Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Viktoria Kirpu

Haigekassa kindlustamata patsientide vigastuste andmete imputeerimine

Bakalaureusetöö (9 EAP)

Juhendajad
Natalja Lepik, PhD
Natalja Eigo, MSc
(*Tervise Arengu Instituut*)

Tartu 2018

Haigekassa kindlustamata patsientide vigastuste andmete imputeerimine

Bakalaureusetöö

Viktoria Kirpu

Lühikokkuvõte. Töö eesmärk on Haigekassale saadetavate andmete täiendamine kasutades lisainformatsiooni allikana tervise infosüsteemi andmeid. Selleks on mõlema andmebaasi andmed omavahel ühendatud ning vastavalt sellele leitud paljudele Haigekassa andmebaasis ravikindlustuseta patsientide vanused. Vaadeldavat tunnust toovad välja tervise infosüsteemi epikriisid, kuid Haigekassa raviarvetel see info puudub. Nendele epikriisidele, kus patsientidele vanust ei õnnestunud leida, imputeeritakse puuduolevad andmed kolmel meetodil: üldine *Hot-Deck* omistus, lähima naabri meetod ja *Hot-Deck* omistus klassis kombineerituna lähima naabri meetodiga. Ühendamise protsessis suudeti leida vanused 5633 ravikindlustuseta patsiendi raviarvele ja vanuseta jäid 3515 raviarvet. Edasise analüüsi tulemuste põhjal otsustati kasutada üldise juhusliku *Hot-Deck* meetodiga saadud väärtusi, sest imputeerimiste simuleerimise katsel andis vaadeldav meetod kõige täpsemaid ja stabiilsemaid tulemusi.

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

Märksõnad: andmeanalüüs, statistiline andmetöötlus, puuduvad andmed, vaatlusvead, imputeerimine, *Hot-Deck* meetod, lähima naabri meetod.

The Estonian Health Insurance Fund's uninsured patients' injury data imputation

Bachelor's thesis

Viktoria Kirpu

Abstract. The objective of this bachelor's thesis is to supplement the data sent to The Estonian Health Insurance Fund's database using Health Information System's data as additional information. To achieve the objective, the data from both databases is linked and as a result a lot of ages of the uninsured patients are found to The Estonian Health Insurance Fund's database. This variable's values are only marked in the Health Information System's data and originally missing from The Estonian Health Insurance Fund's data. For those epicrisis, where the patient's age is still missing, the required variable is imputed with three different methods: general random *Hot-Deck* method, nearest neighbour method and random *Hot-Deck*

imputation within classes combined with nearest neighbour method. As the result of the data linking, there were found ages to 5633 patients and 3515 epicrisis remained without this variable's value. Based on the results of further analysis, it was decided to use the data imputed with general random *Hot-Deck* method, because in the imputation simulation this method gave the most precise and stable results.

CERCS research specialisation: P160 Statistics, operation research, programming, financial and acturial mathematics

Keywords: data processing, statistical data processing, missing data, observation errors, imputation, *Hot-Deck* method, nearest neighbour method.

Sisukord

SISSEJUHATUS

Eesti raviteenuse osutajate ülesanne on esitada ravijuhtude dokumentatsioone ja aruandeid kolme erinevasse süsteemi: Aveeb (Tervise Arengu Instituut), tervise infosüsteem ja Haigekassa. Selleks, et vähendada arstide töökoormust, on otsustatud statistika tegemiseks võtta kasutusele üks süsteem, milleks on tervise infosüsteem. Kahjuks on eelnimetatud süsteemis mitmeid puudusi, millest üheks on see, et arstid esitavad sinna märgatavalt vähem andmeid kui teistesse süsteemidesse.

Praegu esitatakse arstide poolt kõige rohkem ravijuhtude kohta andmeid Haigekassa andmebaasi. Seetõttu on ka selle bakalaureusetöö aluseks võetud just sellest süsteemist saadavad andmed. Kahjuks pole nende andmete põhjal võimalik teha täielikku demograafilist statistikat, kuna ravikindlustuseta patsientide epikriisidel pole märgitud inimese vanust. Töö eesmärgiks on Haigekassa andmebaasi andmete täiustamine kasutades lisainformatsioonina tervise infosüsteemi andmebaasi.

