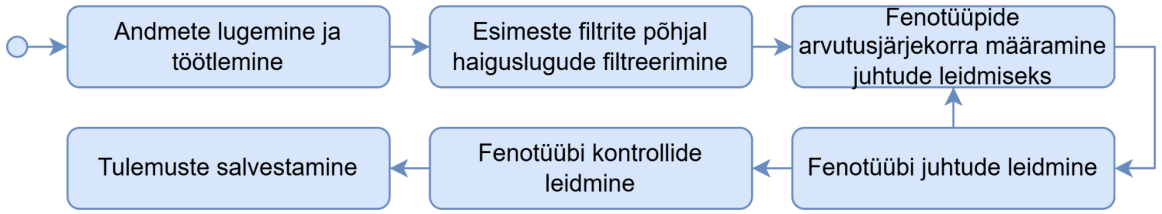
2.1 Töö eesmärgid

Magistritöö eesmärgiks oli luua automatiseeritud töövoog, mille abil saab määrata Eesti Geenivaramu geenidoonorite andmestikust ülegenoomsete assotsiatsioonianalüüside läbiviimiseks juhud ja kontrollid vastavalt etteantud fenotüübi definitsioonidele.

2.2 Materjal ja metoodika

2.2.1 Valminud töövoo ülevaade

Magistritöö tulemusel valmis programmeerimiskeeles *Python* kirjutatud programm, mis leiab iga soovitud fenotüübi definitsiooni kohta ülegenoomseteks assotsiatsioonianalüüsideks sobivad isikud juhtude ja kontrollide gruppi. Selle saavutamiseks kasutatakse informatsiooni, mis on saadud programmile sisendiks antud isikute andmeid sisaldavates failides ja fenotüüpe defineerivas failis leiduvat informatsiooni töödeldes. Programmi töövoo ülevaade on toodud joonisel 4.



Joonis 4. Programmi töövoo ülevaade. Programm loeb sisendiks antud andmefailid ja töötleb neis sisalduvat informatsiooni; filtreerib haiguslood sootunnuste, minimaalsete haigusjuhtude arvude, registrite ja arsti liikide kombinatsioonide järgi; leiab fenotüüpide eeldused ja määrab fenotüüpide arvutusjärjestuse; leiab fenotüübile juhud ning valib järgmise fenotüübi arvutamiseks; leiab fenotüüpidele kontrollid; salvestab tulemused.

Programm loodi programmeerimiskeeles *Python* versioon 3.11.4 ning koodi loomise lihtsustamiseks kasutati integreeritud programmeerimiskeskkonda *PyCharm*, mis on ettevõtte *JetBrains* poolt arendatud rakendus *Pythoni* koodi efektiivseks kirjutamiseks. Loodud programm on talletatud *GitHub* privaatses repositooriumis, mis pärast magistritöö kaitsmist tehakse kättesaadavaks Genoomika Instituudi siseseks kasutamiseks. *GitHub* on versioonihaldustarkvara *Git* kasutav veebimajutusteenus, mis võimaldab salvestada koodi muudatuste ajalugu ja koodile ligi pääseda erinevatest arvutitest (*GitHub*, 2025).

Valminud programm koosneb mitmest *Pythoni* failist, mille nimed ja lühikirjeldused on toodud tabelis 1. Loetavuse parandamiseks on erinevatesse failidesse grupeeritud sarnase eesmärgiga meetodid ja funktsioonid, mis aitavad saavutada programmi väljundit.

Tabel 1. Programmi failide lühikirjeldused.

Faili nimi	Lühikirjeldus
main.py	Programmi sisendpunkt: käivitab programmi tööks vajalikud funktsioonid logimise seadistamiseks, andmete lugemiseks andmefailidest, juhtude ja kontrollide leidmiseks ning tulemuse kirjutamiseks.
doonorite_andmete_lugeja.py	Loeb geenidoonorite andmed andmefailist *Pandase* andmeraamistikku.
ravimite_lugeja.py	Loeb ravimite andmed andmefailist, lisab isikutele määratud ravimite ATC-koodid ja ravimite määramiste kuupäevad doonorite andmeraamistikku, iga ravimite faili rea kohta lisatakse üks rida.
endpointide_lugeja.py	Loeb fenotüüpide definitsioonide andmed *Exceli* vormingus andmefailist *Pandase* andmeraamistikku ning loob RHK-10 koodide ja ATC-koodide põhjal regulaaravaldised.
regexi_koostaja.py	Sisaldab meetodeid RHK-10 ja ATC-koodide põhjal regulaaravaldiste loomiseks.
juhtude_ja_kontrollide_leidja.py	Kasutades eelnevalt moodustatud fenotüüpide definitsioonide andmeraamistikku, leiab eeldused igale otsitavale fenotüübile juhtude ja kontrollide leidmiseks, koostab sõnastiku esimeste filtreerimise sammude kombinatsioonide esimestest haigusjuhtudest, paneb paika fenotüüpide arvutamise järjekorra ja käivitab fenotüübi täitmata eelduste puudumisel juhtude ja kontrollide arvutamise.
juhtude_ja_kontrollide_arvutaja.py	Leiab etteantud fenotüübile doonorite seast juhud ja kontrollid vastavalt fenotüübi defineerimisel määratud reeglitele.
juhud_ja_kontrollid.py	Andmeklass, mis hoiab informatsiooni juhtude, kontrollide ning nendega seotud kuupäevade kohta. Lihtsustab kogu andmekomplekti korraga jagamist erinevate meetodite vahel.
maatriksi_faili_kirjutaja.py	Moodustab leitud juhtude ja kontrollide põhjal tulemuste maatriksi ning kirjutab selle faili.
tulemuste_kirjutaja.py	Programmi töö ajal kogub statistikat programmi töö kohta ning koostab töö lõpus faili, kus on välja toodud programmi töö tulemused, vead ja statistika.
abifunktsioonid.py	Abifunktsioonid, mida kasutavad erinevates programmi failides olevad teised funktsioonid.

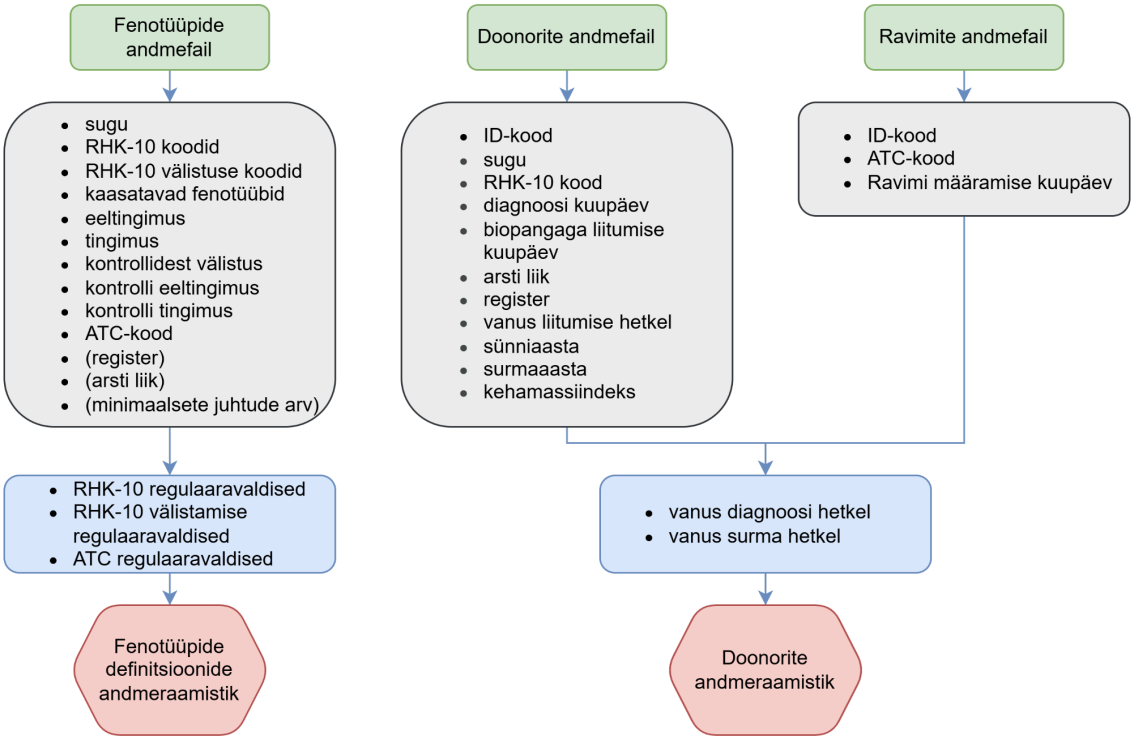
arvutamise_erind.py	Juhtude ja kontrollide arvutamisel tekkivate vigade jaoks defineeritud erind (ingl k *exception*), et eristada need *Python*i enda erinditest.
taimer.py	Sisaldab meetodeid programmi erinevate etappide ajakulu mugavaks mõõtmiseks ja selle logimiseks.
doonorite_genereerija.py	Võimaldab genereerida testandmetega doonorite faile määrates ära parameetrid nagu näiteks doonorite arv. Ei ole vajalik programmi tööks, aga lihtsustab programmi testimist.
test_juhtude_ja_kontrollide_leidmine.py	Kontrollib juhtude ja kontrollide leidmise õigsust ühiktestidega ehk automatiseeritud testidega, millega saab rakenduse loogikat isoleeritult testida.
test_regexi_loomine.py	Kontrollib regulaaravaldiste loomise õigsust ühiktestidega.
test_filtreerimise_sammud.py	Kontrollib iga filtreerimise sammu õigsust ühiktestidega.
testide_abifunktsioonid.py	Abifunktsioonid, mis lihtsustavad erinevate testide jaoks testandmete loomist.
logi_konfiguratsioon.json	Logimise konfiguratsioon.
logija.py	Kasutades eeldefineeritud konfiguratsioonifaili seadistab programmi logimise.
requirements.txt	Pythoni paketihalduri *pip* abifail, mille abil saab failis kirjeldatud pakette ühe käsuga paigaldada.
README.md	Infofail, mis sisaldab samm-sammulist juhendit programmi käivitamiseks ning programmi käivitamiseks vajalikke eeldusi ja selgitusi.

Programm koosneb juhtude ja kontrollide leidmiseks vajalikest failidest, millega loetakse sisse andmed, arvutatakse tulemus ja kirjutatakse see tulemusfailidesse. Lisaks on failid logimise konfigureerimiseks. Programmi üksikute osade testimiseks on loodud ka ühiktestid ning kogu töö testimiseks fail testandmete genereerimiseks. Juhend ja pakette kirjeldav abifail aitavad kasutajal rakendust mugavalt paigaldada ja kasutada.

2.2.2 Andmefailide lugemine

Programmi tööks on vajalik kasutaja poolt sisendiks antud andmefailide lugemine ja neis sisalduva informatsiooni töötlemine. Kasutaja poolt antakse programmile sisendiks kaks või kolm faili: geenidoonorite andmetega fail ja fenotüüpide definitsioonide fail on kohustuslikud, ravimite andmeid sisaldav faili võib programmile sisendiks anda, kuid

programmi saab käivitada ka selleta. Programmi poolt tehtav töötlus nimetatud andmefailidega on täpsemalt kirjeldatud kahes järgnevas alapeatükis. Andmete lugemise ja töötluse tulemusena moodustatakse andmeraamistikud, mida programmi edasistes etappides kasutatakse. Programmi töötlus sisendiks antud failidest kuni edasiseks tööks vajalike andmetega andmeraamistike moodustamiseni on toodud joonisel 5.



Joonis 5. Töövoo ülevaade alates programmile sisendiks antud failidest kuni andmeraamistike moodustamiseni. Rohelisega on märgitud vajalikud andmefailid ja halliga neis sisalduv informatsioon, mida on kasutatud edasiseks töötluseks. Sulgudes olev informatsioon ei ole sisendfailis kohustuslik. Sinisega on märgitud väljad, mis on arvutatud olemasolevate andmete põhjal ning punasega saadud lõpptulemus ehk kaks andmeraamistikku, mida kogu edasises töötluses kasutatakse.

2.2.2.1 Geenidoonorite andmete lugemine

Programmi loomisel kasutati magistritöö juhendaja poolt koostatud näidisfaile, mis ei sisaldanud geenidoonorite isikustatud andmeid, kuid mille struktuur vastab tegelikele andmevormingutele. Programmi mugavaks testimiseks tehti töö autori poolt näidisfailide eeskujul doonorite haiguslugude ja ravimite failide genereerija. Selles saab määrata isikute arvu ning keskmise haigusjuhtude ja ravimite arvu ning nende põhjal genereerida juhuslike andmetega testfailid. Programmi väljundi saavutamiseks vajalikud väljad biopanga andmetest ning nende kirjeldused ja vorming on toodud tabelis 2.

Tabel 2. Programmi poolt kasutatavad biopanga andmed.

Veeru nimi	Kirjeldus	Vorming
person.skood	Isiku identifitseerimiskood.	5 tähte ja 4 numbrit. Näiteks ABCDE1234.
person.gender.code	Isiku soo tunnus.	1 = mees; 2 = naine.
diagnosis_assembled.icd10	Isikul diagnoositud seisundi RHK-10 kood.	Rahvusvaheliselt tunnustatud vorming. Näiteks A12.3.
diagnosis_assembled.diagnosis_date	Isikul diagnoositud seisundi diagnoosi kuupäev.	aaaa-kk-pp.
person.first_agreement_date	Isiku biopangaga liitumise kuupäev.	aaaa-kk-pp.
diagnosis_assembled.speciality	Diagnoosi sisestanud arsti liigi tunnus.	Tekstiline ametinimetus. Filtreerimiseks tuleks see kodeerida sobivateks arstide liikideks (programmi on testitud kasutades kodeeringut 1 = perearst, 2 = eriarst).
source.name	Registri nimi, kust diagnoosi andmed pärinevad.	Vabatekst. Näiteks “Surma põhjuste register” või “Causes of Death Registry”.
person.age_at_agreement	Isiku vanus biopangaga liitumise hetkel.	Täisarv. Näiteks “30”.
person.birthYear	Isiku sünniaasta.	Vorming aaaa või imputeeritud kuupäevaga vorming aaaa-kk-pp või aaaa-kk.
person.deathYear	Isiku surma-aasta.	Vorming aaaa või imputeeritud kuupäevaga vorming aaaa-kk-pp või aaaa-kk.
objectiveInformation.bmi	Isiku kehamassiindeks.	Ujuvkomaarv, näiteks “20.5”.
AssembledMedications.atc.code	Isikule määratud ravimi ATC-kood.	Rahvusvaheliselt tunnustatud vorming. Näiteks A12BC34.
MedicationAssembled.atcDate	Isikule ravimi määramise kuupäev.	aaaa-kk-pp.

Isikute andmete faili lugemiseks kasutati *Pythoni* teeki *Pandas*, mis võimaldab andmeid andmeanalüüsiks kiiresti ja mugavalt lugeda, töödelda, sorteerida ja filtreerida (*pandas documentation*, 2024). Andmefaili konverteerimisel andmeraamistikuks määrati ära failist kasutatavate veergude nimed ning failis andmeid eraldavaks sümboliks tabulaator. Edasise töötluse lihtsustamiseks kasutati faili lugemisel ka parameetrit *parse_dates*, mille väärtuseks määrati järjend diagnoosi kuupäeva ja biopangaga liitumise kuupäeva veergude nimedega. See võimaldas faili konverteerimisel andmeraamistikuks eelmainitud veerud *Pythoni* jaoks sobivaks kuupäeva tüübiks muuta. Teiste kasutatavate veergude väärtustele määrati tüüp kasutades parameetrit *dtype*. Soo ja arsti liigi tähistus muudeti sõne tüübiks (kuigi sugu on kodeeritud arvulise väärtusega ning arsti liik võib samuti olla kodeeritud arvulise väärtusena), sest nendega tehakse programmi töö käigus ainult võrdlusi, mitte arvutusi.

Ravimite kohta informatsiooni saamiseks loetakse sisse ka teine biopanga fail, kui see on kasutaja poolt programmile sisendiks antud. Ravimite faili puhul olid samuti teada veergude nimed ja väärtuste vormingud. Failis on sarnaselt eelnevalt kirjeldatud doonorite andmete failile doonorite identifitseerimiskoodid. Lisaks kasutati sellest isikule määratud ravimite ATC-koode ja nende määramise kuupäevi. Ravimite informatsioon lisatakse doonorite andmeraamistikule – ATC-koodid lisatakse eraldi uude veergu ning ravimi määramise kuupäev lisatakse olemasolevasse diagnoosi kuupäeva veergu, et hilisem töötlus kuupäevade põhjal saaks toimuda korraga nii haiguslugude kui ravimite jaoks. Ülejäänud isikuga seotud andmed täidetakse doonorite andmeraamistiku põhjal isiku identifitseerimiskoodi kaudu.

Osade fenotüüpide definitsioonide puhul on juhtude ja kontrollide välja arvutamiseks vajalik teada ka isiku vanust diagnoosi hetkel ja kui isik on surnud, siis ka isiku vanust surma hetkel. Inimese vanuse arvutamiseks diagnoosi hetkel kasutatakse valemit:

$$vanus\ diagnoosi\ hetkel\ = \frac{diagnoosi\ kuupäev - biopangaga\ liitumise\ kuupäev}{365.25} + vanus\ biopangaga\ liitumise\ hetkel$$

Kõik valemis võrdusmärgist paremal pool olevad väärtused on doonorite andmeraamistikus olemas. Tulemus ümardatakse *round()* funktsiooni kasutades ja muudetakse täisarvu tüübiks, et oleks võimalik vanust võrrelda fenotüüpide defineerimise failis toodud piiranguga vanusele diagnoosi hetkel.

Selleks, et leida vanus surma hetkel, on vaja kasutada doonori sünni- ja surmakuupäeva. Kuna biopanga andmetes mõningal juhul on inimese sünni- ja/või surmakuupäev märgitud ainult aasta või kuu täpsusega, siis võttes eeskujuks osaliste kuupäevade imputatsiooni juhiseid biostatistika alases blogipostituses punktis 2.9.4 (Deng, 2021) määrati neil juhtudel

puuduvaks kuupäevaks 30. juuni ehk umbes keskmine kuupäev aastas ning kui kuupäev on märgitud kuu täpsusega, siis määrati päevaks 15. kuupäev. Seejärel tehti vajalikud arvutused tegelike või oletuslike sünni- ja surmaaegadega. Inimese vanuse arvutamiseks surma hetkel kasutatakse valemit:

$$vanus\ surma\ hetkel\ = \frac{oletuslik\ surmaaeg - oletuslik\ sünniaeg}{365.25}$$

Tulemus ümardatakse ja muudetakse täisarvu tüübiks. Saadud isikute vanused diagnoosi hetkel ja surma hetkel lisatakse uute tulpadena varasemalt loodud andmeraamistikku, et neid oleks võimalik hilisemas töötluses kasutada.

2.2.2.2 Fenotüüpide definitsioonide lugemine

Fenotüüpide definitsioonide eeskujuna kasutati Soome biopanga projekti FinnGen fenotüüpide defineerimise hiljutiseimat faili – magistritöö valmimise hetkel oli selleks fail "*FinnGen core and non-core endpoints and controls DF12"* (*Clinical Endpoints | FinnGen*, 2025). Andmete lugemiseks kasutati sarnaselt doonorite andmete lugemisega *Pandas* teeki, milles on olemas *Exceli* faili andmeraamistikuks konverteeriv funktsioon. Failis on fenotüüpide kohta andmeid rohkem kui oli programmi töö jaoks vajalik, seega on programmis täpsustatud veerud, mida andmeraamistiku koostamisel kasutada. Vältimaks programmi skoobi liiga suureks minemist jäeti välja operatsioonikoodide ja vähispetsiifiliste RHK-O-3 koodide veerud. Samuti ei töötle programm RHK-8 ja RHK-9 koode ja KELA ehk Soome sotsiaalkindlustusameti ravimite maksumuse hüvitamise koode, sest need ei ole Eesti Geenivaramus kasutusel. Programmi jaoks vajalik informatsioon on fenotüübi nimi, soo tähistus, RHK-10 koodid, RHK-10 juhtudest välistamise koodid, fenotüübiga liidetavad teiste fenotüüpide nimed, eeltingimused ja tingimused nii juhtude kui kontrollide leidmiseks ning ravimite ATC-koodid. Lisaks FinnGen definitsioonide failis olevatele veergudele lisati magistritöö käigus loodud programmile võimekus filtreerida juhte registri, arsti liigi ja minimaalse diagnooside arvu põhjal. Vastavad veerud saab lisada fenotüüpide definitsioonide andmefaili ja sobivalt täita ning kui neis on väärtus, siis oskab programm seda arvesse võtta. Programmi tööks kasutatavate veergude nimed ja nende lühikirjeldused on toodud tabelis 3.

Tabel 3. Fenotüüpide definitsioonide faili veergude lühikirjeldused.

Veeru nimi	Lühikirjeldus
NAME	Fenotüübile määratud nimi. Ainuke väli, mille väärtus on kohustuslik.
SEX	Kui lahter on tühi, sobib selle fenotüübi juhtude ja kontrollide gruppi mõlemast soost inimene, kui lahtris on 1, sobib mees, kui lahtris on 2, sobib naine.
COD_ICD_10	Surmapõhjuse RHK-10 kood. Juhtude leidmise ühe sammuna leitakse isikud, kellel on sellel väljal olevale RHK-10 koodile vastavaid haiguslugusid, mis on pärit surmaregistrist (doonorite failis registrina märgitud kas "Surma põhjuste register" või "*Causes of Death Registry"*).
HD_ICD_10	Haiglast pärit RHK-10 kood. Juhtude leidmise ühe sammuna leitakse isikud, kellel on sellel väljal olevale RHK-10 koodile vastavaid haiguslugusid, mis on pärit kõigist teistest registritest peale surmaregistri.
COD_ICD_10_EXCL	Surmapõhjuse RHK-10 kood, mida juhtude gruppi kuuluval isikul ei tohi olla diagnoositud surmaregistrist pärit diagnoosina.
HD_ICD_10_EXCL	Haiglast pärit RHK-10 kood, mida juhtude gruppi kuuluval isikul ei tohi olla diagnoositud kõigist teistest registritest peale surmaregistri pärit diagnoosina.
INCLUDE	Püstkriipsuga eraldatud nimekiri fenotüüpidest, mille juhtude gruppi kuuluvad isikud tuleb kaasata ka selle fenotüübi juhtude gruppi.
PRE_CONDITIONS	Eeltingimus, mis juhtude gruppi kuuluval isikul peab olema täidetud. Programm toetab eeltingimustena isiku vanust diagnoosi saamise hetkel ja diagnoosi saamise aastat.
CONDITIONS	Tingimus, mis juhtude gruppi kuuluval isikul peab olema täidetud. Siin väljal võib olla kombinatsioon fenotüüpide nimedest, mis kas kõik peavad juhul esinema/ühtegi ei tohi esineda või mõni peab esinema/ei tohi esineda. Lisaks võib sellel väljal olla toodud piirang inimese kehamassiindeksile.
CONTROL_ EXCLUDE	Fenotüübid, mida kontrollgruppi kuuluval inimesel ei tohi esineda.
CONTROL_ PRECONDITIONS	Eeltingimus, mis kontrollgruppi kuuluval isikul peab olema täidetud. Programm toetab kontrolli eeltingimusena inimese vanust surma hetkel.
CONTROL_ CONDITIONS	Tingimus, mis kontrollgruppi kuuluval isikul peab olema täidetud. Siin väljal võib olla kombinatsioon fenotüüpide nimedest, mis kas kõik peavad kontrollil esinema/ühtegi ei tohi esineda või mõni peab esinema/ei tohi esineda.

KELA_ATC	Ravimi(te) ATC-kood(id) või koodiosa(d). Juhtude leidmise ühe sammuna leitakse isikud, kellele on määratud ravimeid, mis on kirjeldatud sellel väljal.
REGISTER	Saab määrata, mis registrist/registritest pärit diagnoosidega isikuid valida juhtude gruppi. Tegemist on tekstiväljaga, kuhu filtreerimiseks peab kirjutama registri(te) nime(d) sellisel kujul nagu see/need on kirjas doonorite andmefailis. Mitme registri järgi filtreerimiseks eraldada nimed püstkriipsuga. Veerg ei ole kohustuslik.
SPECIALITY	Saab määrata, kas valida juhtude gruppi ainult perearsti või eriarsti poolt diagnoositud isikuid. Veerg ei ole kohustuslik.
MIN_OCCURRENCES	Saab määrata, mitu korda peab isikul olema RHK-10 või ATC-koodi esinenud, et juhtude gruppi kuuluda. Veerg ei ole kohustuslik.

RHK-10 koodide ja ATC-koodi järgi juhtude ja kontrollide leidmiseks on vajalik eelnevalt fenotüüpide failis vastavas veerus olev koodijada muuta programmile arusaadavaks. Selleks on kasutatud regulaaravaldisi, mis on kindlaks määratud sümbolitest koosnev sõne, millega saab tekstist otsida soovitud mustreid (*Regulaaravaldised — Pythoni materjalid 2025 dokumentatsioon*, 2025). FinnGen fenotüüpide definitsioonide failis eristatakse RHK-10 koode registrite järgi. Sellepärast on välja toodud eraldi tulbad COD_ICD_10 ehk surmapõhjuste registrist pärit RHK-10 koodid ning HD_ICD_10 ehk haiglast pärit RHK-10 koodid. Sama põhimõttega on eristatud ka COD_ICD_10_EXCL ja HD_ICD_10_EXCL väljadel olevad RHK-10 koodid. Suures osas on olenemata registrist nende väljade koodijadad samad, kuid alati need siiski ei kattu. Seda arvestades koostatakse eraldi surmaregistri RHK-10 regulaaravaldised ja haiglast pärit RHK-10 regulaaravaldised.

Programm suudab töödelda koodijadasid, milles sisalduvad koodid vastavad õigele vormingule (RHK-10 ja ATC-koodide õiged vormingud on kirjeldatud tabelis 2) või on vormingus, millest on võimalik soovitud koodivahemik tuletada. Kui märgitud on koodijada, mida programm ei suuda töödelda, siis kirjutatakse selle kohta vastav teade nii programmi töö käigus tehtavasse logifaili kui ka tulemuste faili. Nii saab pärast programmi töö lõppu kontrollida, kas kõik fenotüüpide defineerimise failis märgitud RHK-10 koodijadad on suudetud töödelda ja on arvesse võetud juhtude ja kontrollide moodustamisel.

FinnGen fenotüüpide defineerimise failis on RHK-10 koodijadades kasutatud mitmeid erinevaid märke. Näiteks püstkriipsuga tähistatakse “või”, sidekriipsuga “kuni”, &-märgiga “ja”. Olenevalt märkidest oleneb ka see, millised RHK-10 koodid sellesse koodijadasse

kuuluvad. Näiteks "A1[2-4]" koodijada tähistab koode A12, A13 ja A14. Samas koodijada "A1[2|4]" tähistab koode A12 ja A14. Kui pärast kantsulge on koodijadas veel numbreid, siis tähistab see pärast punkti lisanduvat täpsustust. Seega koodijada "A1[2-4]5" tähistab koode A12.5, A13.5, A14.5. Ilma märkideta koodijada “A1234” tähistab koodi A12.34 ning ainult kantsulgudega koodijada “A12[34]” tähistab koode A12.3 ja A12.4. FinnGen fenotüüpide defineerimise failis on koodijadadena märgitud ka suuremaid vahemikke. Näiteks A10–A15 tähistab kõiki koode, mis sellesse vahemikku jäävad. Esineb ka vahemikuga koodijadasid, kus kategooria täht on erinev. Näiteks A95–B05 tähistab kõiki koode, mis sellesse vahemikku jäävad, arvestusega, et kõige suurem number saab olla 99 ja väikseim 00. Seega viimane A-kategooria kood selles vahemikus oleks A99 ja esimene B-kategooria kood B00.

Erinevaid koodijadade esitlusviise esineb andmefailis veel, kuid kõigi erinevate variantide regulaaravaldiste moodustamisel jälgiti reeglit, et RHK-10 koodi esimene sümbol on alati suurtäht, mis tähistab haiguse/seisundi peatükki ja sellele järgneb kaks numbrit, seejärel võib järgneda punkt ja veel täpsustavaid numbreid. See, mitu numbrit koodis on märgitud, oleneb fenotüübi spetsiifilisusest. Kui märgitud on ainult suurtäht, siis see tähendab, et antud fenotüübi juhtudel peaks olema diagnoosina märgitud RHK-10 kood, mis algab sama tähega, kuid see, millised konkreetsed numbrid tähele järgnevad, ei mõjuta juhtudesse kuulumist. Samas kui fenotüübi defineerimisel on märgitud täht ja numbrid, mis sellele peavad järgnema, siis loetakse juhtude hulka ainult sellised isikud, kellel konkreetse tähe ja numbritega RHK-10 kood olemas. Fenotüüpide defineerimise andmefailis märgitud RHK-10 koodi järgi sobivate RHK-10 koodidega juhtude leidmine on kirjeldatud tabelis 4.

Tabel 4. Fenotüübi RHK-10 koodi struktuuri järgi juhtude määramine.

RHK-10 koodi struktuur	**Juhtude hulka kuuluvate isikute koodid**	**Juhtude hulka mittekuuluvate isikute koodid**
A	Kõik tähega “A” algavad koodid, olenemata sellest, mis numbrid järgnevad.	Kõik “A”-st erineva tähega koodid, olenemata mis numbrid järgnevad.
A1	Kõik tähega “A” algavad koodid, millele järgneb number 1, aga sellest järgnevad numbrid kuuluvust ei mõjuta.	Kõik “A”-st erineva tähega koodid ja kõik “A”- koodid, millele ei järgne number 1.
A12	Kõik tähega “A” algavad koodid, millele järgneb number 12, aga sellest järgnevad numbrid kuuluvust ei mõjuta.	Kõik “A”-st erineva tähega koodid ja kõik “A”- koodid, millele ei järgne number 12.

A12.3	Kõik tähega “A” algavad koodid, millele järgneb number 12, punkt ja number 3. Sellest järgnevad numbrid kuuluvust ei mõjuta.	Kõik “A”-st erineva tähega koodid ja kõik “A”- koodid, millele ei järgne number 12, punkt ja number 3.
A12.34	Kõik tähega “A” algavad koodid, millele järgneb number 12, punkt ja number 34.	Kõik “A”-st erineva tähega koodid ja kõik “A”- koodid, millele ei järgne number 12, punkt ja number 34.
A12–A34	Kõik tähega “A” algavad koodid, mis jäävad antud vahemikku. Sobivad koodid võivad olla ka täpsemad, näiteks A12.00 või A34.99.	Kõik “A”-st erineva tähega koodid ja kõik “A”- koodid, mis antud vahemikku ei jää.
A95–B05	Kõik tähega “A” ja “B” algavad koodid, mis jäävad antud vahemikku, arvestades, et suurim number on 99 ja vähim 00. Sobivad koodid võivad olla ka täpsemad, näiteks A95.00 või B05.99.	Kõik “A”-st ja “B”-st erineva tähega koodid ja kõik “A” ja “B”- koodid, mis antud vahemikku ei jää.

ATC-koodijadadesse kuuluvate koodide leidmise põhimõte on sarnane RHK-10 koodijadadesse kuuluvate koodide leidmisega. Erinevusena RHK-10 koodijadadest on FinnGen fenotüüpide defineerimise failis ATC-koodijadades märkidest kasutatud ainult “kuni”-märki. Koodijada “A12B” tähistab kõiki ATC-koode, mille 4 esimest sümbolit on samasugused, jada “A12BC” neid, mille esimesed 5 sümbolit on samasugused, jada “A12BC3” neid, mille esimesed 6 sümbolit on samasugused ja jada “A12BC34” tähistab neid ATC-koode, mille kõik 7 sümbolit ehk kogu ATC-kood on samasugune. Koodijada “A12BC3[4-6]” tähistab ATC-koode A12BC34, A12BC35 ja A12BC36.

Testimaks, et programm suudab RHK-10 koodide ja ATC-koodide regulaaravaldisi õigesti moodustada, lisas töö autor programmile ühiktestid, mis võimaldavad kontrollida, kas testitavad koodiosad töötavad nii nagu ette nähtud (*Unittest*, 2025). Töö autor koostas erinevaid ühikteste vastavalt sellele, milliseid RHK-10 ja ATC koodijadasid FinnGen fenotüüpide defineerimise andmefailis esines ning hindas, kas programmi loodud regulaaravaldised on õigesti moodustatud. Ühiktestid on abiks ka siis, kui programmi neid osi on vaja muuta, mis regulaaravaldisi koostavad. Pärast muudatusi on võimalik ühiktestidega veenduda, et loodud mustritega on jätkuvalt võimalik õigeid juhte ja kontrolle leida.

2.2.3 Juhtude ja kontrollide leidmine

Juhtude ja kontrollide leidmiseks on vaja arvesse võtta fenotüüpide defineerimisel seatud reegleid. Näiteks kui fenotüübi definitsioonis on olemas RHK-10 kood või koodiosa, siis juhud moodustuvad inimestest, kellel on sellele vastav kood doonorite andmeraamistikus olemas. Kui lisaks on näiteks täpsustatud eeltingimusena inimese vanus diagnoosimise hetkel või tingimusena mõni teine diagnoos, mis samuti juhuks sobival isikul peab olemas olema või ei tohi olla, siis tuleb ka seda arvesse võtta. Samuti on paljude fenotüüpide definitsioonides toodud teiste fenotüüpide nimed, mille juhtudeks loetud isikud tuleb ka esimese fenotüübi juhtudeks lugeda. Seega on fenotüüpe, mis sõltuvad teistest fenotüüpidest ja mille juhtude ja kontrollide leidmisel on vaja eelnevalt leida nende fenotüüpide juhud. Sarnaselt on ka kontrollidele seatud kindlad reeglid, mis peavad täidetud olema.

Eeskujuna juhtude ja kontrollide leidmisel kasutati Soome biopanga FinnGen projekti *Risteys* töövoogu (*Risteys*, 2024). *Risteys* võimaldab sirvida FinnGen andmeid fenotüübi tasemel tuues välja statistika iga fenotüübi juhtude alla kuuluvate inimeste kohta. Magistritöö raames tehtud programm võimaldab leida juhud ja kontrollid kas kõigile fenotüüpide definitsioonide failis olevatele fenotüüpidele või määrata konkreetsed fenotüübid, millele juhud ja kontrollid leida. Kui kasutaja soovib leida juhte ja kontrolle kindlatele fenotüüpidele, siis saab nende fenotüüpide nimed lisada projekti kaustas *konfiguratsioon* olevasse faili *arvutatavad_endpointid.txt* (projektis on fenotüübile viidates kasutatud sõna "*endpoint*", sest FinnGen fenotüüpide defineerimise failis on neid nimetatud *endpointideks*), kusjuures igal real peab olema üks fenotüübi nimi samal kujul nagu need on fenotüüpide defineerimise failis. Programmi alguses kontrollitakse selle faili sisu ning kui seda faili ei ole või on see tühi, siis arvestatakse, et kasutaja soovib leida juhte ja kontrolle kõikidele fenotüüpidele fenotüüpide defineerimise failis.

Fenotüüpide juhud ja kontrollid leitakse eraldi, sest kontrolle ei saa leida, kui juhud on leidmata. Kuna kontrollide leidmise kahes viimases sammus on fenotüübi enda juhtude eemaldamine kontrollide seast ning vajadusel ka teiste fenotüüpide juhtude eemaldamine kontrollide seast, siis leitakse kõigepealt seotud fenotüüpide juhud ning seejärel alustatakse kontrollide leidmisega. Fenotüüpide juhtude leidmise järjekorra tuvastamiseks kasutatakse järjekorra andmestruktuuri (*deque*). Alguses lisatakse sinna hetkel töödeldava fenotüübi nimi ning seejärel tsükli igal sammul võetakse järjekorra esimene element välja ning tehakse järgnevad tegevused. Esimesena kontrollitakse, kas fenotüüp on vigane. Fenotüüp on vigane siis, kui selle nimi puudub fenotüüpide defineerimise failis või kui sellele ei õnnestunud leida

regulaaravaldist, sest kas RHK-10 kood või ATC-kood ei vastanud õigele vormingule. Kui fenotüüp on vigane, siis juhtude leidmine katkestatakse ja sellest teavitatakse kasutajat programmi väljundis. Kui vigu ei leitud, siis leitakse täitmata eeldused ehk fenotüübi defineerimisel kasutatud teised fenotüübid, millele ei ole veel juhte leitud. Kui leiti täitmata eelduseid, siis lisatakse nii eeldused kui hetkel töödeldav fenotüüp uuesti järjekorda ning nende juurde tullakse hiljem tagasi. Hiljem võib selguda, et ka mõnel eeldusel on eeldused, seega võib keerulisemate fenotüüpide puhul olla vaja välja arvutada mitu taset eelduseid enne kui on võimalik kasutaja poolt soovitud fenotüübi jaoks juhte leida. Kui täitmata eelduseid ei leitud ehk kui selle fenotüübi defineerimisel ei ole ühtegi teist fenotüüpi kasutatud või on nende jaoks juhud juba leitud, siis alustatakse juhtude leidmist vastavalt fenotüübi definitsioonile. Seega iga otsitava fenotüübi puhul leitakse juhud esmalt kõikidele fenotüüpidele, mida otsitava fenotüübi definitsioonis on kasutatud ning kui need kõik on leitud ehk eeldused on täidetud, saab leida otsitava fenotüübi enda juhud.