Kvaliteetse statistika tegemiseks on oluline koguda maksimaalselt kättesaadavat infot. Selleks on otsustatud ühendada omavahel Haigekassa ja tervise infosüsteemi andmed ja selle põhjal saada paljudele ravikindlustuseta patsientide epikriisidele ravitava vanused. Nendele ravijuhtudele, millele patsiendi vanust ei suudetud andmestike ühendamise protsessis leida, imputeeritakse vanused *Hot-Deck* meetoditega ja lähima naabri meetodiga.

Töö jaguneb neljaks peatükiks. Esimeses peatükis tutvustatakse lähemalt imputeerimise protsessi ja kasutatud meetodeid. Teises peatükis kirjeldatakse töös kasutatatud andmete edastamise süsteemide eripärasid. Töö kolmas, neljas ja viies peatükk on töö praktiline osa, kus kirjeldatakse andmete ühendamise ja imputeerimise protsessi. Kolmandas peatükis on välja toodud kõik erinevad kombinatsioonid, mida on andmete ühendamisel kasutusele võetud. Neljandas peatükis on imputeeritud analüüsis kasutatavatele andmetele puuduvad väärtused. Viiendas peatükis on läbi viidud imputeerimise katsed, mille põhjal on välja valitud parim imputeerimise meetod.

Töö praktilise osa läbiviimiseks on kasutatud andmete ühendamiseks rakendustarkvara STATA ja andmete imputeerimiseks rakendustarkvara R.

1 IMPUTEERIMINE

Mahukate uuringute korral on tihti probleemiks mittetäielikud andmed [1]. Puuduvate väärtustega andmed tekivad siis, kui valimisse sattunud subjekt ei vasta küsitlusele täielikus mahus ehk kui toimub nn **objekti** kadu (ingl k *unit non-response*) või kui küsitlus jääb osaliselt vastamata ehk kui toimub nn **tunnuse väärtuse** kadu (ingl k *item non-response*) [1, 7]. Kao kompenseerimiseks objekti tasemel kasutatakse tavaliselt kaalumismeetodeid, mis eeldavad taustinfo kasutamise võimalust (registrid, eelmised samalaadsed uuringud jms) [7]. Kõige laialdasemalt tunnuse väärtuse kao kompenseerimiseks kasutatav meetod on aga **imputeerimine**. Selle meetodi rakendamisel leitakse puuduvatele väärtustele hinnangud, et lõpptulemuseks saavutada täielikud andmed, mida saaks analüüsida traditsiooniliste analüüsimeetoditega. [1]

Tavaliselt viiakse valikuuringud läbi eesmärgiga leida rahvastikku kirjeldavad karakteristikud, näiteks keskmised, korrelatsioonid, regressioonikoefitsiendid. Seejuures üksikjuhtumite väärtused andmetes ei ole esmatähtsad. Lühidalt öeldes on imputeerimise eesmärk mitte niivõrd saada puuduvatele väärtustele parimad prognoosid, vaid asendada need piisavalt usaldusväärsete väärtustega, et lõplikult saadud täielik andmestik oleks rahvastikku kirjeldavate karakteristikute leidmiseks võimalikult tõepärane. [1]

1.1 Imputeerimise olulisus

Andmete puudumise tõttu ei kao ainult vajaminev informatsioon ega vähene uuringu võimsus, vaid tekivad nihkega hinnangud. Avastamata kadunud vaatluste mahtu ehk ravijuhtude arvu, mille kohta dokumentatsiooni ei esitatud tervise infosüsteemi ja kontrolli käigus nende puudumist ei avastatud, on tähtis minimiseerida. Vastasel korral statistilised järeldused, näiteks tulemuse usaldusintervall, on ilmselt vigane. Kaoga mitte arvestamine suurendab saadud hinnangute nihet. Kvaliteetse statistika jaoks on vajalik omada nihketa hinnangut või tuleks nihe muuta võimalikult väikseks. Mida väiksem on nihe, seda paremini peegeldavad statistilised tulemused reaalset olukorda. [5]

Näiteks jääb tervise infosüsteemis arstide poolt suuremal määral esitamata ravijuhu erakorralisuse tüüp. Kui tekib olukord, kus arstid jätavad erakorraliste ravijuhtude korral vaadeldavat tunnust märkimata, siis jääb mulje, et erakorralisi ravijuhte esineb meie riigis vähe. Sellises olukorras saame olla kindlad, et saadud statistika ei kirjelda tegelikkust ning oleme saanud nihkega hinnangud. Samal põ-

himõttel tekivad nihked ka siis, kui mõned epikriisid on koguni jäänud dokumenteerimata. [5]