Juhtude arvutamisel rakendatakse juhtudele seatud nõudeid kindlas järjekorras. Juhtude leidmise töövoo esimesetes filtreerimise sammudes teisi fenotüüpe kasutatud ei ole, seega ei ole neil sammudel vaja ka veel eelduseid arvesse võtta. Nendes sammudes filtreeritakse haiguslood soo, minimaalsete haigusjuhtude arvu, registrite ja arsti liigi järgi ning seejärel leitakse iga sama diagnoosi või ravimi koodi varaseim kirje, sest edasistes sammudes ei ole hilisemaid kirjeid vaja enam arvesse võtta. Tihtipeale on nende filtrite väärtused mitme fenotüübi peale samad, mistõttu on kasulik kõikide fenotüüpide üleselt need sammud valmis ära arvutada enne, kui eelmises lõigus mainitud iga fenotüübi spetsiifilisi samme tegema hakatakse. Nende sammude jaoks koostatakse sõnastik, kus iga filtreerimise kombinatsiooni kohta, mida arvutatavate fenotüüpide või nende eelduste definitsioonides on kasutatud, leitakse sobivad esmased haiguslood. Seega näiteks kui fenotüübil on defineeritud sugu mees, minimaalsete haigusjuhtude arv 3, register Haigekassa ja arsti liik eriarst, siis leitakse kõik haigusjuhud, mis sellisele filtreerimise kombinatsioonile vastavad ning nende seast filtreeritakse omakorda esmased haiguslood ja salvestatakse need sõnastikku. Selline lahendus säästab aega – nii ei pea iga fenotüübi jaoks seda aeganõudvat filtreerimist uuesti tegema, vaid saab sõnastikust valida fenotüübi jaoks sobiva kombinatsiooni järgi leitud juhtude haiguslood mis võetakse sisendiks edasisel filtreerimisel.

Soo järgi filtreerimiseks kontrollitakse, kas fenotüüpi kuuluvad juhud peaksid olema kindlast soost. Sel juhul peab fenotüübi veeru SEX lahtris olema märgitud kas 1 või 2, vastavalt mehe või naise kohta. Seega on kolm võimalikku filtreerimise varianti: 1, 2, ja tühi. Iga kasutatud varianti arvestatakse eelnevalt kirjeldatud kombinatsioonide moodustamisel. Iga soo kohta

lisatakse sõnastikku kõik sobivad haiguslood, sest neid on hiljem vaja kasutada kontrollide leidmisel ning seejärel jätkatakse juhtude jaoks sobivate haiguslugude filtreerimisega, võttes järgmises sammus arvesse vaid iga kasutatud sootunnusega järgi jäänud haiguslugusid ja fenotüüpe.

Minimaalsete haigusjuhtude arvu järgi filtreerimisega saab määrata, mitu korda peab inimesel olema haigus diagnoositud või ravim määratud. Selleks kirjutatakse fenotüübi veeru MIN_OCCURRENCES lahtrisse soovitud diagnooside esinemiste arv inimesel. Eelnevas filtreerimise sammus järgi jäänud haiguslood grupeeritakse isikute ja RHK-10 koodide põhjal ning kui doonorite andmeraamistikus on ravimite informatsioon, siis ka isikute ja ATC-koodide põhjal ning leitakse need haiguslood, mida on sama palju või rohkem kui välja MIN_OCCURRENCES välja väärtus.

Registri järgi filtreerimiseks peab fenotüübi veeru REGISTER lahtrisse kirjutama registri nime sellisel kujul nagu see on kirjutatud isikute andmefailis. Kui tahetakse filtreerida juhte mitmest registrist korraga, siis tuleb registrite nimed eraldada püstkriipsuga. Kuna doonorite ravimite andmetel ei ole registri informatsiooni kaasas, siis ravimite andmeid registri järgi ei filtreerita.

Juhtude filtreerimiseks arsti liigi järgi on võimalus fenotüübi veeru SPECIALITY lahtrisse kirjutada näiteks 1 või 2, kui eelnevalt on isiku haiguslugude puhul perearsti diagnoosid kodeeritud sõnega 1 ja ning eriarsti diagnoosid sõnega 2. Kuna doonorite ravimite andmetel ei ole arsti liigi informatsiooni kaasas, siis ravimite andmeid arsti liigi järgi ei filtreerita

Esimese haigusjuhu filtreerimine tähendab, et iga isiku jaoks grupeeritakse haiguslood RHK-10 põhjal ning jäetakse alles ainult kuupäeva järgi esimene. Kui ravimite informatsioon on andmeraamistikus olemas, siis tehakse sama grupeerimine ka ATC-koodi põhjal ning ühendatakse RHK-10 esimeste haigusjuhtudega. See filtreerimine on vajalik selleks, et kui fenotüübi definitsioonis on seatud piirang inimese vanusele diagnoosi saamise hetkel, siis programm oskaks arvestada vanust kõige esimese diagnoosi saamise ajal mitte korduvdiagnooside ajal. Lisaks suurendab see hilisemate sammude efektiivsust, kuna korduvad haiguslood filtreerimisel midagi juurde ei anna, kuid samas töödeldavate andmete hulk on väiksem. Iga unikaalse filtrite kombinatsiooni jaoks lisatakse esmased haiguslood eelnevalt kirjeldatud sõnastikku. Sõnastiku võtmed moodustatakse arvutatavate fenotüüpide ja nende eelduste defineerimisel kasutatud filtreerimise kombinatsioonidest nii, et esimesel kohal on sootunnus, teisel minimaalsete haigusjuhtude arv, kolmandal register ja neljandal arsti liik ning iga filtreerimise tunnuse vahel on #-sümbol.

Seejärel jätkub juhtude leidmise töövoog tingimuste järgi filtreerimisega, kus on juba üleval mainitud eelduseid vaja arvesse võtta. Selles sammus kasutatakse sisendiks eelnevalt leitud haiguslugusid vastavalt konkreetse fenotüübi filtrite väärtustele. Kontrollitakse tingimuste veeru CONDITIONS sisu. Kui seal olev tingimus on piirang kehamassiindeksile (näiteks BMI>25), siis eemaldatakse edasisest töötlusest isikud, kelle kehamassiindeks tingimusele ei vasta. Lisaks kehamassiindeksi kontrollile lubab programm tingimuste veergu lisada ka teiste fenotüüpide nimesid. Sellise tingimuse koostamisel on lubatud kasutada sümboleid "|", "&", "!" ja ümarsulud. Püstkriips tähistab disjunktsiooni ehk "või"-operaatorit ja see tähendab, et vähemalt üks kahel pool püstkriipsu olevatest tingimustest peab olema täidetud. Ampersand tähistab konjunktsiooni ehk "ja"-operaatorit ja see tähendab, et mõlemad tingimused selle ümber peavad olema täidetud. Hüüumärk tingimuse või sulgudega ümbritsetud tingimuste ees eitab tingimuse või tingimused ehk selliseid fenotüüpe ei tohi juhtude gruppi kuuluval isikul olla.

Selleks, et tingimusi arvesse võtta, on vaja tingimuste jadast välja lugeda eraldi tingimused. Kui tingimuses on kasutatud sulge, avatakse need nii, et terve tingimuse mõte jääks samaks. Selleks kasutati De Morgani seaduseid ehk lausearvutuse samaväärsuseid, mis kirjeldavad kuidas konjuktsiooni või disjunktsiooni eitust esitada samaväärselt ilma sulgudeta (*De Morgan's Laws*, s.a.). Need seadused ütlevad, et konjunktsiooni eitus on samaväärne selle üksikute komponentide eituste disjunktsiooniga ning disjunktsiooni eitus on samaväärne selle üksikute komponentide eituste konjunktsiooniga.

De Morgani seaduseid väljendatakse matemaatilisel kujul järgnevalt:

$$\neg(X \wedge Y) = \neg X \vee \neg Y$$

$$\neg(X \vee Y) = \neg X \wedge \neg Y$$

kus X ja Y tähistavad tingimusi, $\wedge$ tähistab konjunktsiooni, $\vee$ tähistab disjunktsiooni ja $\neg$ tähistab eitust.

Seega enne juhtude filtreerimist tingimuse järgi on vaja tingimusest eemaldada sulud. Hüüumärki kasutatakse programmeerimise kontekstis loogilise eitusena ehk näiteks !T1D tähendab, et isikul ei tohi olla diagnoositud I tüübi diabeet. Vastavalt De Morgani samaväärsuste seadustele sulgude avamisel läheb hüüumärk ehk eitus iga sulu sees oleva tingimuse ette ning sulgude sees olevad operaatorid vahetuvad: konjuktsioon disjunktsiooniks ja vastupidi. Näiteks tingimus !(F5_SCHIZO|F5_SCHIZOAFF|F5_BIPO) muudetakse tingimuseks !F5_SCHIZO&!F5_SCHIZOAFF&!F5_BIPO, mida on programmil lihtsam

töödelda, sest programm saab eeldada, et eitus võib vaid iga üksiku tingimuse ees olla ning sulge ei pea enam arvesse võtma.

Programm jagab regulaaravaldist kasutades töödeldud tingimuste jada osadeks. Nii tekib tingimuste jadast järjend, kus on fenotüüpide nimed (koos hüüumärgiga, kui see on olemas) ning püstkriipsu ja ampersandi operaatorid. Seejärel töödeldakse tsüklis tingimuse osasid võttes tsükli igal sammul töötlusesse ühe osa. Kui osaks on operaator, siis jäetakse see järgmisteks tsükli sammudeks meelde. Kui osaks on fenotüübi nimi, siis leitakse juba varem valmis arvutatud juhtude seast selle fenotüübi juhud. Seejärel leitakse need isikud, kellega edasist töötlust teha: kui fenotüübi nime ees hüüumärki ei ole, ongi sobivateks isikuteks selle fenotüübi juhud, kuid kui nime ees on hüüumärk, on sobivad isikud need, kes selle fenotüübi juhtude seas ei ole.

Järgnevalt leitakse ühisosa või ühend varasematel tsükli sammudel leitud isikutega, vastavalt kas meelde jäetud operaator oli konjunktsioon või disjunktsioon. Selleks saab kasutada tehteid hulkadega. Hulk on *Pythonis* andmetüüp, mis on sarnane järjendiga, kuid erinevalt järjendist saavad elemendid esineda ainult ühes korduses ja elementide järjekord ei ole oluline. Hulkadega saab teha tehteid nagu näiteks ühisosa ja ühendi leidmine. Kui tingimuste osade vahel on konjunktsioon, siis kasutatakse hulkade ühisosa leidmist ehk leitakse need isikud, kellel on mõlemad tingimused täidetud. Kui tingimuste jadas on püstkriips, kasutab programm ühendi leidmist ehk leiab need isikud, kellel on kas üks või teine tingimus täidetud. Seejärel leitakse varasemates sammudes filtreeritud haiguslugude seast nende isikute haiguslood, kes vastavad tingimustele. Näiteid andmefailis esinevatest tingimustest ja nende tähendustest on toodud tabelis 5.

Tabel 5. Näited FinnGen fenotüüpide defineerimise failis esinevate tingimuste veeru CONDITIONS väärtustest ja nende tähendused.

Tingimuste lahtri sisu	Lihtsustatud esitus	Tähendus
J10_ASTHMA	A	Juhul peab olema diagnoositud astma.
!C3_BREAST	!A	Juhul ei tohi olla diagnoositud rinnavähk.
ASTHMA_MIXED|ASTHMA_NAS	A|B	Juhul peab olema diagnoositud kas segavormiline astma või kategoriseerimata astma (sobib ka siis, kui on diagnoositud mõlemad).

K11_CHOLANGI_STRICT&K11_IBD_STRICT	A&B	Juhul peab olema diagnoositud sapiteede põletik ja põletikuline soolehaigus (ei sobi, kui on diagnoositud ainult üks neist).
!(F5_SCHIZO|F5_SCHIZOAFF|F5_BIPO)	!(A|B|C)	Juhul ei tohi olla diagnoositud skisofreenia ega skisoafektiivne häire ega bipolaarne häire.
!I9_IHD&!I9_HYPERTROCARDMYOP	!A&!B	Juhul ei tohi olla diagnoositud isheemiline südamehaigus ega hüpertooniline kardiomüopaatia.
K11_KELAIBD&!K11_CROHN	A&!B	Juhul peab olema diagnoositud põletikuline soolehaigus ja ei tohi olla diagnoositud Crohni tõbi.
!C3_LIVER_INTRAHEPATIC_BILE_DUCTS|!C3_GALLBLADDER|!C3_BILIARY_TRACT	!A|!B|!C	Juhul ei tohi olla diagnoositud maksa ja sapijuhade vähk ja sapipõievähk ja sapiteede vähk (kui 0-2 neist on diagnoositud, siis sobib juhuks, aga kui kõik 3 korraga, siis ei sobi).

Edasi kontrollitakse eeltingimustele vastavust veeru PRE_CONDITIONS väärtuse põhjal. FinnGen definitsioonide failis on sellel väljal kasutatud piirangut inimese vanusele diagnoosi hetkel ja diagnoosi väljastamise aastale, mida ka magistritöö tulemusel valminud programm toetab. Eeltingimuse vastavuse hindamiseks kontrollitakse esmalt kas tegemist on nõudega vanusele või aastale, seejärel leitakse operaator(id) ning kontrollitakse nõude täitmist biopanga faili abil välja arvutatud inimese vanuse järgi diagnoosi hetkel ja diagnoosi sisestamise aasta järgi. Näited FinnGen fenotüüpide defineerimise failis esinevatest fenotüüpide eeltingimustest on toodud tabelis 6.

Tabel 6. Näited FinnGen fenotüüpide defineerimise failis esinevate eeltingimuste veergude PRE_CONDITIONS ja CONTROL_PRECONDITIONS väärtustest ja nende tähendused.

Eeltingimuste lahtri sisu	Tähendus
EVENT_AGE<16	Juhtudesse kuuluv isik peab esmase diagnoosi saamise hetkel olema noorem kui 16-aastane.
16<=EVENT_AGE<=44	Juhtudesse kuuluv isik peab esmase diagnoosi saamise hetkel olema vähemalt 16-aastane, aga noorem kui 45-aastane.
EVENT_AGE>=50	Juhtudesse kuuluv isik peab esmase diagnoosi saamise hetkel olema kas 50-aastane või vanem.

EVENT_YEAR<1994	Juhtudesse kuuluvale isikule peab esmane diagnoos olema väljastatud enne 1994. aastat.
DEATH_AGE>=65	Kontrollidesse kuuluv isik peab olema surnud ja surma hetkel vähemalt 65-aastane.

Kui kõik eelnevad sammud on tehtud, siis on doonorid eelfiltreeritud. See tähendab, et ettevalmistus järgnevateks sammudeks on tehtud ja järgmises kolmes sammus kasutatakse sisendina eelfiltreerimise tulemust.

Järgmiseks võetakse arvesse fenotüübile defineeritud ravimid. Selleks kontrollitakse, kas fenotüübi ravimite veerus ATC_MUSTRID on varasemalt loodud regulaaravaldis KELA_ATC välja põhjal. Kui seal on väärtus olemas, kuid kasutaja programmile ravimite andmefaili sisendiks ei andnud, siis tõstatakse vastava teatega erind ning selle fenotüübi jaoks juhte ja kontrolle ei leita ning vastav teade kirjutatakse ka tulemuste faili. Kui aga ravimite info on olemas, siis leitakse siiani filtreeritud juhtude seast need inimesed, kellel on sama ATC-koodiga ravim määratud. Mõningatel juhtudel ei ole KELA_ATC veeru väärtuseks kindla ravimi ATC-kood vaid koodiosa, mis tähistab kindlat gruppi ravimeid. Sellisel juhul leitakse inimesed, kellele on selle grupi ravimeid määratud. Kuna juhtude ja kontrollide maatriksi faili märgitakse juhtudesse kuuluva isiku esimese sobiva diagnoosi kuupäev, siis salvestatakse see ka siin etapis. Maatriksisse märgitakse ka aeg, mis kulus isiku biopangaga liitumisest konkreetse diagnoosi saamiseni, mis samuti selles sammus välja arvutatakse. Aja arvutamiseks kasutatakse esmakordse diagnoosi saamise kuupäeva, millest lahutatakse isiku biopangaga liitumise kuupäev ning vastus salvestatakse päevades. Kui diagnoos on saadud enne biopangaga liitumist, siis on see väärtus negatiivne.

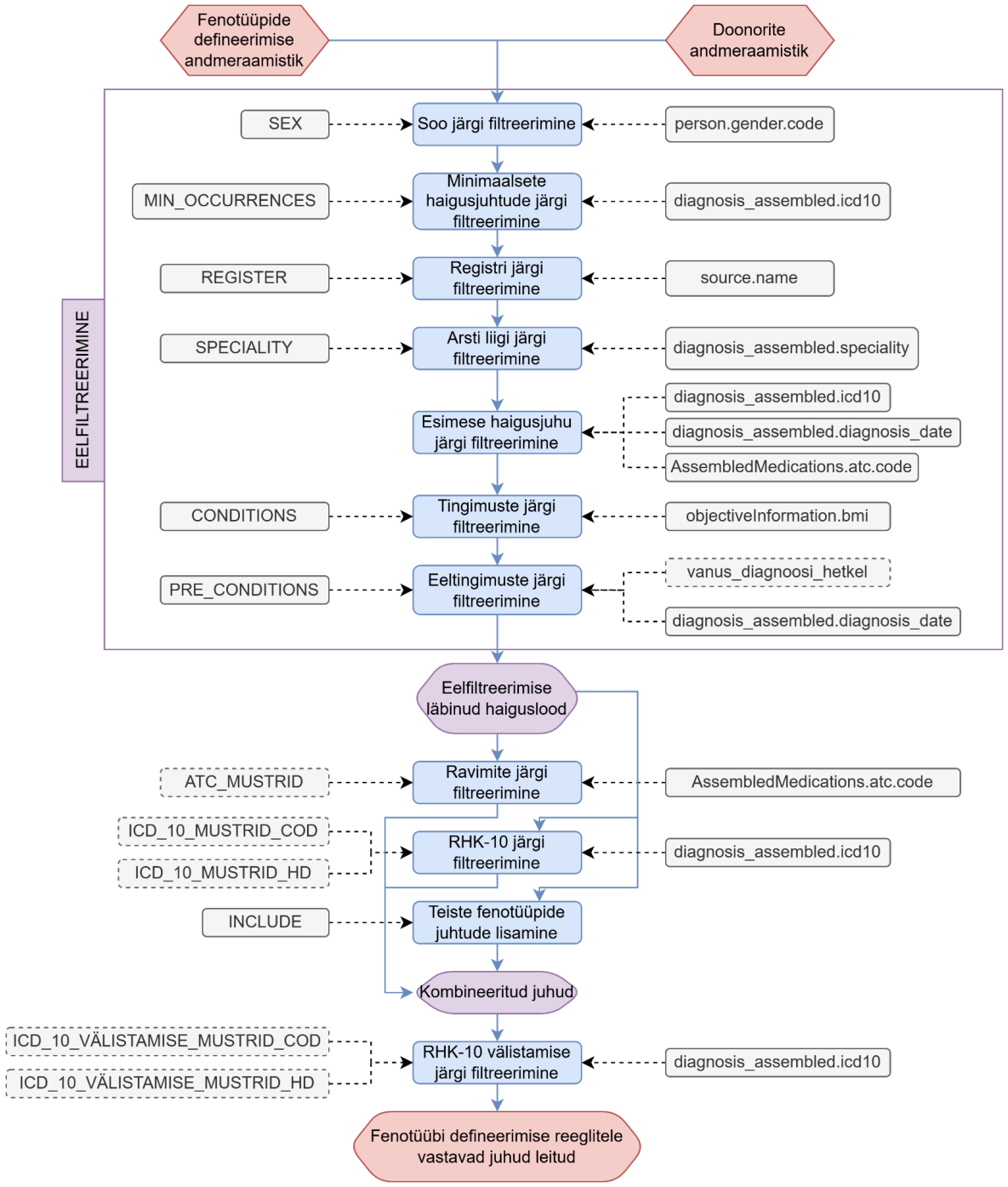
Edasi jätkub filtreerimine RHK-10 põhjal, kus kontrollitakse RHK-10 koodile vastavust veergude COD_ICD_10 ja HD_ICD_10 põhjal. RHK-10 alusel juhtude filtreerimiseks kasutatakse eelnevalt leitud regulaaravaldisi ja leitakse nende abil eelfiltreeritud isikute haiguslugude seast need haiguslood, kus on sobiv RHK-10 kood biopanga failis olevates sobivatest registritest pärit haigusjuhtudes olemas. COD_ICD_10 veeru puhul on selleks registriks surmaregister (doonorite failis nimega “Surma põhjuste register” või *“Causes of Death Registry”*) ja HD_ICD_10 veeru puhul kõik ülejäänud registrid. Edasi filtreeritakse sobivate haiguslugude ühendit. Nii nagu eelneval sammul, salvestatakse ka siin isiku esimese sobiva diagnoosi kuupäev ja aeg haigestumiseni ning leitakse salvestatud kuupäevadest ja aegadest diagnoosi jaoks väikseim. Leitud sobivad isikud ühendatakse ravimite järgi

filtreerimise sammus leitud sobivate isikutega ning kui isik on nii RHK-10 kui ravimi järgi sobiv, siis leitakse nende kahest kuupäevast varaseim.

Järgmise sammuna lisatakse RHK-10 koodide ja ravimite koodide abil leitud juhtudele juurde veeru INCLUDE lahtris märgitud fenotüüpidesse kuuluvad juhud. Fenotüüpide failis on selles veerus märgitud teiste fenotüüpide nimed, mille juhud tuleb lisada hetkel töödeldava fenotüübi juhtudeks. Selle saavutamiseks eraldatakse kõik jadas olevad fenotüüpide nimed püstkriipsu koha pealt ning iga fenotüübi kohta kaasatakse juba varem leitud juhud. Seejärel leitakse iga fenotüübi juhtude ning eelfiltreeritud juhtude ühisosa, et mitte lisada juhtudesse neid isikuid, kes eelmistele nõuetele ei vasta. Lisaks täiendatakse siin ka esmakordse diagnoosi kuupäeva fenotüübi iga juhu jaoks, et kehtima jääks sobivatest diagnoosidest varaseim kuupäev.

Viimase sammuna eemaldatakse RHK-10 välistamise veerus COD_ICD_10_EXCL ja HD_ICD_10_EXCL lahtris olevate koodidega inimesed juhtude seast. Seda filtreerimist viiakse läbi samuti eelnevalt koostatud regulaaravaldise abil, kuid need inimesed kellel vastavast registrist sobiv kood või koodid leitakse, eemaldatakse juhtudest.

Sellega on kõik sammud juhtude leidmiseks tehtud. Iga sammu järel kirjutab programm logisse informatsioon selle kohta, mitu isikut on algsest doonorite arvust alles jäänud. Juhud salvestatakse mällu ning seejärel minnakse tagasi eelmisse etappi, kus leitakse juhud ka teistele fenotüüpidele, millele kasutaja juhte ja kontrolle leida soovib või fenotüüpidele, mis on arvutatava fenotüübi eelduseks. Töövoo ülevaade juhtude leidmiseks on kujutatud joonisel 6.

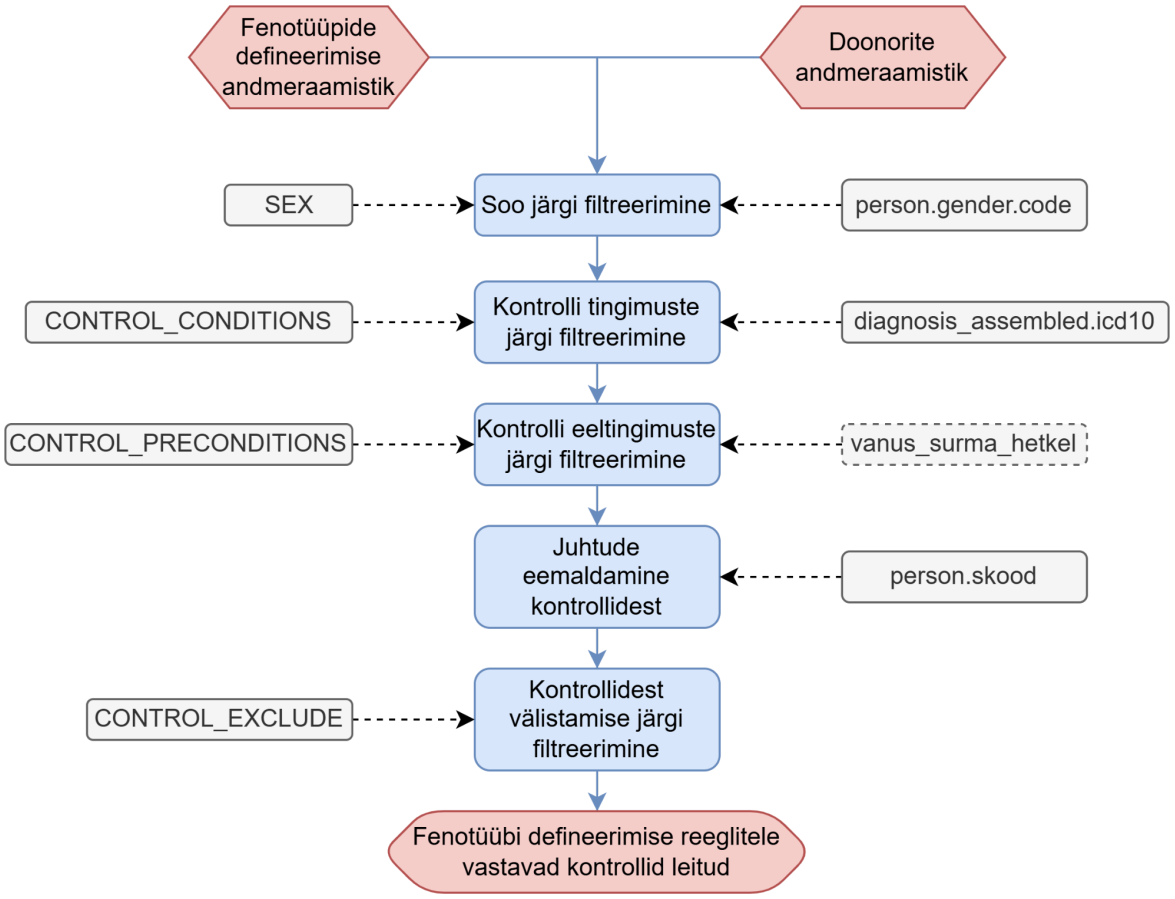


Joonis 6. Töövoog juhtude leidmiseks. Sisendina kasutatakse doonorite ja fenotüüpide andmeraamistikke. Sinisega on toodud filtreerimise etapid. Halliga on toodud vastavast andmeraamistikust filtreerimise etapi jaoks kasutatava informatsiooni sisaldava veeru nimi. Veeru nimi on tähistatud punktiirjoonega, kui seda ei ole algselt andmefailis olemas vaid see on programmi poolt välja arvutatud väärtus. Lillaga on märgitud filtreerimise vahetulemused, mida kasutatakse mõne filtreerimise sammu sisendina.

Kui kõikide soovitud fenotüüpide ja nende eelduste jaoks on juhud leitud, siis alustatakse kontrollide leidmisega. Nii nagu juhtude jaoks on kindlad defineerimise reeglid, on need ka

kontrollide jaoks. Esmalt hinnatakse iga fenotüübi jaoks, kas sellele on võimalik kontrolle leida. Olukord, kus kontrolle ei ole võimalik leida tekib siis, kui selle fenotüübi juhtude leidmine ei õnnestunud või kui fenotüübi kontrollide defineerimisel on kasutatud teise fenotüübi nime, millele ei ole definitsiooni või on sisestatud vigane kontrollide tingimus või eeltingimus. Kui kontrolle leida ei saa, siis eemaldatakse ka juhud, kui need olid leitud. See on vajalik selleks, et ekslikult ei tunduks, et fenotüübile leiti juhud ja kontrollid, kui tegelikult kontrollide leidmine ebaõnnestus. Programm kontrollib kontrollide nõuetele vastavust kindlas järjekorras. Kontrollide leidmise esimene samm on samuti soo järgi filtreerimine, mis tehakse ära juba enne juhtude leidmist. Seejärel tehakse kontrolli tingimuste järgi filtreerimine ja kontrolli eeltingimuste järgi filtreerimine, mis toimuvad identselt juhtude leidmise vastavatele sammudele. Ainus erinevus on, et FinnGen fenotüüpide definitsioonide failis on kontrollide eeltingimuste väljal CONTROL_PRECONDITIONS kasutatud tingimust doonori vanusele surma hetkel, mille implementatsiooni töö autor ka eeltingimuste järgi filtreerimisele lisas. Vanus surma hetkel oli välja arvutatud varasemalt biopanga doonorite faili andmete põhjal ning selle vastavuse hindamine toimub samamoodi nagu juhtude eeltingimuse vastavuse hindamine.

Järgmise sammuna eemaldatakse kontrollide seast kõik inimesed, kes on sama fenotüübi juhtude seas. Seejärel eraldatakse kontrollidest välistamise veerus CONTROL_EXCLUDE väljal olevad püstkriipsuga eraldatud fenotüüpide nimed ning iga fenotüübi juhud eemaldatakse samuti kontrollide seast. Töövoo ülevaade kontrollide leidmiseks on kujutatud joonisel 7.



Joonis 7. Töövoog kontrollide leidmiseks. Sisendina kasutatakse doonorite ja fenotüüpide andmeraamistikke. Sinisega on toodud filtreerimise etapid. Halliga on toodud vastavast andmeraamistikust filtreerimise etapi jaoks kasutatava informatsiooni sisaldava veeru nimi. Veeru nimi on tähistatud punktiirjoonega, kui seda ei ole algselt andmefailis olemas, vaid see on programmi poolt välja arvutatud väärtus.

Juhtude ja kontrollide leidmine lõpeb, kui järjekorras ei ole ootel enam ühtki fenotüüpi ja seega on fenotüüpidele juhud ja kontrollid leitud või on tehtud otsus, et neid leida ei ole võimalik.

Fenotüübid võivad olla üksteisega seotud ehk ühe fenotüübi definitsioonis võib sisalduda teise fenotüübi nimi. Kui seotud fenotüübi juhtude leidmisel tekib viga (näiteks on selle RHK-10 vorming vale), siis tulemuste õigsuse tagamiseks ei leita juhte ka esimesele fenotüübile. Mõningal juhul võib tekkida aga olukord, kus fenotüübi definitsioonis vigu ei ole, aga puudu on nii RHK-10 koodid, ATC-kood ning ka INCLUDE väli on tühi. Programmi jaoks on vajalik vähemalt üks neist väljadest, et fenotüübile juhud leida. Nii võib tekkida olukord, kus ühe fenotüübi tõttu, millele ei ole võimalik juhte ja kontrolle leida, jäävad leidmata ka kõigi seda fenotüüpi definitsioonis sisaldavate fenotüüpide juhud ja kontrollid.

Sellise olukorra vältimiseks tehti lahendus, kus märgitakse see fenotüüp arvutatuks ilma juhtude ja kontrollideta isegi kui seda töödelda ei ole võimalik. Tulemusena suudetakse leida juhud ja kontrollid võimalikult paljudele fenotüüpidele. Vastav märge, et fenotüüp sisaldas definitsioonis sellist fenotüüpi, millele juhte ja kontrolle ei olnud võimalik leida, lisatakse tulemuste faili. FinnGen fenotüüpide defineerimise failis võib fenotüüp olla RHK-10 koodi, ATC-koodi või INCLUDE fenotüüpide asemel defineeritud näiteks operatsioonikoodiga, vähispetsiifilise RHK-O-3 koodiga, RHK-8 või RHK-9 koodiga. Näiteks fenotüüp AMPUTATION_TOE on defineeritud ainult operatsioonikoodiga ^NHQ40 ja fenotüüp C3_CERVIX_ADENO on defineeritud RHK-O-3 vähi topograafia koodiga C53 ja vähi morfoloogia koodiga 8140.

Juhtude ja kontrollide leidmise õigsuse kontrollimiseks loodi töö autori poolt ühiktestid, kasutades Pythoni moodulit *unittest* (*Unittest*, 2025). Näide ühest ühiktestist on olukord, kus fenotüübi jaoks on defineeritud ainult RHK-10 kood ning teisi reegleid ei ole. Ühiktestiga antakse sisendiks sobivate andmetega isikud ning veendutakse, et pärast töötlust on isikud jaotatud õigesti juhtude ja kontrollide gruppidesse. Sarnaselt on testid ka teiste fenotüübi defineerimise võimalike väljade kontrollimiseks – RHK-10 välistamine, erinevad tingimused ja eeltingimused, kontrollidest välistamise reeglid, kontrollide tingimused ja eeltingimused, kehamassiindeksi piirangud, registri, arsti liigi ja miinimum diagnooside arvu määramine, biopangaga liitumise hetkest diagnoosi saamiseni kulunud aja arvestamine, esmase diagnoosi kuupäeva leidmine, viimase diagnoosi kuupäeva või surmaregistri kande kuupäeva leidmine. Üleval kirjeldatud testid vaatlevad kogu juhtude ja kontrollide leidmise voogu tervikuna, aga loodi ka ühiktestid iga üksiku arvutamise sammu kontrollimiseks, kus muud sammud tulemust ei mõjuta. Testide lisaväärtus seisneb ka selles, et kui programmi koodi on vaja muuta või täiendada, siis saab nende testide abil veenduda, et juhtude ja kontrollide leidmine töötab jätkuvalt samamoodi.

Programmi töö katsetuste käigus tuli FinnGen fenotüüpide definitsioonide andmefailist välja erinevaid sisestusvigu. Näiteks oli ühest RHK-10 koodist puudu kategooria täht, koodijadast puudu püstkriips või olid andmed nihkes nii, et RHK-9 kood oli kirjutatud RHK-10 koodi asemel. Samuti olid failis RHK-10 koodid, kus oli ühe tähe asemel kaks tähte. Need on Soome spetsiifilised RHK-10 koodid, mis tähistavad vigastuse/haiguse asukohta kehal. Kuna selliseid koode ei ole RHK-10 rahvusvahelises klassifikatsioonis defineeritud ning Eestis neid ei kasutata, siis nende koodidega fenotüüpide definitsioonidele juhte ja kontrolle ei leitud. FinnGen fenotüüpide definitsioonide andmefailis leitud vead ja erisused on toodud lisas 3.

2.2.4 Abifailid

Lisaks programmi tööks vajalikele põhifailidele lõi töö autor mitmeid abifaile, mis programmile lisaväärtust loovad. Varasemalt on kirjeldatud kolme testimise faili – üks regulaaravaldiste moodustamise õigsuse kontrollimiseks, teine juhtude ja kontrollide erinevate reeglite täitmise õigsuse kontrollimiseks ja kolmas kõikide filtreerimise sammude õigsuse kontrollimiseks üksteisest sõltumatult. Lisaks testimisele on programmile loodud lisafunktsionaalsust erinevate abistavate failidena.

Üheks lisafunktsionaalsuseks on programmile lisatud taimer. See ei ole otseselt vajalik juhtude ja kontrollide leidmiseks, aga sellest on abi programmi töö kontrollimisel. Ajakulu mõõtmine on selle programmi puhul eriti tähtis, kuna isikute andmeid sisaldav fail võib olla väga suur - näiteks Eesti Geenivaramu doonorite failis on 212 000 inimese kohta 24 miljonit esmatasandi diagnoosi, 1,4 miljonit teise tasandi diagnoosi, 29 miljonit välja kirjutatud ravimit ja lisaks veel muid terviseandmeid (Milani et al., 2025). Failis *taimer.py* defineeritakse funktsioon taimeri alustamiseks, millega fikseeritakse algusaeg ning funktsioon taimeri lõpetamiseks, milles lisaks lõpuaja fikseerimisele lisatakse logifaili tekst selle kohta, kui kaua mingile programmi tööetapile aega kulus. Aeg väljastatakse sekundites.