Levinud on väärarusaam, et kui vastamismäär on kõrge, siis pole oluline arvestada andmete kaoga. Näiteks on valikuuringu läbiviimisel sageli aktiivsemad vastajad just vanemad inimesed ja nooremate seas on mittevastamise määr kõrgem. Vastajate arvu suurendamisel on aga võimalik saada valimisse veel rohkem vanemas eas inimesi ja selline olukord võib viia nihkega hinnanguteni, mis ei kirjelda üldkogumit. Seega ei oleks õige statistikas keskenduda vaid vastamismäärale kui indikaatorile, mis vähendab kaost põhjustatud nihet. Erinevalt hinnangu dispersioonist ei pruugi nihe läheneda valimimahu suurenedes nullile. Kaost põhjustatud nihke vähendamiseks on oluline kasutada vastavaid hindamismeetodeid. [5]

1.2 Doonoripõhised imputeerimismeetodid

Käesolevas bakalaureusetöös uuritakse praktikas väga levinud doonoripõhiseid imputeerimismeetodeid, st puuduvatele väärtustele omistatakse reaalselt eksisteerivad väärtused doonorgrupist, mis on mõne teise objekti puhul väärtuseks saadud. Sellise meetodi plussiks on see, et imputeeritud väärtus on ka reaalselt võimalik. [9]

1.3 *Hot-Deck* imputeerimismeetod

Väga levinud doonoripõhine imputeerimismeetod on *Hot-Deck* meetod, mille korral asendatakse iga puuduv väärtus antud objektiga sarnase objekti olemasoleva väärtusega [1]. *Hot-Deck* imputeerimise protseduuri korral tähistame elemendi k imputeeritud väärtust $\hat{y}_k = y_{l(k)}$, kus $l(k)$ on juhuslikult valitud doonor kõikvõimalikest doonorelementidest $l \in r_i$, kus r_i tähistab kõikvõimalike doonorelementide hulka. Meetodil on ka oma miinus: kuigi visuaalsel vaatlusel näeb imputeeritud tunnuse jaotus välja üsna loomulik, võib esineda imputeerimisnihe, kuna vastanud objektid võivad oluliselt erineda mittevastanud objektidest. [9]

1.4 Näiteid *Hot-Deck* imputeerimismeetoditest

Üldiselt eristatakse järgnevaid *Hot-Deck* meetodeid.

1. *Juhuslik* Hot-Deck *omistus klassis* on imputeerimismeetod, kus abitunnuse põhjal moodustatakse andmetest kõigepealt doonorgrupid ja seejärel puuduv tunnuseväärtus asendatakse vastavast doonorgrupist võetud olemasoleva väärtusega. Sageli on valik doonorgrupist tehtud juhuslikult. Abitunnuse rolli sobivad vaid sellised registritunnused, mille väärtused on teada kõikide valimiobjektide kohta (näiteks sugu, elukoht, vanuseklass jne). [8]
2. *Üldise juhusliku* Hot-Deck *omistuse* korral omistatakse puuduvale väärtusele kõikide vastanute seast juhuslikult valitud objekti väärtus [8]. Selle meetodi korral objekte gruppideks ei jagata ning tulemus on robustsem võrreldes eelmise variandiga.
3. *Järjestikune* Hot-Deck *omistus* on imputeerimismeetod, kus kõik valimi objektid järjestatakse tausttunnuse järgi. Puuduvale väärtusele omistatakse sellele järjekorras eelneva samasse klassi kuuluva objekti olemasolev väärtus. [8] Erinevalt 1. ja 2. meetodist on see deterministlik[1], mittejuhuslik omistuse meetod.

1.5 Doonorgrupi loomine

Eespool mainitud esimese meetodi jaoks tuleb kõigepealt luua mittekattuvaid imputeerimisgruppe ehk doonorgruppe (ingl k *Donor pools*) [1, 9]. Need imputeerimisgrupid moodustatakse kasutades abitunnuseid, mille väärtused on teada kõikide valimiobjektide jaoks [1]. Iga grupi sees rakendatakse puuduvate väärtuste leidmiseks sageli ühte ja sama imputeerimismeetodit, kuid võib esineda ka erijuhte [9].