Programmi töövoo kuvamiseks ja programmi vigade lihtsamaks lahendamiseks on programmile lisatud logija. Failis *logija.py* seadistatakse *Pythoni logging* (*Logging*, 2025) mooduli abil logimine. Selle jaoks kasutatakse ettemääratud vormingus faili *logi_konfiguratsioon.json*, mida eelmainitud logimise moodul oskab kasutada seadistuseks. Logimise konfigureerimisel salvestatakse kindel ajahetk programmi käivitumisel ning määratakse tekkiva logifaili nimeks programmi käivitamise kuupäev koos kellaajaga. Nii tekitatakse uus logifail igal programmi käivitamisel. Failis *logi_konfiguratsioon.json* on konfigureeritud ka konsooli logimine. Programmi töö käigus kirjutatakse mõlemasse väljundisse informatsiooni programmi töö kohta. Programmis on kasutatud kolme logimistaset: DEBUG, INFO ja ERROR. Soovitud logimistaseme saab määrata failis *logi_konfiguratsioon.json*. ERROR taseme logides on vaid programmi vead (ja juhtude ja kontrollide leidmisel tekkinud vead), INFO taseme logides on informatsioon selle kohta, milliseid juhte või kontrolle parajasti leitakse ning arenduse hõlbustamiseks on logitase DEBUG, mille puhul logitakse ka iga fenotüübi juhtude ja kontrollide leidmise iga sammu staatus. Näide logifaili sisust on toodud joonisel 8.

```
2025-05-24 09:58:18 - INFO - Proovin arvutada juhte endpointile I9_CHD
2025-05-24 09:58:18 - INFO - Endpointi I9_CHD jaoks on täitmata eelduseid.
2025-05-24 09:58:18 - INFO - Proovin arvutada juhte endpointile I9_MI
2025-05-24 09:58:18 - DEBUG - Alustan juhtude leidmist endpointile I9_MI
2025-05-24 09:58:18 - DEBUG - Pärast soo, min haigusjuhtude arvu, registrite ja arsti liigi järgi filtreerimist doonoreid järgi: 2000
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Juhtude tingimuste järgi filtreerimine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Juhtude tingimuste järgi filtreerimine doonoreid järgi: 2000
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Juhtude eeltingimuse järgi filtreerimine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Juhtude eeltingimuse järgi filtreerimine doonoreid järgi: 2000
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Juhtude ravimi järgi filtreerimine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Juhtude ravimi järgi filtreerimine doonoreid järgi: 0
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Juhtude icd-10 järgi filtreerimine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Juhtude icd-10 järgi filtreerimine doonoreid järgi: 137
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Include juhtude lisamine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Include juhtude lisamine doonoreid järgi: 137
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Juhtude icd-10 välistamise järgi filtreerimine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Juhtude icd-10 välistamise järgi filtreerimine doonoreid järgi: 137
2025-05-24 09:58:18 - INFO - Endpointi I9_MI jaoks on juhud arvutatud.
2025-05-24 09:58:18 - INFO - Proovin arvutada juhte endpointile I9_REVASC
2025-05-24 09:58:18 - INFO - Endpointi I9_REVASC jaoks on täitmata eelduseid.
2025-05-24 09:58:18 - INFO - Proovin arvutada juhte endpointile I9_CABG
2025-05-24 09:58:18 - INFO - Endpointi I9_CABG arvutamiseks on vajalik info (ICD-10, INCLUDE, KELA_ATC) puudu.
2025-05-24 09:58:18 - INFO - Proovin arvutada juhte endpointile I9_ANGIO
2025-05-24 09:58:18 - INFO - Endpointi I9_ANGIO arvutamiseks on vajalik info (ICD-10, INCLUDE, KELA_ATC) puudu.
2025-05-24 09:58:18 - INFO - Proovin arvutada juhte endpointile I9_CHD
2025-05-24 09:58:18 - INFO - Endpointi I9_CHD jaoks on täitmata eelduseid.
2025-05-24 09:58:18 - INFO - Proovin arvutada juhte endpointile I9_REVASC
2025-05-24 09:58:18 - DEBUG - Alustan juhtude leidmist endpointile I9_REVASC
2025-05-24 09:58:18 - DEBUG - Pärast soo, min haigusjuhtude arvu, registrite ja arsti liigi järgi filtreerimist doonoreid järgi: 2000
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Juhtude tingimuste järgi filtreerimine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Juhtude tingimuste järgi filtreerimine doonoreid järgi: 2000
2025-05-24 09:58:18 - DEBUG - Alustan arvutussammuga: Juhtude eeltingimuse järgi filtreerimine
2025-05-24 09:58:18 - DEBUG - Pärast sammu Juhtude eeltingimuse järgi filtreerimine doonoreid järgi: 2000
```

Joonis 8. Näide logifaili sisust. Failis on toodud igal sammul kuupäev ja kellaaeg. Üldisema informatsiooni ees on sõna “INFO”, täpsema informatsiooni ehk pigem programmi arendajale kasutatava informatsiooni ees sõna “DEBUG”.

Programmi töö testimiseks erinevate isikute ja haigusjuhtude arvudega koostati abifail *doonorite_genereerija.py*. Sellega on võimalik genereerida fail isikute andmetega ja neile määratud ravimitega, mis on mõlemad programmi tööks sobiliku struktuuriga. Mõlemas failis on kõik vajalikud veerud ning kindlaid reegleid järgides (näiteks, et RHK-10 koodid oleksid genereeritud õige vorminguga ning et diagnoosi kuupäev ei saaks olla enne isiku sündi või pärast tema surma) loodud juhuslikud andmed igasse veergu. Töö autor kasutas seda faili peamiselt erinevate isikute ja haiguslugude arvudega failide genereerimiseks, et saada aimu, kui kaua programmi erinevad etapid võiksid aega võtta päris biopanga faili kasutades.

Rakenduse kasutamise lihtsustamiseks lisas töö autor ka faili *requirements.txt*, mis on *Python*i paketihalduri *pip* abifail, millega saab programmi jaoks vajalikke pakette paigaldada ühe käsuga. Seda faili kasutades saab käsureal *pip install -r requirements.txt* käsku kasutades alla laadida kõik vajalikud paketid korraga, mis teeb programmi esmakordse käivitamise mugavamaks.

Projekti paigaldamise ja käivitamise kohta loodi juhend failis *README.md*. Selles on ära kirjeldatud projekti eeldused ehk millised failid peavad olemas olema ja mis väljad peavad olema nende failide päises. Samuti on seal juhend juhuks, kui soovitakse juhte filtreerida registri, arsti liigi või minimaalsete diagnooside arvu järgi ning selleks, kui kasutaja soovib kõikide fenotüüpide defineerimise failis olevate fenotüüpide asemel leida juhte ja kontrolle

ainult mingitele kindlatele fenotüüpidele. Lisatud on ka samm-sammuline juhend projekti jaoks *Pythoni* virtuaalse keskkonna loomiseks (*Venv — Creation of Virtual Environments*, 2025), teekide paigaldamiseks ja programmi käivitamiseks operatsioonisüsteemidel Windows ja Linux.

2.3 Tulemused ja arutelu

2.3.1 Maatriksi moodustamine

Selleks, et programmi tulemust pärast programmi tööd ka kasutada, on vaja see salvestada. Juhendajaga lepiti kokku, et tulemus salvestatakse faili maatriksi kujul, kus maatriksi esimeses veerus on doonorite identifitseerimiskoodid ja päises fenotüüpide nimed, iga fenotüübi kohta aeg haigestumiseni, isiku esmase haigestumise kuupäev ja kontrolli hiljutiseima diagnoosi kuupäev. Faili kirjutatakse juhud ja kontrollid ainult nende fenotüüpide kohta, mida kasutaja soovis leida ning mille puhul juhtude ja kontrollide leidmine õnnestus. Samuti on iga isiku kohta märgitud, kas esineb kanne surmaregistris ehk kas isik on surnud. Surmakande korral on vastavas lahtris isikul märgitud 1, kande puudumisel 0.

Fenotüübi juhtudeks sobivad isikud on maatriksis vastava fenotüübi veerus märgitud numbriga 1, kontrollideks sobivad isikud numbriga 0 ning kui isik ei sobi ei juhuks ega kontrolliks, siis on vastav fenotüübi lahter tühi. Kui isik sobib fenotüübi juhuks, on selle kõrvallahtris “(fenotüübi_nimi)_aeg_haigestumiseni” märgitud mitu päeva oli tema Eesti Geenivaramuga liitumise ja diagnoosimise vahel. Kui isiku diagnoos on saadud enne geenidoonoriks saamiseks nõusoleku andmist, on lahtri väärtus negatiivne. Selle lahtri kõrval on ka selle fenotüübi jaoks isiku esmase diagnoosi kuupäeva lahter nimega “(fenotüübi_nimi)_diagnoosi_kp”, mis juhu puhul on täidetud vormingus aaaa-kk-pp. Kui isik juhuks ei sobi, on mõlemad nimetatud lahtrid ilma väärtuseta.

Kui isik on fenotüübi jaoks kontroll, siis on lahtris “(fenotüübi_nimi)_uusima_diagnoosi_kp” tema kõige viimase diagnoosi kande kuupäev vastavast filtreeritud registrist või kui selle fenotüübi juhte registrite järgi ei filtreerita, siis arvestatakse kuupäev üle kõigi registrite kannete. Kui kontrollgruppi kuuluv isik on surnud, siis on tema viimase diagnoosi kande kuupäeva lahtris tema viimase surmaregistri kande kuupäev.

Binaarsete tunnuste uurimiseks tavapärasel ülegenoomsel assotsiatsiooniuuringul piisaks maatriksist, kus on märgitud iga isiku kohta, kas isik on fenotüübi suhtes juht või kontroll (või mõlemaks sobimatu). Eelnevalt kirjeldatud lisaväljad aga pakuvad informatsiooni

elukestusanalüüsi mudelite rakendamiseks. Näide maatriksi failis olevast informatsioonist ja väärtuste vormingutest on toodud tabelis 7.

Tabel 7. Näide maatriksi failis olevast informatsioonist ja vormingutest.

	on_surma-registri_kanne	***T1D***	***T1D*_aeg_haigestumiseni**	***T1D*_diagnoosi_kp**	***T1D*_uusima_diagnoosi_kp**
AAAAA1111	0	1	123	2016-01-03	
BBBBB2222	0	0			2025-04-29
CCCCC3333	0				
DDDDD4444	1	1	-4567	2003-06-20	
EEEEE5555	1	0			2021-11-13

Esimene isik tabelis 7 on elus, fenotüübi suhtes juht ning esimese sobiva diagnoosi sai 123 päeva pärast Eesti Geenivaramuga liitumist diagnoosi kuupäeva lahtris toodud kuupäeval. Teine isik on elus ja fenotüübi suhtes kontroll ning tema hiljutiseim diagnoos on märgitud uusima diagnoosi kuupäeva lahtrisse. Kolmas isik on elus, kuid ei sobi fenotüübi suhtes juhuks ega kontrolliks. Selline olukord tekib näiteks siis, kui fenotüüp on defineeritud soo järgi ning isik on teisest soost või on isik mingi muu tingimuse tõttu kontrollide hulgast välistatud. Neljas isik on surnud, fenotüübi suhtes juht ning vastava diagnoosi sai 4567 päeva enne Eesti Geenivaramuga liitumist diagnoosi kuupäeva lahtris märgitud kuupäeval. Viies isik on surnud, fenotüübi suhtes kontroll ning tema uusima diagnoosi kuupäeva lahtris on märgitud tema surma kuupäev.

2.3.2 Programmi testimine päris andmetega

Magistritöö tulemusel valminud programmi testiti magistritöö juhendaja Kristi Läll poolt Eesti Geenivaramu geenidoonorite andmetega. Testiti programmi erinevaid funktsioone nagu valitud fenotüüpide määramine, registrite ja arsti liigi järgi juhtude filtreerimine ning võrreldi programmi tulemusi juhendaja poolt kirjutatud skriptiga saadud fenotüüpide juhtude ja kontrollide arvudega.

Töö testimise lihtsustamiseks lisas töö autor programmi uue käsurea parameetri, millega määratakse tulemusfailide salvestamise jaoks kausta asukoht ja selle nimi. Nii saab iga kord programmi erinevate filtritega käivitades salvestada tulemusfailid uude sobiva nimega kausta, et oleks võimalikult mugav tulemusi võrrelda.

Esialgu valiti välja 5 fenotüüpi: T1D (I tüüpi diabeet), T2D (II tüüpi diabeet), I9_CHD (südame isheemiatõbi), C3_BREAST (rinnavähk), C3_PROSTATE (eesnäärmevähk), mille nimed kirjutati programmi konfiguratsiooni kaustas *arvutatavad_endpointid.txt* faili. Seejärel käivitati programm FinnGen fenotüüpide defineerimise failiga "*FinnGen core and non-core endpoints and controls DF12"* ja Eesti Geenivaramu geenidoonorite andmefailidega. Programmi tulemusena tekkisid tulemusfail, logifail ja maatriksi fail. Failid andsid edasi informatsiooni selle kohta, et T1D ja T2D fenotüüpidele juhte ja kontrolle leida ei õnnestunud, sest fenotüüpide defineerimise failis on viga. Teistele valitud fenotüüpidele leiti juhud ja kontrollid.

Fenotüüpide defineerimise failis esinevad vead kirjeldatakse programmi töö käigus tekkivas tulemuste failis. Failis on välja toodud, millise fenotüübi juhtude ja kontrollide leidmisel viga on tekkinud ning mille tõttu. Kirjeldatud on kogu jada alates fenotüübist, millele juhte ja kontrolle leida ei olnud võimalik, kuni põhjuseni, miks viga tekkis. See teeb kasutaja jaoks lihtsamaks probleemide leidmise, sest mõnel fenotüübil on palju eelduseid mitmel tasemel (eeldustel on omakorda eeldused) ja neid ükshaaval läbi kontrollida on väga aeganõudev.

Iga veaga lõppenud fenotüübi puhul on esimesel real nooltega eraldatud jada kuni vigase fenotüübini ning järgmistel ridadel taandes vigade kirjeldused. Näide programmi tulemuste failis kuvatavatest I ja II tüüpi diabeedi fenotüüpide defineerimise vigadest on toodud joonisel 8. Nii I kui II tüüpi diabeedi fenotüüpide defineerimisel jõutakse läbi teiste fenotüüpide välja DM_OTHER_WIDE nimelise fenotüübini, mille RHK-10 kood on FinnGen fenotüüpide defineerimise failis vigane. Nii COD_ICD_10 kui HD_ICD_10 väljal on DM_OTHER_WIDE fenotüübil märgitud E13|14, kus on puudu number 14 eest kategooria täht E, kuid õige oleks E13|E14. Kui viga fenotüüpide defineerimise failis parandati, siis suutis programm nii I kui II tüüpi diabeedi fenotüübile juhud ja kontrollid leida.

```
Ei suutnud leida juhte ja kontrolle järgmistele fenotüüpidele:
        T1D
        T2D
Järgmistel fenotüüpidel esinevad vead:
        T1D:
                T1D -> CONDITIONS: E4_DM2 -> INCLUDE: E4_DM2COMA -> CONTROL_EXCLUDE: DM_OTHER_WIDE
                        Vigane ICD-10 COD kood
                        Vigane ICD-10 HD kood
                T1D
                        Kontrolle ei ole võimalik leida, sest juhud selle fenotüübi jaoks ei ole leitud.
        T2D:
                T2D -> CONDITIONS: E4_DM1 -> INCLUDE: E4_DM1COMA -> CONTROL_EXCLUDE: DM_OTHER_WIDE
                        Vigane ICD-10 COD kood
                        Vigane ICD-10 HD kood
                T2D
                        Kontrolle ei ole võimalik leida, sest juhud selle fenotüübi jaoks ei ole leitud.
```

Joonis 8. Näide programmi tulemuste failis kuvatavatest fenotüübi defineerimise vigadest. Failis on näha, et I ja II tüüpi diabeedi defineerimisel esinevad vead. I tüüpi diabeedi tingimuste lahtris on fenotüüp E4_DM2, mille INCLUDE lahtris on fenotüüp E4_DM2COMA, mille kontrollidest välistamise lahtris on fenotüüp DM_OTHER_WIDE, mille RHK-10 kood nii surmaregistri kui haigla lahtris on vigane. Sama tulemuseni jõuab II tüüpi diabeedi fenotüüp.

Programmi testimise käigus tuli välja ka erisus surmapõhjuste registri RHK-10 koodides ja haigla RHK-10 koodides. Algselt luges programm isiku RHK-10 koodi sobivaks olenemata sellest, mis registrist koodi kohta informatsioon on andmebaasi lisatud. Testimise käigus leidis töö autor, et see ei ole korrektne lahendus ning tegi paranduse, et COD_ICD_10 tulbast võetud RHK-10 koodijada puhul vaadeldakse ainult isikute surma põhjuste registrist võetud kandeid ning HD_ICD_10 tulbast võetud RHK-10 koodijada puhul vaadeldakse kandeid kõikidest registritest peale surma põhjuste registri. Sarnane loogika tehti ka COD_ICD_10_EXCL ja HD_ICD_10_EXCL tulpade RHK-10 koodijadade puhul. Lisaks tehti täiendus, et surma põhjuste registri nimi võib isikute haiguslugudes olla kirjutatud kas eesti või inglise keeles. See on vajalik seetõttu, et Eesti Geenivaramu andmebaasi sisse logides valitud keelest muutuvad ka andmefailide väljade väärtused vastavalt.

Selle muudatuse järgselt uuesti testides ilmnes, et C3_BREAST ja C3_PROSTATE fenotüüpidele juhte leiti vähem kui varasemalt. Põhjuse uurimisel selgus, et FinnGen fenotüüpide defineerimise failis on mõlematel nimetatud fenotüüpidel RHK-10 kood ainult COD_ICD_10 lahtris ning HD_ICD_10 lahter on tühi. See tähendab, et inimene sobib juhuks ainult siis, kui tema RHK-10 kood on pärit surmaregistrist, mistõttu on loogiline, et juhte leiti nüüd vähem. Lisaks on nende puhul kasutatud RHK-O-3 vähi morfoloogia ja topograafia koode, mida programm hetkel kasutada ei võimalda.

Seega on oluline jälgida, kuidas on fenotüüp defineeritud. Kui soovida leida kõik rinnavähiga või eesnäärmevähiga juhud, olenemata millisest registrist RHK-10 kood on pärit, siis tuleks

leida kas FinnGen fenotüüpide failist selline fenotüüp, mis on vastavalt defineeritud või ise sellised fenotüübid defineerida. FinnGen failis on olemas näiteks C3_BREAST_EXALLC ja C3_PROSTATE_EXALLC fenotüübid, mille puhul on RHK-10 koodid nii COD_ICD_10 kui HD_ICD_10 lahtrites ja seega neid kasutades leiaks programm juhud kõikidest registritest pärit RHK-10 koodide alusel. Samas nende fenotüüpide definitsioonid välistavad kontrollide hulgast kõik erinevate vähkidega inimesed, mida C3_BREAST ega C3_PROSTATE fenotüüpide puhul ei tehta.

Hilisemates testides suurendati fenotüüpide arvu ning lisaks testiti registrite ja arsti liigi järgi juhtude filtreerimist. Registrite puhul katsetati nii ühe registriga filtreerimist (Haigekassa) kui mitme registriga filtreerimist (E-Tervis|PERH|TÜ Kliinikum). Tulemuste failide võrdlemisel oli näha, et juhtude ja kontrollide arvud muutusid ehk filtreerimine toimus. Samuti testiti arsti liigi puhul nii perearsti kui eriarsti poolt diagnoositud juhtude filtreerimist, mille puhul oli samuti tulemuste failides näha, et juhtude arvud muutusid vastavalt. Tabelis 8 on näidatud võrdlus programmi poolt leitud juhtude arvudes arsti liigi järgi filtreerides. Iga fenotüübi kohta on toodud, mitu juhtu leiti Eesti Geenivaramu andmestikust kokku ning mitu juhtu leiti siis, kui arvestati vaid perearsti/-õe poolt pandud diagnoose või ainult eriarsti/-õe poolt pandud diagnoose. Samuti on toodud ülekate koguandmete osas, perearsti/-õe ja eriarsti/-õe poolt diagnoositud juhtude ülekatete arv ja osakaal.

Tabel 8. Programmi poolt leitud juhtude arvude võrdlus arsti liigi järgi filtreerides.

Fenotüübi nimi	Kirjeldus	Kõik juhud	Perearst /-õde juhud	Eriarst /-õde juhud	Ülekate kogu-andmete osas	Ülekate juhtude osas	Ülekatte osakaal kogu-juhtudest
T2D	Teist tüüpi diabeet	26208	26949	25771	98.98%	24658	94.09%
GOUT	Podagra	14941	13138	7792	96.02%	6134	41.05%
I9_CHD	Südame isheemia-tõbi	9434	4886	6633	98.11%	3759	39.85%
I9_HYPTENS	Kõrgvere-rõhutõbi	82670	76180	68715	90.71%	62607	75.73%
K11_APPENDACUT	Pimesoole-põletik	15781	5888	8811	97.51%	4627	29.32%

Andmestike võrdlemisel on näha, et teist tüüpi diabeedi puhul on haigust diagnoositud nii perearstide kui ka eriarstide poolt ning juhud kattuvad sõltumata allika valikust ~94% ulatuses. Kusjuures arsti liigi järgi filtreerimisel peab arvestama, et kui fenotüübil on defineeritud välistused (teist tüüpi diabeedi puhul esimest tüüpi diabeedi diagnoosiga inimesed), siis osad välistuse diagnoosid võivad ära kaduda teise arsti liigi tõttu (kui esimest tüüpi diabeet on isikul diagnoositud eriarsti poolt, aga arsti liigiks määratakse perearst, siis isikut ei välistata teist tüüpi diabeedi juhtudest). Tulemusena võib programm leida arsti liigi filtrit kasutamata väiksema arvu juhte kui arsti liigi järgi filtreerides (nagu tabelis 8 teist tüüpi diabeedi korral, kus juhtude koguarv on väiksem kui perearsti/-õe poolt diagnoositud juhtude arv).

Podagra puhul on juhud ülekattes ~41% juhtudest ning suurem osa juhtudest pärineb perearstidelt või pereõdedelt, mis arvestades haiguse iseloomu on oodatav – podagra puhul manageerib haiguse kulgu pikka aega perearstisüsteem. Südame isheemiatõve korral kattuvad juhud umbes 40% ulatuses, suurem arv diagnoose pärineb eriarstidelt/-õdedelt. See on ka oodatav, sest I9_CHD puhul otsitakse suur osa diagnoose vaid surmapõhjuste registrist (kus diagnoos ei saa pärineda perearstilt) ning teistest registritest otsitakse peamiselt esmase või sekundaarse infarkti diagnoosi. Kõrgvererõhu tõve puhul kattuvad juhud ~75% ulatuses ning veidi rohkem esineb diagnoose perearstidelt. Pimesoolepõletiku puhul on rohkem diagnoose pärit eriarstidelt kui perearstidelt. Üllatava tulemusena on pimesoolepõletiku puhul kattuvaid juhte vähe (~29%). Spekuleerida võiks, et perearstide poolt pandud diagnoos võib osutada kahtlusele, aga eriarstile suunates diagnoos ei kinnitunud, kuid seda peaks põhjalikumalt uurima.

Magistritöö juhendaja võrdles programmi katsetuste käigus erinevatele fenotüüpidele leitud juhtude ja kontrollide arve enda kirjutatud skriptiga saadud juhtude ja kontrollide arvudega. Fenotüübile, kus sooviti leida II tüüpi diabeedi juhud, kelle seast on välistatud isikud, kellel on ka I tüüpi diabeet (HD_ICD_10 ja COD_ICD_10 lahtrites on kirjas E11 ning HD_ICD_10_EXCL ja COD_ICD_10_EXCL lahtrites on kirjas E10) leidis magistritöö käigus valminud programm Eesti Geenivaramu andmestikust 15837 juhtu ning täpselt sama arvu leidis ka juhendaja enda poolt koostatud skriptidega. Samuti kattusid juhtude arvud ka teiste võrreldud fenotüüpide puhul – südame isheemiatõve (I9_CHD) korral leiti nii töö autori poolt loodud programmiga kui juhendaja poolt loodud skriptiga 9434 juhtu, rinnavähi (C3_BREAST_EXALLC) korral 3771 juhtu ja eesnäärmevähi (C3_PROSTATE_EXALLC) korral 3206 juhtu. Seega näitasid tulemused, et programm suudab juhud õigesti leida.

Programmi katsetamistel tekkinud tulemusfailidest sai näha ka seda, kui kaua programmi töö aega võtab. Näiteks 15 valitud fenotüübile juhtude ja kontrollide leidmiseks kulus ligikaudu 3 tundi, kuid arvestama peab, et fenotüüpidel on ka eeldused ning sel juhul oli 15 soovitud fenotüübi juhtude ja kontrollide leidmiseks vaja tegelikult arvutada juhud 308 fenotüübile. Samuti näeb tulemuste failidest, et kõige kauem võtab filtreerimise sammudest aega esimeste haiguslugude järgi filtreerimine, kuid seda on vajalik teha ainult mõned korrad üle kõikide fenotüüpide, samas kui teised filtreerimise sammud alates tingimuste filtreerimisest tehakse iga fenotüübi jaoks eraldi. Lisaks peab arvestama, et programmi tööks kulunud aeg sõltub valitud fenotüüpide hulgast ja nende definitsiooni keerukusest. Samuti peab arvestama, et programmi tööaeg ei suurene lineaarselt, sest esmaste filtreerimise sammude (sugu, minimaalsete haigusjuhtude arv, register, arsti liik ja esimeste haiguslugude filtreerimine) töötlus tehakse programmi alguses olenemata arvutatavate fenotüüpide arvust. Seega mida rohkematele fenotüüpidele korraga juhte ja kontrolle otsitakse, seda efektiivsemalt programm töötab. Näiteks autori arvutis genereeritud 2000 isiku testandmetega katsetades võttis 10 fenotüübi juhtude ja kontrollide leidmine kokku aega 204 sekundit ehk 20,4 sekundit ühe fenotüübi kohta ning 50 fenotüübi puhul 676 sekundit ehk ainult 13,5 sekundit ühe fenotüübi kohta ehk 33,7% kiiremini.

2.3.3 Tehtud töö mõju analüüsides

Fenotüüpide defineerimine omab mõju kõikidele tehtavatele analüüsidele (Burstein et al., 2023). Nagu tabelist 8 näha, on ainuüksi arsti eriala järgi filtreerimine tugevalt mõjutav, milline inimene analüüsis oleks juht ja milline kontroll. Osade diagnooside puhul on eriarsti poole pöördumine vältimatu (näiteks vähid), samas kui teiste haiguste puhul jõutakse eriarsti vastuvõtule alles siis, kui haiguse olemus on muutunud raskemaks. Erinevad biopangad kasutavad eri registritest või arstidelt pärit diagnoose – näiteks Ühendkuningriigi Biopangas on RHK koodidena olemas vaid haiglasse võetud patsientide andmed (*UK Biobank*, 2025) ning FinnGen projekti enamik GWAS uuringuid süstemaatiliselt jätavad esmatasandi diagnoosid välja ning kasutavad vaid haiglast pärit (nii eriarstide vastuvõttude kui haiglasse võetud patsientide visiitide) informatsiooni koos teiste riiklike registritega (*FinnGen*, 2025). Eesti Geenivaramu kasutab enamasti kõikidest registritest pärit andmeid ning esmatasandi diagnoose süstemaatiliselt ei välista.

Hiljutised uuringud on näidanud, et heterogeensus fenotüüpide defineerimisel omab mõju nii GWASi tulemustes kui ka polügeensete riskiskooride uuringutes (Monti et al., 2023) – näiteks on näidatud, et ilma esmatasandi diagnoosideta on mitmete haiguste riskiskooride efektid

suuremad kui siis, kui esma- ja teise tasandi diagnoosid kombineerida (Jermy et al., 2024). Sõltuvalt uuringu eesmärgist ei pruugi olla alati halb idee vaid ekstreemsemaid juhte uuringus juhtudena käsitleda, kuid osade haiguste puhul (näiteks podagra) tekitab see tugevat väärklassifitseerimist juhtude ja kontrollide määramisel.

2.3.4 Programmi edasiarenduse võimalused

Tulevikus saaks programmi edasi arendada nii, et see võtaks arvesse ka vähispetsiifilised RHK-O-3 koodid ja kirurgiliste protseduuride koodid. RHK-O-3 koodidega klassifitseeritakse kasvajaid topograafia ehk kasvaja asukoha ja morfoloogia põhjal. RHK-O-3 ja RHK-10 koodid erinevad struktuuri poolest – RHK-O-3 koodid kirjeldavad täpsemalt kasvaja histoloogilist tüüpi. Näiteks RHK-10 klassifikatsiooni järgi ei ole võimalik eristada kopsu adenokartsinoomi ja kopsu lamerakulist kartsinoomi, mõlema RHK-10 kood on C34.9, samal ajal kui RHK-O-3 klassifikatsioonis topograafia kood jääb küll samaks, aga lisandub morfoloogia kood ja viimase numbrina kasvaja iseloom (kas healoomuline, halvaloomuline, primaarne, metastaatiline või muud tüüpi). Seega näiteks kopsu adenokartsinoom RHK-O-3 klassifikatsiooni järgi on tähistatud koodiga C34.9, 8140/3 ja kopsu lamerakuline kartsinoom koodiga C34.9, 8070/3 (Allen, 2013). FinnGen fenotüüpide defineerimise failis on RHK-O-3 jagatud kolme veergu – topograafia kood (CANC_TOPO), morfoloogia kood (CANC_MORPH) ja kasvaja iseloomu kood (CANC_BEHAV). Programmi edasiarenduses saaks implementeerida neist veergudest väärtustele vastavate doonorite filtreerimise Eesti Geenivaramu andmestikest. Sarnaselt saaks implementeerida kirurgiliste protseduuride koodide (FinnGen fenotüüpide defineerimise failis veerus OPER_NOM) arvestamise juhtude ja kontrollide leidmisel.

Kui Eestis võetakse kasutusele RHK-11 koodid, siis saaks programmile täienduse teha ka nende töötlemiseks. Hetkel kasutab programm ainult RHK-10 koode, sest Eesti Geenivaramu andmestikes on samuti kasutusel ainult RHK-10 koodid. RHK-11 koodid on täiendatud versioon RHK-10 koodidest ning nende vorming on teistsugune. Programmi täiendamisel saaks nende vormingu arvesse võtta, koostada RHK-10 koodide töötlemise eeskujul regulaaravaldised, mille abil leida doonorite haiguslugude seast vasteid.

Edasiarendusena saaks uurida võimalusi, kuidas muuta programm efektiivsemaks nii mälukasutuse kui ajakasutuse osas. Saaks vähendada kasutatavate andmete hulka hilisemates filtreerimise sammudes, muuta ja täiendada aeglasemaid kohti rakenduses, et neid kiiremaks muuta ning ka lisada paralleelarvutuse (*parallel computing*) võimekus. See tähendab, et programm võiks suuta kasutada mitut lõime (*multithreading*) või protsessorit

(*multiprocessing*) samaaegselt ja iga lõim või protsessor saaks paralleelselt teha arvutusi ilma, et peaks ootama eelmise arvutuse tulemusi.

Samuti saaks uurida, kas oleks võimalik kasutada programmi sisendina mõnda teise formaadiga fenotüüpide definitsioonide faili. Kuna FinnGen fenotüüpide definitsioonide failid on kõigile kättesaadavad, siis loodi programm selle faili struktuuri arvestades, kuid fenotüüpide definitsioone pakuvad mingil määral ka teised biopangad. Samas tuleb arvestada, et need ei pruugi olla vabalt kättesaadavad nagu FinnGen failid ning neis võib olla kasutatud teistsuguseid klassifitseerimissüsteeme ja andmevälju.

KOKKUVÕTE

Magistritöö eesmärgiks oli luua automatiseeritud töövoog juhtude ja kontrollide määramiseks Eesti Geenivaramu geenidoonorite seast vastavalt fenotüüpide definitsioonidele. Selleks tutvuti FinnGen fenotüüpide raamatukoguga ja FinnGen projekti rakendusega Risteys, et aru saada, kuidas komplekshaigusi defineeritakse ning kuidas selle töövoog juhtude ja kontrollide leidmiseks on üles ehitatud. Samuti uuriti, milliseid erinevaid kodeerimise süsteeme on olemas, kuidas neid omavahel harmoniseeritakse ning miks on tähtis erinevate biopankade andmete ühildamine ülegenoomsetel assotsiatsiooniuuringutel.

Töö tulemusel valmis programm, mille abil on võimalik leida fenotüüpide definitsioonidele vastavad juhud ja kontrollid geenidoonorite seast. Programm kasutab sisendina fenotüüpide definitsioonide faili ja Eesti Geenivaramu andmefaile. Programmi töö tulemusena salvestatakse maatriks, milles on iga otsitava fenotüübi kohta märgitud juhud ja kontrollid ning juhtude puhul ka isikute aeg haigestumiseni ja esmase diagnoosi kuupäev, kontrollide puhul nende kõige viimase diagnoosi kuupäev. Samuti on maatriksis märgitud iga doonori kohta, kas doonoril on kanne surmaregistris või mitte. Lisaks luuakse tulemuste fail, kus on iga fenotüübi kohta kirjas mitu juhtu ja kontrolli leiti, mitu doonorit oli kokku, kui kaua programmi erinevad etapid keskmiselt aega võtsid, mis probleeme programmi töös esines ja iga doonorite failis esineva registri kohta kõige uuema diagnoosi kuupäev. Täpsem informatsioon programmi iga tehtud sammu kohta kajastub logide failis, mis samuti programmi töö käigus luuakse.

Valminud programmi testiti Eesti Geenivaramu andmestikega ning tehti vajalikud parandused ja täiendused, et seda oleks võimalikult mugav kasutada. Lõpptulemusena valmis programm, mis võiks pakkuda täpset ja mugavat võimalust suure arvu fenotüüpide definitsioonidele vastavate juhtude ja kontrollide leidmiseks, mida hiljem sisendina kasutada kõikvõimalikeks analüüsideks, sealhulgas ülegenoomsete assotsiatsiooniuuringute või polügeensete riskiskooride arendamise läbiviimiseks.

Workflow for defining phenotypes in the Estonian Biobank dataset

Eret Orumaa

SUMMARY

Complex diseases are influenced by genetic variation and environmental factors and are characterized by a combination of multiple traits. They can be studied using genome-wide association studies, which try to find statistical associations between phenotypes and genetic variants. When studying binary traits individuals are categorized as cases or controls based on their phenotype and their genetic differences are compared.

These studies use biological samples and health-related data which are stored in biobanks. To improve the statistical power of the studies, data from various biobanks can be combined through meta-analyses, which require the data to be harmonized and the phenotypes to be uniformly defined. Phenotypes are defined using requirements for data origin, clinical codes for diagnoses, procedures and medications and other health related data. Accurate phenotype definitions are essential for selecting the right groups of individuals for the study as cases and controls to ensure the reproducibility and reliability of the study.

Complex disease phenotype libraries have been created to host curated phenotypes. One well-known library is the Finnish Biobank project FinnGen's clinical phenotype library, which provides continuously updated phenotype definitions and a tool for exploring case-control identification workflows.

The aim of this master's thesis was to develop an automated workflow for identifying suitable cases and controls among Estonian Biobank donors based on predefined phenotype definitions. To achieve this, the FinnGen clinical phenotype library was examined to understand how complex diseases are defined and how their workflow for identifying cases and controls is structured. Additionally, the author explored different medical coding systems and the way they are harmonized and why the harmonization of data from various biobanks is important in genome-wide association studies.

The work resulted in an application that uses a phenotype definition file and Estonian Biobank data files as input. The output of the program is a matrix where cases and controls are identified for each phenotype. For cases the time to disease onset and the date of first diagnosis are also included in the matrix and for controls the most recent diagnosis date is displayed. The matrix also includes whether each donor has a record in the death registry.