Eraldi gruppides imputeeritakse peamiselt kahel põhjusel. Esiteks võivad valimi erinevates alagruppides olla erinevad seosed ja seetõttu tunnus, mis sobib imputeerimistunnuseks ühes grupis, ei ole teises grupis sobilik. Sobilike gruppide määramine eeldab head olukorra hindamise võimet ning teema tundmist. [9]

Teine põhjus seisneb selles, et alati ei ole kõigi tunnuste jaoks teada ühesugune abiinfo. Mingi kindla imputeerimismeetodi jaoks vajalikud tunnused ei pruugi olla teada kogu valimi s jaoks. Näiteks oletame, et leidub tugevalt seotud imputeerimisvektor $\mathbf{x}$, kuid ainult ühe hulga korral valimist. Sel juhul saab selle alagrupi

[1] deterministlik – kõigi sündmuste ja nähtuste objektiivsest põhjuslikust tingitusest lähtuv [4]

korral rakendada regressioon- või lähima naabri meetodit. Ülejäänud gruppide imputeerimiseks tuleb kasutada paremuselt halvemaid imputeerimisvektoreid. Vähese abiinfo korral võib viimase abivahendina kasutada ka vastanute keskmisega imputeerimist või *Hot-Deck* protseduuri. [9]

1.6 Lähima naabri imputeerimine

Lähima naabri meetodi korral püütakse leida imputeerimistunnus või tunnused, mis oleksid seotud imputeeritava tunnusega ning selle läbi vähendada tekkida võivat viga. Idee seisneb selles, et eeldatakse, et kaks sarnase x-väärtusega objekti omavad ka sarnaseid y-väärtusi. Doonorelement k leitakse kauguse minimiseerimise meetodil. [9]

Pidevad muutujad

Pidevate muutujate jaoks on absoluutne kaugus jagatud kogu vaadeldava vahemiku pikkusega:

$$d_{i,j,k} = \frac{|x_{i,k} - x_{j,k}|}{r_k},$$

kus $x_{i,k}$ on k-nda muutuja väärtus i-ndal vaatlusel ja r_k on k-nda muutuja vahemik [6].

Diskreetsed muutujad

Järjestustunnused on muudetud arvtunnusteks ja seejärel on kogu kaugus jagatud vahemiku pikkusega, mis on arvutatud. Nominaaltunnuseid käsitletakse seejuures nagu need oleks ühel kaugusel. [6]

Nominaal- ja binaarsete tunnuste jaoks kasutatakse lihtsat $0/1$ kaugust:

$$d_{i,j,k} = \begin{cases} 0, \text{ kui } x_{i,k} = x_{j,k}, \\ 1, \text{ kui } x_{i,k} \neq x_{j,k}. \text{ [6]} \end{cases}$$

1.7 *Hot-Deck* meetodite eelised ja puudused

Vaatamata sellele, et *Hot-Deck* meetod on praktikas väga laialdaselt kasutusel, pole selle kohta selgeid teoreetilisi tulemusi [1]. Tänapäeval on välja töötatud mitmeid imputeerimismeetodeid, mis on ka teoreetilisest aspektist hästi uuritud. Siiski on *Hot-Deck* meetodite suureks eeliseks nende lihtsus ja kiirus, mistõttu rakendatakse seda meetodit eriti suurte andmestike korral. Lähima naabri meetodiga imputeerimine võtab palju rohkem aega, sest iga vaatluse jaoks on vaja välja arvutada kaugust puuduvatest väärtustest, et leida k lähimat naabrit. [6]

2 EESTI TERVISHOIUTEENUSTEGA SEOTUD ANDMEBAASIDE ERIPÄRAD

Käesoleval hetkel on tervishoiuteenuste osutajatel kohustus esitada ravijuhtude dokumentatsiooni eraldi mitmesse süsteemi. Toimub samade ravijhtude andmete dubleeriv esitamine, mida tegelikult statistika tegemiseks pole vaja teha. Tervise Arengu Instituudi (TAI) üheks prioriteediks on aruandeesitajate koormuse vähendamine. Üheks potentsiaalseks statistika andmeallikaks peetakse tervise infosüsteemi (TIS) ehk Digilugu. See võimaldab mitte ainult vähendada aruandeesitajate koormust, vaid ka esitada mitmekesisemat ja detailsemat statistikat tarbijatele ning tõsta tervisestatistika kvaliteeti. Sellel eesmärgil hindab TAI regulaarselt TIS-i andmekvaliteeti. [5]

Haigekassale esitatakse andmeid ravijuhu põhiselt, kuid selles andmebaasis võib ühe ravijuhu raames olla esitatud dokumentatsioon mitu korda (vt Lisa 3.a). Tervise infosüsteemi on tervishoiu teenuste osutamised dokumenteeritud ka ravijuhtude põhiselt, kuid seal süsteemis on palju vähem epikriise (vt Lisa 3.b). Erinevalt tervise infosüsteemist maksab Haigekassa tervishoiuteenuste osutajatele andmete esitamise eest raha. Sellest tingituna on Haigekassasse laekunud palju rohkem andmeid kui tervise infosüsteemi ning see annab põhjust kahtlustada, et viimasena mainitud süsteemi on jäetud olulisel määral andmeid esitamata. Kui aga andmeid on jäetud dokumenteerimata, siis on oluline statistika tegemisel sellega arvestada ja rakendada vastavaid statistilisi meetodeid. Vastasel juhul tehakse puudulike andmete põhjal reaalsele olukorrale mittevastavad järeldused. [5] Seetõttu tuleb enne ühele süsteemile üleminekut kindlaks teha, et sinna esitatakse kõik vajalikud andmed.

Kuna Haigekassasse esitatakse rohkem andmeid, siis on seda valitud analüüsi jaoks ja statistika tegemine õigustatud just sinna laekuvate andmete põhjal. Kuid ka sellesse süsteemi esitatavad andmed pole täielikud. Seetõttu soovib TAI Haigekassale laekuvaid andmeid täiustada kasutades tervise infosüsteemi andmeid.

2.1 Eesti Haigekassa andmebaas

Avalik-õigusliku Eesti Haigekassa tähtsaim ülesanne on korraldada ravikindlustust, et võimaldada kindlustatud isikutele saada tervishoiteenust Haigekassa ravikindlustuse hüvitise eest. Institutsiooni ülesanneteks on veel aidata kaasa ka ravistandardite ja ravijuhiste koostamisele, motiveerida tervishoiuasutusi arendama tervishoiuteenuste kvaliteeti, korraldada ravikindlustust ja Haigekassat puudutavate välislepingute täitmist; osaleda tervishoiu planeerimisel; avaldada arvamust Haigekassa ja ravikindlustusega seotud õigusaktide ja välislepingute eelnõude kohta ning anda nõu ravikindlustusega seonduvates küsimustes. Lisaks sellele kogub Haigekassa tervishoiuteenust osutavatelt asutustelt ravijuhtude raviarvete kohta dokumentatsioone, et saada paremat ülevaadet ravikindlustuste kohta. (vt Lisa 3.a) [5]

Haigekassa ravikindlustuse andmekogust on antud bakalaureusetöös kasutatud järgnevaid andmeid:

- *ttokood_HK* – tervishoiuteenuse osutaja äriregistri kood (nt 90003434);
- *ID_HK* – isikut eristav unikaalne kood (mitte isikukood) (nt 10734533);
- *sugu_HK* – isiku sugu (Mees/Naine);
- *vanus_HK* – isiku vanus tervishoiuteenuse saamise ajal (nt 34, 81);
- *mk_HK* – isiku elukoht maakonna tasemel tervishoiuteenuse saamise ajal (nt “Järvamaa”, “Viljandimaa”, “välismaa”);
- *pohidgn_HK* – raviarvel olev põhi- ja kaasuv diagnoos RHK-10 haiguste klassifikatsiooni järgi (nt “S00.01”);
- *valispohjus_HK* – raviarvel olev välispõhjuse kood RHK-10 haiguste klassifikatsiooni järgi (nt “W00.01”);
- *algus_HK* – raviarve alguse kuupäev (pp/kk/aaaa);
- *lopp_HK* – raviarve lõpu kuupäev (pp/kk/aaaa);
- *summa_HK* – raviarve summa eurodes (nt 215 või 2125);
- *emo_HK* – erakorralise meditsiini osakonna abi (jah/ei);
- *tuup_HK* – tervishoiuteenuse tüüp (ambulatoorne/statsionaarne);
- *valtimatu_HK* – vältimatu abi (jah/ei);
- *ravikindl_HK* – patsiendi ravikindlustuse olemasolu (jah/ei).