In addition, a statistics file is generated, listing the number of identified cases and controls, the total number of donors, the average duration of each processing step, issues encountered during the program execution, and the most recent diagnosis date for each registry present in the donor files. More detailed information about each step of the program is recorded in a log file.

The program was tested using datasets from the Estonian Biobank and bug fixes and necessary improvements were made according to the feedback. The final result is a tool that provides a precise and convenient way to identify phenotype specific cases and controls for conducting genome-wide association studies.

The developed program can be further improved by adding functionality to identify cases based on surgical procedure codes as well as cancer specific ICD-O-3 codes. If Estonia transitions to using ICD-11 in the future, the program can be adapted to accommodate this classification system as well. Additionally, there is potential to optimize the program for better speed and efficiency and exploring alternative input file formats could also broaden its usability.

KASUTATUD KIRJANDUS

Allen, P. W. (2013). ICDO — International Classification of Diseases for Oncology. *Pathology*, *23*(3), 280. https://doi.org/10.1016/S0031-3025(16)36112-8

Annaratone, L., De Palma, G., Bonizzi, G., Sapino, A., Botti, G., Berrino, E., Mannelli, C., Arcella, P., Di Martino, S., Steffan, A., Daidone, M. G., Canzonieri, V., Parodi, B., Paradiso, A. V., Barberis, M., & Marchiò, C. (2021). Basic principles of biobanking: From biological samples to precision medicine for patients. *Virchows Archiv*, *479*(2), 233–246. https://doi.org/10.1007/s00428-021-03151-0

Batheja, D., Goel, S., Fransman, W., Mantsoki, A., Ongarello, S., & Laxminarayan, R. (2023). Understanding the value of biobank attributes to researchers using a conjoint experiment. *Scientific Reports*, *13*(1), 22728. https://doi.org/10.1038/s41598-023-49394-6

Bennett, S. N., Caporaso, N., Fitzpatrick, A. L., Agrawal, A., Barnes, K., Boyd, H. A., Cornelis, M. C., Hansel, N. N., Heiss, G., Heit, J. A., Kang, J. H., Kittner, S. J., Kraft, P., Lowe, W., Marazita, M. L., Monroe, K. R., Pasquale, L. R., Ramos, E. M., van Dam, R. M., … Williams, K. (2011). Phenotype harmonization and cross-study collaboration in GWAS consortia: The GENEVA experience. *Genetic epidemiology*, *35*(3), 159–173. https://doi.org/10.1002/gepi.20564

Bouttle, K., Ingold, N., & O'Mara, T. A. (2024). Using Genetics to Investigate Relationships between Phenotypes: Application to Endometrial Cancer. *Genes*, *15*(7), 939. https://doi.org/10.3390/genes15070939

Burstein, D., Hoffman, G., Mathur, D., Venkatesh, S., Therrien, K., Fanous, A. H., Bigdeli, T. B., Harvey, P. D., Roussos, P., & Voloudakis, G. (2023). Detecting and Adjusting for Hidden Biases due to Phenotype Misclassification in Genome-Wide Association Studies. *medRxiv: The Preprint Server for Health Sciences*, 2023.01.17.23284670. https://doi.org/10.1101/2023.01.17.23284670

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics

to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00424

Chapman, M., Mumtaz, S., Rasmussen, L. V., Karwath, A., Gkoutos, G. V., Gao, C., Thayer, D., Pacheco, J. A., Parkinson, H., Richesson, R. L., Jefferson, E., Denaxas, S., & Curcin, V. (2021). Desiderata for the development of next-generation electronic health record phenotype libraries. *GigaScience*, *10*(9), giab059. https://doi.org/10.1093/gigascience/giab059

Colwill, M., Baillie, S., Pollok, R., & Poullis, A. (2024). Biobanks and biomarkers: Their current and future role in biomedical research. *World Journal of Methodology*, *14*(4), 91387. https://doi.org/10.5662/wjm.v14.i4.91387

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., Cooper, L. D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T. A., Dettai, A., Diogo, R., … Mabee, P. (2015). Finding Our Way through Phenotypes. *PLoS Biology*, *13*(1), e1002033. https://doi.org/10.1371/journal.pbio.1002033

Denaxas, S., Liu, G., Feng, Q., Fatemifar, G., Bastarache, L., Kerchberger, Eric. V., Hingorani, A. D., Lumbers, T., Peterson, J. F., Wei, W.-Q., & Hemingway, H. (2022). Mapping the Read2/CTV3 controlled clinical terminologies to Phecodes in UK Biobank primary care electronic health records: Implementation and evaluation. *AMIA Annual Symposium Proceedings*, *2021*, 362–371.

Dey, R., Zhou, W., Kiiskinen, T., Havulinna, A., Elliott, A., Karjalainen, J., Kurki, M., Qin, A., Lee, S., Palotie, A., Neale, B., Daly, M., & Lin, X. (2022). Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature Communications*, *13*(1), 5437. https://doi.org/10.1038/s41467-022-32885-x

Faro, V. L., Bhattacharya, A., Zhou, W., Zhou, D., Wang, Y., Läll, K., Kanai, M., Lopera-Maya, E., Straub, P., Pawar, P., Tao, R., Zhong, X., Namba, S., Initiative, the G. B. M., Sanna, S., Nolte, I. M., Okada, Y., Ingold, N., MacGregor, S., … Hirbo, J. (2022). *Genome-wide association meta-analysis identifies novel ancestry-specific primary*

open-angle glaucoma loci and shared biology with vascular mechanisms and cell proliferation (lk 2021.12.16.21267891). medRxiv. https://doi.org/10.1101/2021.12.16.21267891

Friedman, N. P., Banich, M. T., & Keller, M. C. (2021). Twin studies to GWAS: There and back again. *Trends in cognitive sciences*, *25*(10), 855–869. https://doi.org/10.1016/j.tics.2021.06.007

Fung, K. W., Xu, Julia, & Bodenreider, O. (2020). The new International Classification of Diseases 11th edition: A comparative analysis with ICD-10 and ICD-10-CM. *Journal of the American Medical Informatics Association : JAMIA*, *27*(5), 738–746. https://doi.org/10.1093/jamia/ocaa030

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., … O'Leary, T. J. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, *70*, 214–223. https://doi.org/10.1016/j.jclinepi.2015.09.016

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., … Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444. https://doi.org/10.1038/ng.3247

Hirsch, J. A., Nicola, G., McGinty, G., Liu, R. W., Barr, R. M., Chittle, M. D., & Manchikanti, L. (2016). ICD-10: History and Context. *AJNR: American Journal of Neuroradiology*, *37*(4), 596–599. https://doi.org/10.3174/ajnr.A4696

Jermy, B., Läll, K., Wolford, B. N., Wang, Y., Zguro, K., Cheng, Y., Kanai, M., Kanoni, S., Yang, Z., Hartonen, T., Monti, R., Wanner, J., Youssef, O., Lippert, C., van Heel, D., Okada, Y., McCartney, D. L., Hayward, C., Marioni, R. E., … Ripatti, S. (2024). A unified framework for estimating country-specific cumulative incidence for 18

diseases stratified by polygenic risk. *Nature Communications*, *15*(1), 5007. https://doi.org/10.1038/s41467-024-48938-2

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., … Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, *613*(7944), 508–518. https://doi.org/10.1038/s41586-022-05473-8

Lazareva, T. E., Barbitoff, Y. A., Changalidis, A. I., Tkachenko, A. A., Maksiutenko, E. M., Nasykhova, Y. A., & Glotov, A. S. (2022). Biobanking as a Tool for Genomic Research: From Allele Frequencies to Cross-Ancestry Association Studies. *Journal of Personalized Medicine*, *12*(12), Article 12. https://doi.org/10.3390/jpm12122040

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P. C., Mägi, R., Milani, L., Fischer, K., & Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*, *44*(4), 1137–1147. https://doi.org/10.1093/ije/dyt268

Linder, J. E., Bastarache, L., Hughey, J. J., & Peterson, J. F. (2021). The Role of Electronic Health Records in Advancing Genomic Medicine. *Annual Review of Genomics and Human Genetics*, *22*(Volume 22, 2021), 219–238. https://doi.org/10.1146/annurev-genom-121120-125204

Liu, S., Zheng, H., Gu, Y., Yang, Z., Liu, Y., Wei, Y., Guo, X., Chen, Y., Hu, L., Chen, X., Zhang, F., Chen, G.-B., Qiu, X., Huang, S., Zhen, J., & Wei, F. (2024). *The Dynamic Genetic Atlas of 122 Gestational Phenotypes* (lk 2024.10.15.24315491). medRxiv. https://doi.org/10.1101/2024.10.15.24315491

López-Cortegano, E., & Caballero, A. (2019). GWEHS: A Genome-Wide Effect Sizes and Heritability Screener. *Genes*, *10*(8), 558. https://doi.org/10.3390/genes10080558

Milani, L., Alver, M., Laur, S., Reisberg, S., Haller, T., Aasmets, O., Abner, E., Alavere, H., Allik, A., Annilo, T., Fischer, K., Hofmeister, R., Hudjashov, G., Jõeloo, M., Kals, M., Karo-Astover, L., Kasela, S., Kolde, A., Krebs, K., … Metspalu, A. (2025). The Estonian Biobank's journey from biobanking to personalized medicine. *Nature*

Communications, *16*(1), 3270. https://doi.org/10.1038/s41467-025-58465-3

Mills, M. C., & Rahal, C. (2019). A scientometric review of genome-wide association studies. *Communications Biology*, *2*(1), 1–11. https://doi.org/10.1038/s42003-018-0261-x

Monti, R., Eick, L., Hudjashov, G., Läll, K., Kanoni, S., Wolford, B. N., Wingfield, B., Pain, O., Wharrie, S., Jermy, B., McMahon, A., Hartonen, T., Heyne, H., Mars, N., Genes &amp, Team, H. R., Hveem, K., Inouye, M., Heel, D. A. van, … Lippert, C. (2023). *Evaluation of polygenic scoring methods in five biobanks reveals greater variability between biobanks than between methods and highlights benefits of ensemble learning* (lk 2023.11.20.23298215). medRxiv. https://doi.org/10.1101/2023.11.20.23298215

Næss, M., Kvaløy, K., Sørgjerd, E. P., Sætermo, K. S., Norøy, L., Røstad, A. H., Hammer, N., Altø, T. G., Vikdal, A. J., & Hveem, K. (2024). Data Resource Profile: The HUNT Biobank. *International Journal of Epidemiology*, *53*(3), dyae073. https://doi.org/10.1093/ije/dyae073

O'Donoghue, S., Dee, S., Byrne, J. A., & Watson, P. H. (2022). How Many Health Research Biobanks Are There? *Biopreservation and Biobanking*, *20*(3), 224–228. https://doi.org/10.1089/bio.2021.0063

Partanen, J. J., Häppölä, P., Zhou, W., Lehisto, A. A., Ainola, M., Sutinen, E., Allen, R. J., Stockwell, A. D., Oldham, J. M., Guillen-Guio, B., Flores, C., Noth, I., Yaspan, B. L., Jenkins, R. G., Wain, L. V., Ripatti, S., Pirinen, M., Initiative (GBMI), G. B. M., Kaarteenaho, R., … Koskela, J. T. (2021). *Leveraging global multi-ancestry meta-analysis in the study of Idiopathic Pulmonary Fibrosis genetics* (lk 2021.12.29.21268310). medRxiv. https://doi.org/10.1101/2021.12.29.21268310

Pedersen, E. M., Agerbo, E., Plana-Ripoll, O., Steinbach, J., Krebs, M. D., Hougaard, D. M., Werge, T., Nordentoft, M., Børglum, A. D., Musliner, K. L., Ganna, A., Schork, A. J., Mortensen, P. B., McGrath, J. J., Privé, F., & Vilhjálmsson, B. J. (2023). ADuLT: An efficient and robust time-to-event GWAS. *Nature Communications*, *14*(1), 5553. https://doi.org/10.1038/s41467-023-41210-z

Spataro, N., Rodríguez, J. A., Navarro, A., & Bosch, E. (2017). Properties of human disease

genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Human Molecular Genetics*, *26*(3), 489–500. https://doi.org/10.1093/hmg/ddw405

Surakka, I., Wu, K.-H., Hornsby, W., Wolford, B. N., Shen, F., Zhou, W., Huffman, J. E., Pandit, A., Hu, Y., Brumpton, B., Skogholt, A. H., Gabrielsen, M. E., Walters, R. G., The TOPMed Stroke Working Group, M. V. P. (MVP), Hveem, K., Kooperberg, C., Zöllner, S., Wilson, P. W. F., Sutton, N. R., … Initiative (GBMI), the G. B. M. (2022). *Multi-ancestry meta-analysis identifies 2 novel loci associated with ischemic stroke and reveals heterogeneity of effects between sexes and ancestries* (lk 2022.02.28.22271647). medRxiv. https://doi.org/10.1101/2022.02.28.22271647

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Thayer, D. S., Mumtaz, S., Elmessary, M. A., Scanlon, I., Zinnurov, A., Coldea, A.-I., Scanlon, J., Chapman, M., Curcin, V., John, A., DelPozo-Banos, M., Davies, H., Karwath, A., Gkoutos, G. V., Fitzpatrick, N. K., Quint, J. K., Varma, S., Milner, C., Oliveira, C., … Jefferson, E. (2024). Creating a next-generation phenotype library: The health data research UK Phenotype Library. *JAMIA Open*, *7*(2), ooae049. https://doi.org/10.1093/jamiaopen/ooae049

Tsuo, K., Zhou, W., Wang, Y., Kanai, M., Namba, S., Gupta, R., Majara, L., Nkambule, L. L., Morisaki, T., Okada, Y., Neale, B. M., Initiative, G. B. M., Daly, M. J., & Martin, A. R. (2022). *Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity* (lk 2021.11.30.21267108). medRxiv. https://doi.org/10.1101/2021.11.30.21267108

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J.

(2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wolford, B. N., Zhao, Y., Surakka, I., Wu, K.-H. H., Yu, X., Richter, C. E., Bhatta, L., Brumpton, B., Desch, K., Thibord, F., Klarin, D., Johnson, A. D., Trégouët, D.-A., Damrauer, S. M., Smith, N. L., Invent, Mvp, Faro, V. L., Tsuo, K., … Willer, C. J. (2022). *Multi-ancestry GWAS for venous thromboembolism identifies novel loci followed by experimental validation in zebrafish* (lk 2022.06.21.22276721). medRxiv. https://doi.org/10.1101/2022.06.21.22276721

Wu, K.-H. H., Douville, N. J., Konerman, M. C., Mathis, M. R., Hummel, S. L., Wolford, B. N., Surakka, I., Graham, S. E., Joo, H., Hirbo, J., Cox, N. J., Lee, S., Preuss, M., Loos, R. J. F., Daly, M. J., Neale, B. M., Zhou, W., Hornsby, W. E., Willer, C. J., & Initiative, on behalf of the G. B. M. (2021). *Polygenic risk score from a multi-ancestry GWAS uncovers susceptibility of heart failure* (lk 2021.12.06.21267389). medRxiv. https://doi.org/10.1101/2021.12.06.21267389

KASUTATUD VEEBIAADRESSID

Clinical endpoints | FinnGen. (2025). Salvestatud 10. aprill 2025, https://www.finngen.fi/en/researchers/clinical-endpoints

De Morgan's laws. (s.a.). Ada Computer Science. Salvestatud 24. mai 2025, https://adacomputerscience.org/concepts/boolean_de_morgans

deCode genetics. (2024). deCODE Genetics. Salvestatud 20. aprill 2025, https://www.decode.com/

Deng, W. B. F. D. (2021, juuli 18). *On Biostatistics and Clinical Trials: Imputation of partial dates for adverse events, concomitant medications, and disease diagnosis*. On Biostatistics and Clinical Trials. Salvestatud 12. aprill 2025, https://onbiostatistics.blogspot.com/2021/07/imputation-of-partial-dates-for-adverse.html

Diversity in Genomic Research. (s.a.). Salvestatud 15. detsember 2024, https://www.genome.gov/about-genomics/fact-sheets/Diversity-in-Genomic-Research

Endpoints. (2024). Salvestatud 20. aprill 2025, https://finngen.gitbook.io/documentation/methods/endpoints

FinnGen. (2025). Salvestatud 20. aprill 2025, https://www.finngen.fi/en

GBMI Home. (2024). Global Biobank Meta. Salvestatud 3. aprill 2025, https://www.globalbiobankmeta.org

GitHub. (2025). GitHub. Salvestatud 23. mai 2025, https://github.com/

International Classification of Diseases (ICD). (2025). Salvestatud 18. mai 2025, https://www.who.int/standards/classifications/classification-of-diseases

Lifelines biobank for researchers. (2025). Salvestatud 20. aprill 2025, https://www.lifelines-biobank.com/

Logging. (2025). Python Documentation. Salvestatud 20. aprill 2025, https://docs.python.org/3/library/logging.html

MVP Home. (2025). Salvestatud 20. aprill 2025, https://www.mvp.va.gov/pwa/

National Biopank of Korea. (s.a.). Salvestatud 20. aprill 2025,

https://biobank.nih.go.kr/eng/../main/main.html

Pandas documentation. (2024). Salvestatud 20. aprill 2025, https://pandas.pydata.org/docs/index.html

Regulaaravaldised—Pythoni materjalid 2025 dokumentatsioon. (2025). Salvestatud 23. märts 2025, https://pydoc.pages.taltech.ee/extra/regex.html

RHK ehk rahvusvaheline haiguste klassifikatsioon | Tervise Arengu Instituut. (2025). Salvestatud 3. märts 2025, https://www.tai.ee/et/instituudist/meditsiiniterminoloogia-kompetentsikeskus/who-klassifikaatorid-rfk-ja-rhk/rhk-ehk

Risteys. (2024). Salvestatud 20. aprill 2025, https://risteys.finngen.fi/

SNOMED CT. (2025). NHS England Digital. Salvestatud 3. märts 2025, https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct

UK Biobank. (2025). Salvestatud 20. aprill 2025, https://www.ukbiobank.ac.uk

Unittest. (2025). Python Documentation. Salvestatud 20. aprill 2025, https://docs.python.org/3/library/unittest.html

venv—Creation of virtual environments. (2025). Python Documentation. Salvestatud 25. mai 2025, https://docs.python.org/3/library/venv.html