Haigekassas genereeritakse raviarvete piires patsiendile vanus isikukoodi põhjal, kuid kõigile ravikindlustuseta ravitavatele on vanus jäetud arvutamata. Põhjus on selles, et Haigekassas genereeritakse patsiendile automaatselt vanus isikukoodist ja kuna ravikindlustuseta inimesed on enamjaolt välismaalased, siis nende isikukoodide eripärade tõttu on otsustatud vanust ravikindlustuseta inimestele mitte genereerida. Kvaliteetse statistika tegemiseks on aga oluline teada võimalikult palju informatsiooni patsientide kohta.

Kokku oli Haigekassa andmebaasis 2016. aasta kohta vigastusi sisaldavate andmete hulgas 294 744 raviarvet.

2.2 Tervise infosüsteem ehk Digilugu

Aastal 2008 loodud tervise infosüsteem (TIS) ehk Digilugu, mida haldab ja arendab Tervise ja Heaolu Infosüsteemide Keskus (TEHIK), on erinevaid lahendusi hõlmav tervishoiusektori koostöömudel, mille üheks oluliseks osaks on riigi infosüsteemi kuuluv andmekogu. Sellega liidestunud tervishoiuasutused saadavad sinna haiguslugude kokkuvõtteid (ehk epikriise) ning teisi meditsiinidokumente, et vahetada omavahel teavet. TIS-is töödeldakse tervishoiuvaldkonnaga seotud andmeid, muuhulgas tervislikku seisundit kajastavate registrite pidamiseks ja tervisestatistika tegemiseks. Tervise infosüsteemi vastutav töötleja on Sotsiaalministeerium ning volitatud töötleja Tervise ja Heaolu Infosüsteemide Keskus. (vt Lisa 3.b) [5]

Tervise infosüsteemist on antud bakalaureusetöös kasutatud järgnevaid andmeid:

- *ttokood_TIS* – tervishoiuteenuse osutaja äriregistri kood (nt 90003434);
- *doknr_TIS* – epikriise eristav unikaalne kood (nt 603114342);
- *ID_TIS* – isikut eristav unikaalne kood (mitte isikukood) (nt 476756325);
- *sugu_TIS* – isiku sugu (Mees/Naine);
- *vanus_TIS* – isiku vanus tervishoiuteenuse saamise ajal (nt 34, 81);
- *mk_TIS* – isiku elukoht maakonna tasemel tervishoiuteenuse saamise ajal (nt "Järvamaa", "Viljandimaa", "välismaa");
- *pohidgn_TIS* – epikriisil olev põhi- ja kaasuv diagnoos RHK-10 haiguste klassifikatsiooni järgi (nt "S00.01");
- *valispohjus1_TIS*, *valispohjus2_TIS*, ... – epikriisil olevad välispõhjused RHK-10 haiguste klassifikatsiooni järgi (nt "W00.01");
- *algus_TIS* – ravijuhu alguse kuupäev (pp/kk/aaaa);
- *lopp_TIS* – ravijuhu lõpu kuupäev (pp/kk/aaaa);
- *tuup_TIS* – teenuse osutamise viis (ambulatoorne/statsionaarne);
- *valtimatu_TIS* – erakorraline abi (jah/ei);
- *ravikindl_TIS* – patsiendi ravikindlustuse olemasolu (jah/ei);

Tervise infosüsteemi on patsientidele vanus arvutatud isikukoodist või võetud arstide poolt käsitsi sisestatud andmetest. Seega vanuse väärtus leidub kõikidel patsientidel. Seetõttu on otsustatud kasutada just tervise infosüsteemi andmeid Haigekassa andmete täiustamiseks.

2016. aastal oli tervise infosüsteemi vigastuste kohta andmebaasi saadetud $186\,283$ ravijuhu dokumenti.