LISAD

Lisa 1: Biopankade võrdlus

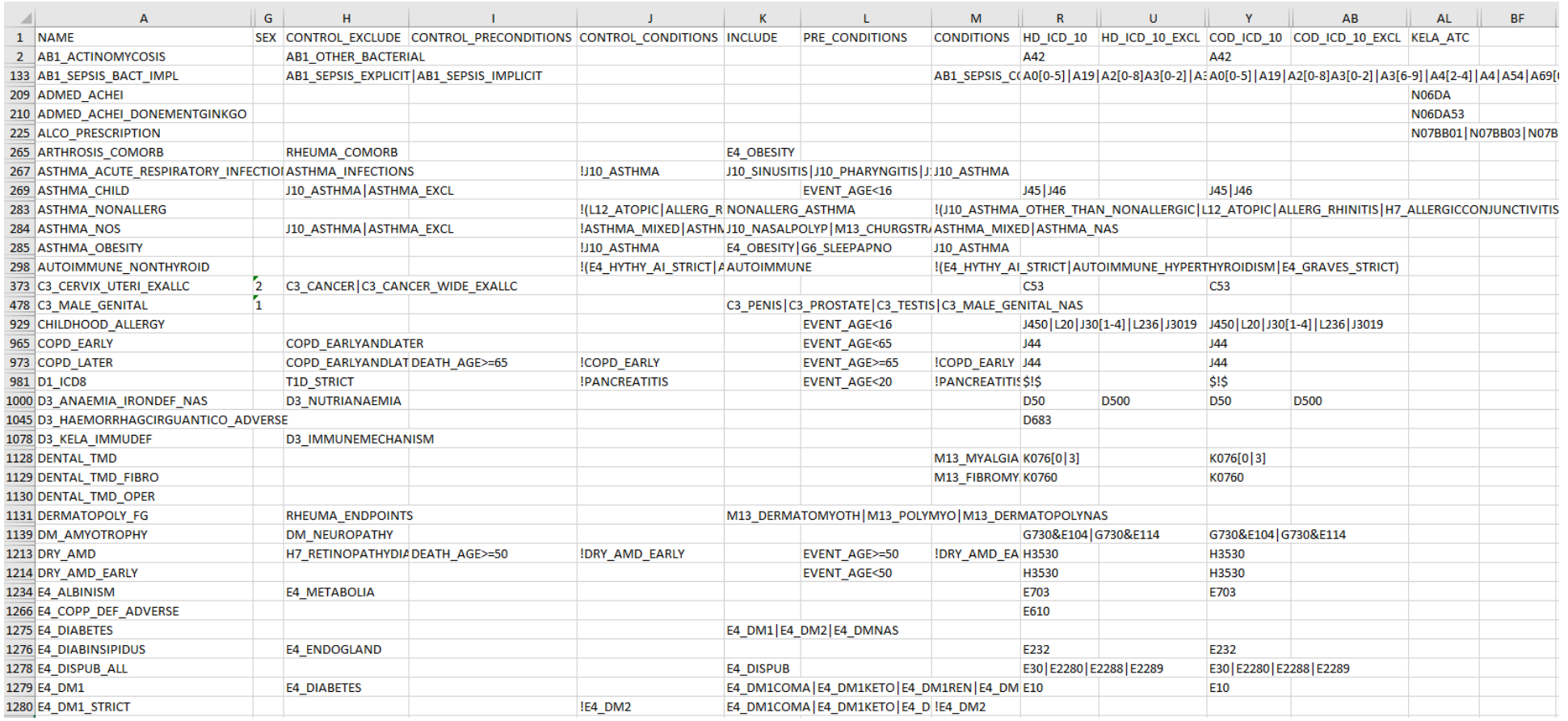
Tabel 9. Biopankade võrdlus (kohandatud tabelist Lazareva et al., 2022; lisainformatsioon biopankade kodulehtedelt)

Biopank	Asukoht	Osalejate hulk	Biopanga tüüp	Värbamine	Kohordi vanus	Oomika andmed	Viide
Ühendkuning-riigi biopank	Ühendkuning-riik	500 000	populatsioonipõhine	lõppenud	40-69	genotüpiseerimine, WGS, WES, metaboloomika, telomeeride pikkus	Lazareva et al., 2022
Jaapani biopank	Jaapan	267 307	patsiendipõhine	lõppenud	-	genotüpiseerimine, WGS, metaboloomika	Lazareva et al., 2022
FinnGen	Soome	538 600	populatsioonipõhine ja 15 haigusspetsiifilist kohorti	käimasolev	piiranguteta, keskmine vanus 53 aastat	genotüpiseerimine	Lazareva et al., 2022, *FinnGen*, 2025
Eesti Geenivaramu	Eesti	200 000	populatsioonipõhine	käimasolev	18+	WGS, WES, genotüpiseerimine, metaboloomika, RNA sekveneerimine, ülegenoomsed metülatsiooni kiibid, geeniekspressiooni kiibid	Lazareva et al., 2022

Hiina Kadoorie biopank	Hiina	512 891	populatsioonipõhine, osalejad 10 provintsist	lõppenud	30-79	genotüpiseerimine, WGS	Lazareva et al., 2022
Tohoku Medical Megabank projekt	Jaapan	157 000	populatsioonipõhine, osalejad 2 prefektuurist, 3 generatsiooni	lõppenud	18+	WGS, genotüpiseerimine, metaboloomika, metülatsiooni kiibid	Lazareva et al., 2022
Taiwani biopank	Taiwan	181 635	patsiendipõhine	lõppenud	20-70	genotüpiseerimine, WGS, DNA metülatsioon, HLA tüpiseerimine, metaboloomika	Lazareva et al., 2022
Lifelines Cohort Study	Holland	167 000	populatsioonipõhine, osalejad riigi põhja osast	lõppenud	lapsed, täiskasvanud, vanurid	genotüpiseerimine, WGS, mikrobioomi andmed	Lazareva et al., 2022, *Lifelines Biobank for Researchers*, 2025
Korea rahvuslik biopank	Lõuna- Korea	1 051 787	populatsiooni- ja patsiendipõhine	käimasolev	-	genotüpiseerimine	Lazareva et al., 2022, *National Biopank of Korea*, s.a.
Karolinska biopank	Rootsi	700 000	kollektsioonipõhine	-	-	genotüpiseerimine	Lazareva et al., 2022
HUNT biopank	Norra	120 000	populatsioonipõhine, osalejad 1 maakonnast	käimasolev	13-100	genotüpiseerimine	Lazareva et al., 2022, Næss et al., 2024

Kanada Partnership for Tomorrow's Health	Kanada	331 359	populatsioonipõhine, osalejad 9 provintsist	käimasolev	30-74	genotüpiseerimine	Lazareva et al., 2022
All of us Research Program	USA	348 000	populatsioonipõhine, vähemusrahvus	käimasolev	18+	WGS, genotüpiseerimine	Lazareva et al., 2022
BioVU	USA	275 000	patsiendipõhine	käimasolev	lapsed ja täiskasvanud	genotüpiseerimine	Lazareva et al., 2022
Penn Medicine biopank	USA	52 853	patsiendipõhine	käimasolev	18+	genotüpiseerimine, WES	Lazareva et al., 2022
Million Veteran Program	USA	1 000 000	populatsioonipõhine, USA sõjaväes teeninud veteranid	käimasolev	18+	genotüpiseerimine	*MVP Home*, 2025
deCODE	Island	160 000	populatsioonipõhine	-	18+	genotüpiseerimine, WGS	*deCode Genetics*, 2024

Lisa 2: FinnGen fenotüüpide definitsioonide fail



	A	G	H	I	J	K	L	M	R	U	Y	AB	AL	BF
1	NAME	SEX	CONTROL_EXCLUDE	CONTROL_PRECONDITIONS	CONTROL_CONDITIONS	INCLUDE	PRE_CONDITIONS	CONDITIONS	HD_ICD_10	HD_ICD_10_EXCL	COD_ICD_10	COD_ICD_10_EXCL	KELA_ATC	
2	AB1_ACTINOMYCOSIS		AB1_OTHER_BACTERIAL						A42		A42			
133	AB1_SEPSIS_BACT_IMPL		AB1_SEPSIS_EXPLICIT|AB1_SEPSIS_IMPLICIT					AB1_SEPSIS_C(	A0[0-5]|A19|A2[0-8]A3[0-2]|A3		A0[0-5]|A19|A2[0-8]A3[0-2]|A3[6-9]|A4[2-4]|A4|A54|A69[(			
209	ADMED_ACHEI												N06DA	
210	ADMED_ACHEI_DONEMENTGINKGO												N06DA53	
225	ALCO_PRESCRIPTION												N07BB01|N07BB03|N07B	
265	ARTHROSIS_COMORB		RHEUMA_COMORB			E4_OBESITY								
267	ASTHMA_ACUTE_RESPIRATORY_INFECTIOI		ASTHMA_INFECTIONS		!J10_ASTHMA	J10_SINUSITIS|J10_PHARYNGITIS|J		J10_ASTHMA						
269	ASTHMA_CHILD		J10_ASTHMA|ASTHMA_EXCL				EVENT_AGE<16		J45|J46		J45|J46			
283	ASTHMA_NONALLERG				!(L12_ATOPIC|ALLERG_R	NONALLERG_ASTHMA		!(J10_ASTHMA_OTHER_THAN_NONALLERGIC|L12_ATOPIC|ALLERG_RHINITIS|H7_ALLERGICCONJUNCTIVITIS						
284	ASTHMA_NOS		J10_ASTHMA|ASTHMA_EXCL		!ASTHMA_MIXED|ASTHN	J10_NASALPOLYP|M13_CHURGSTR/		ASTHMA_MIXED|ASTHMA_NAS						
285	ASTHMA_OBESITY				!J10_ASTHMA	E4_OBESITY|G6_SLEEPAPNO		J10_ASTHMA						
298	AUTOIMMUNE_NONTHYROID				!(E4_HYTHY_AI_STRICT|A	AUTOIMMUNE		!(E4_HYTHY_AI_STRICT|AUTOIMMUNE_HYPERTHYROIDISM|E4_GRAVES_STRICT)						
373	C3_CERVIX_UTERI_EXALLC	2	C3_CANCER|C3_CANCER_WIDE_EXALLC						C53		C53			
478	C3_MALE_GENITAL	1				C3_PENIS|C3_PROSTATE|C3_TESTIS|C3_MALE_GENITAL_NAS								
929	CHILDHOOD_ALLERGY						EVENT_AGE<16		J450|L20|J30[1-4]|L236|J3019		J450|L20|J30[1-4]|L236|J3019			
965	COPD_EARLY		COPD_EARLYANDLATER				EVENT_AGE<65		J44		J44			
973	COPD_LATER		COPD_EARLYANDLAT	DEATH_AGE>=65	!COPD_EARLY		EVENT_AGE>=65	!COPD_EARLY	J44		J44			
981	D1_ICD8		T1D_STRICT		!PANCREATITIS		EVENT_AGE<20	!PANCREATITIS	$!$		$!$			
1000	D3_ANAEMIA_IRONDEF_NAS		D3_NUTRIANAEMIA						D50	D500	D50	D500		
1045	D3_HAEMORRHAGCIRGUANTICO_ADVERSE								D683					
1078	D3_KELA_IMMUDEF		D3_IMMUNEMECHANISM											
1128	DENTAL_TMD							M13_MYALGIA	K076[0|3]		K076[0|3]			
1129	DENTAL_TMD_FIBRO							M13_FIBROMY	K0760		K0760			
1130	DENTAL_TMD_OPER													
1131	DERMATOPOLY_FG		RHEUMA_ENDPOINTS			M13_DERMATOMYOTH|M13_POLYMYO|M13_DERMATOPOLYNAS								
1139	DM_AMYOTROPHY		DM_NEUROPATHY						G730&E104|G730&E114		G730&E104|G730&E114			
1213	DRY_AMD		H7_RETINOPATHYDIA	DEATH_AGE>=50	!DRY_AMD_EARLY		EVENT_AGE>=50	!DRY_AMD_EA	H3530		H3530			
1214	DRY_AMD_EARLY						EVENT_AGE<50		H3530		H3530			
1234	E4_ALBINISM		E4_METABOLIA						E703		E703			
1266	E4_COPP_DEF_ADVERSE								E610					
1275	E4_DIABETES					E4_DM1|E4_DM2|E4_DMNAS								
1276	E4_DIABINSIPIDUS		E4_ENDOGLAND						E232		E232			
1278	E4_DISPUB_ALL					E4_DISPUB			E30|E2280|E2288|E2289		E30|E2280|E2288|E2289			
1279	E4_DM1		E4_DIABETES			E4_DM1COMA|E4_DM1KETO|E4_DM1REN|E4_DM			E10		E10			
1280	E4_DM1_STRICT				!E4_DM2	E4_DM1COMA|E4_DM1KETO|E4_D		!E4_DM2						

Joonis 9. Ekraanitõmmis FinnGen fenotüüpide definitsioonide failist *"FinnGen core and non-core endpoints and controls DF12".* Päisest on peidetud veerud, mida magistritöö käigus valminud programmi töös ei kasutata. Samuti on peidetud ridu, et võimalikult palju näidata erinevate veergude võimalikke väärtuseid.

Lisa 3: FinnGen fenotüüpide definitsioonide failis leitud vead ja erisused

Tabel 10. FinnGen fenotüüpide definitsioonide failis "*FinnGen core and non-core endpoints and controls DF12"* leitud vead ja erisused.

Fenotüübi nimi	Veerg	Originaalfaili lahtri sisu	Parandatud sisu
AB1_SEPSIS_BACT_IMPL	COD_ICD_10	A0[0-5]|A19|A2[0-8]A3[0-2]|A3[6-9]|A4[2-4]|A4|A54|A69[0-1]|A699|A70|A7[4-5]|A7[7-9]|B9[56]|G0[01]|J1[3-5]	Lisada püstkriips: A0[0-5]|A19|A2[0-8]|A3[0-2]|A3[6-9]|A4[2-4]|A4|A54|A69[0-1]|A699|A70|A7[4-5]|A7[7-9]|B9[56]|G0[01]|J1[3-5]
DM_OTHER_WIDE	COD_ICD_10	E13|14	Lisada kategooria täht: E13|**E**14
INFLUENZA	COD_ICD_10_EXCL	4870	Andmete nihe parandada, RHK-9 kood 4870 asendada sellel väljal olema pidanud RHK-10 koodijadaga: **J100|J110**
O15_POSTPART_HAEMORRH_ANATOMY	CONTROL_EXCLUDE	!(NUMBER_OF_OFFSPRING>0)|!(O15_DELIVERY|O15_DELIVERY_PROXY)	Lahtri sisu tõsta CONTROL_CONDITIONS veergu, muutes märgid vastavaks: NUMBER_OF_OFFSPRING>0**&**!O15_DELIVERY**&**!O15_DELIVERY_PROXY
O15_POSTPART_HAEMORRH_PLACENTA			
O15_POSTPART_HEAMORRH_LB			
O15_ICP_EXMORE O15_ICP_WIDE_EXMORE	CONTROL_EXCLUDE	!(O15_DELIVERY|O15_DELIVERY_PROXY)	Lahtri sisu tõsta CONTROL_CONDITIONS veergu, muutes märgid vastavaks: O15_DELIVERY**&**O15_DELIVERY_PROXY
I9_CHD_NOREV	INCLUDE	**K**.I9_MI	Eemaldada "K.": I9_MI

OFFICIAL_TILASTOKESKUS_DEATH	INCLUDE	K.ALCOHOLMENTAL|K.ALCOHOLACUTE10|K.ALCOHOLACUTE9|K.ALCONEURODEGEN|K.EPISEIZALCO|K.ALCOPOLYNEU|K.ALCOMYOP|K.CARMYOPATHALCO|K.ALCOGASTRITIS|K.ALCOLIVER|K.ALCOPANCACU|K.ALCOPANCCHRON|K.MATERNCAREFETUSALCO|K.P16_FET_NEWB_MATERN_USE_ALCOHOL|K.Q17_FETAL_ALCOHOL_SYNDR_DYSMORPHIC|K.VWXY20_ACCIDENTAL_POISO_EXPOS_ALCOHOL	Eemaldada "K.": ALCOHOLMENTAL|ALCOHOLACUTE10|ALCOHOLACUTE9|ALCONEURODEGEN|EPISEIZALCO|ALCOPOLYNEU|ALCOMYOP|CARMYOPATHALCO|ALCOGASTRITIS|ALCOLIVER|ALCOPANCACU|ALCOPANCCHRON|MATERNCAREFETUSALCO|P16_FET_NEWB_MATERN_USE_ALCOHOL|Q17_FETAL_ALCOHOL_SYNDR_DYSMORPHIC|VWXY20_ACCIDENTAL_POISO_EXPOS_ALCOHOL
Z21_ADDIT_CODES_LOCAT_DEFECT_INJURY_ILLNE	COD_ICD_10	Z[AB]	Eemaldada kõik kahe kategooria tähega koodidega fenotüüpide read. Kokku 20 fenotüüpi "ZA" ja ZB" koodidega. Tegemist ei ole otseselt veaga, vaid Soome-spetsiifiliste koodidega, mida Eestis ei kasutata ja seega ka programmi töös vastavatele fenotüüpidele juhte ja kontrolle leida ei ole võimalik.
Z21_ANKLE_FOOT		ZA70	
Z21_BILATERAL		ZB02	
Z21_DEFECT_INJURY_ILLNE_CHEST		ZA83	
Z21_DEFECT_INJURY_ILLNE_FACE		ZA81	
Z21_DEFECT_INJURY_ILLNE_LUMBAR_SPINE		ZA84	
Z21_DEFECT_INJURY_ILLNE_NECK		ZA82	

Z21_DEFECT_INJURY_ILLNE_SKULL		ZA80	
Z21_FOREARM		ZA30	
Z21_HAND		ZA40	
Z21_HIP_THIGH		ZA51	
Z21_LATERAL		ZB10	
Z21_LEFT_SIDE		ZB00	
Z21_LOWER_LEG		ZA60	
Z21_MEDIAL		ZB11	
Z21_MULTIPLE_SITES		ZA00	
Z21_PELVIC_REGION		ZA50	
Z21_RIGHT_SIDE		ZB01	
Z21_SHOUL_REGION		ZA10	
Z21_UPPER_ARM		ZA20	

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Eret Orumaa,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "**Töövoog fenotüüpide defineerimiseks Eesti Geenivaramu andmestikus"**, mille juhendajad on kaasprofessor Triin Laisk ja teadur Kristi Läll, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Eret Orumaa

29.05.2025