2.3 RHK-10 koodid

RHK-10 haiguste klassifikatsiooni võib defineerida kui jaotiste süsteemi, millesse haigused (või haiguste nimetused) on määratud vastavalt kehtestatud kriteeriumitele. RHK-10 peamiseks eesmärgiks on võimaldada eri aegadel kogutud suremuse ja haigestumuse andmete süstemaatiline registreerimine, analüüsimine, interpreteerimine ja võrdlemine. Seda klassifikatsiooni kasutatakse haiguste diagnooside ja muude terviseprobleemide ülekandmiseks sõnadest tärkkoodi. See võimaldab andmete hõlpsat säilitamist, otsingut ja analüüsi ka rahvusvahelisel tasandil. [3]

Käesolevas bakalaureusetöös on vaatluse alla võetud ainult need ravijuhud, mis on seotud vigastustega. Selleks on välja sorteeritud ainult nende ravijuhtude andmeid, mille haigust tekitanud välispõhjused algavad RHK koodides tähega "V", "W", "X" või "Y" või mille põhidiagnoosi RHK kood algab tähega "S" või "T".

RHK-10 klassifikatsiooni järgi paiknevad põhidiagnoosi koodide "$S00$–$T98$" all vigastused, mürgitused ja teatavad muud välispõhjuste toime tagajärjed. Imputeerimise jaoks on need omakorda jaotatud mittelõikuvatesse gruppidesse:

- pea- ja kehapiirkonna vigastused;
- kätepiirkonna vigastused;
- jalapiirkonna vigastused;
- muud täpsustamata piirkonna ja muud liiki vigastused või tüsistused (vt Lisa 1). [3]

Välispõhjuse koodide "$V01$–$Y98$" all on haigestumise ja surma välispõhjused. Tähega "V" algavate välispõhjuste koodide all paiknevad täpsemalt sõidukiõnnetuses vigastuse saanud patsientide ravijuhtude andmed. Tähega "W" algavate välispõhjuste koodidega on eraldatud füüsikaliste faktorite poolt põhjustatud vigastuste ravijuhud (nt kukkumised, elekter jms). "$X00$–$Y34$"-ga on välja toodud erinevatest loodusnähtustest ja muudest teguritest põhjustatud vigastuse andmed (nt tuli, põletused, mürgitused, jms). Koodide "$Y35$–$Y98$" all asuvad aga kõik inimeste ja muudest faktoritest põhjustatud vigastuste andmed. Vastavalt eelnevale kirjeldusele on välispõhjuste koodid ära grupeeritud (vt Lisa 2). [3]

3 ANDMETE ÜHENDAMINE

Bakalaureusetöö esimeseks sammuks oli Haigekassa ja TIS andmeesituse süsteemidesse laekunud ravijuhtude dokumentatsioonide omavaheline ühendamine. Peamiseks probleemiks osutus see, et samad patsiendid olid erinevate ID koodidega erinevates andmestikes. Probleemi põhjustas erinev patsiendi ID koodi genereerimise algoritm Eesti Haigekassa ja TIS-i andmebaasides. Kuna Tervise Arengu Instituut ei ole ametlikult riikliku statistika teostaja, siis neil ei ole õigust saada ühesuguste ID-koodidega patsientide andmeid, sest Andmekaitse Inspektsiooni põhimõtete järgi on tegemist andmete lekkega. Seetõttu polnud kahjuks haiguslugusid võimalik mõlemast andmebaasist ühendada patsiendi ID-koodide põhjal.

Selleks, et ühendamise protsessi siiski läbi viia, üritati ravijuhtude dokumenti erinevatest andmestikest ühendada järgnevate tunnuste põhjal:

- tervishoiuteenuse osutaja äriregistri kood (*ttokood*);
- isiku sugu (*sugu*);
- isiku vanus (*vanus*);
- isiku elukoht maakonna tasemel tervishoiuteenuse saamise ajal (*mk*);
- ravijuhul olevad põhi- ja kaasuvad diagnoosid (*pohidgn*);
- ravijuhul olevad välispõhjused (*valispohjus*);
- ravijuhu alguse kuupäev (*algus*);
- ravijuhu lõpu kuupäev (*lopp*);
- tervishoiuteenuse tüüp (*tuup*);
- vältimatu abi (jah/ei) (*valtimatu*);
- patsiendi ravikindlustuse olemasolu (*ravikindl*).

Andmete esialgse analüüsi tulemusena avastati, et nii Haigekassas kui ka tervise infosüsteemi vigastusjuhtumite andmetes esineb palju duplikate vaadeldavate tunnuste põhjal. Antud analüüsis kasutati rakendustarkvara STATA funktsiooni *Merge*, mis ei lubanud korduvaid andmeid ühendada. Ülesande lihtsustamise eesmärgil võeti TAI tervisestatistika osakonna koosolekul analüütikute poolt vastu otsus, et igat kombinatsiooni katsetades ühendatakse omavahel ainult selliseid ravijuhte, mis on vastava vaadeldava kombinatsiooni põhjal unikaalsed.

Enne vaatluste ühendamist uuriti Haigekassa andmetest, millistel ravikindlustuseta patsientidel on vastava ID-koodi põhjal mõne muu raviarve korral vanuse väärtus olemas. Täpsemate ühendamise tulemuste saamise eesmärgil omistati sellistele raviarvetele ajutiselt olemasolevate andmetega raviarvelt patsiendi vanuse väärtus.

3.1 Sisestusvead ja nendega arvestamine

Üheks hüpoteesiks, mida sooviti ka kontrollida andmete ühendamisega oli see, et arstid teevad andmete sisestamisel vigu ehk sisestavad andmeid sama ravijuhu kohta mõlemasse süsteemi erinevalt (väärtused on erinevad või on väärtus ühte süsteemi esitatud ja teise mitte). Vastavate erinevuste ja iseärasustega oli tarvis arvestada, et leida võimalikult paljudele Haigekassa ravikindlustuseta patsientide ravijuhu vanused tervise infosüsteemi andmebaasist.

Vead, mida ühendamise protsessis suudeti avastada (esinesid mõlema süsteemi andmetes) olid järgmised:

- isiku elukoht maakonna tasemel tervishoiuteenuse saamise ajal on erinev või puudu;
- ravijuhul olev põhidiagnoosi kood sisestatud erinevalt või jäetud märkimata;
- ravijuhul olevad välispõhjuse kood sisetatud erinevalt või jäetud märkimata;
- ravijuhu alguse kuupäeva muutus;
- ravijuhu lõpu kuupäeva muutus;
- isiku vanuse muutus, kui ravijuht on erinevate kuupäevadega;
- tervishoiuteenuse tüüp on määratud erinevalt;
- vältimatu abi on määratud erinevalt või jäetud märkimata;
- patsiendi ravikindlustuse olemasolu on erinev.

Ühendamise protsessis pandi tähele, et mida rohkem arvestada arstide poolt tehtavate vigadega, seda rohkem esineb ühendatud andmete hulgas sellist olukorda, kus ühest andmestikust on ühele ID-koodile seatud vastavusse teise andmestiku kaks erinevat ID-koodi. Seetõttu oli tarvis välja valida kõige mõistlikumad vigade esinemise kombinatsioonid, mida ühendamise protsessis arvestada.

Selleks, et ühendamise protsessis kõige mõistlikumad kombinatsioonid välja valida, moodustati meeskond, kuhu kuulusid Tervise Arengu Instituudi tervisestatistika osakonna vanemanalüütikud ja analüütikud.

3.2 Andmebaaside vaatluste ühendamine

Kokkuvõttes otsustati 187 erineva kombinatsiooni kasuks (vt Lisa 4). Vaadeldavate kombinatsioonide korral lubati teatud tunnustel erineda ühendamise protsessis, kuna eeldati, et arstid võisid teha vigu andmete sisestamisel.

Ühendamise protsessis avastati, et arstid võisid esitada dokumentatsiooni ühe ja sama ravijuhu kohta erineval ajal ehk ühte süsteemidest hiljem teisest. Üheskoos TAI tervisestatistika osakonna meeskonnaliikmetega võeti vastu otsus määrata maksimaalseks lubatud ajavahemikuks kahe ühe ja sama epikriisi vahel 30 päeva.

Ühendamise protsessis avastati, et kui lubati tunnusel *vanus* erineda kõikide Haigekassa ravijuhtude korral, siis ühendamises võis tekkida selliseid vigu, kus näiteks ühildus ühe andmebaasi 17-aastase ja teise andmestiku 80-aastase patsiendi ravijuht. Seetõttu võeti ühiselt vastu otsus lubada patsiendi vanusetunnusel erineda ainult Haigekassa vanuse väärtuseta inimeste ravijuhtudel (vt Lisa 4 “(vanusega)”, “*vanus* puudub”).

Ühendamisel arvestati veel infoga, et epikriisil pannakse patsiendi vanus vastavalt ravijuhtumi alguskuupäevale. Kui aga epikriis esitati ühte süsteemidest hiljem kui teise, siis võis inimene vaadeldaval perioodil saada ka aasta vanemaks (vt Lisa 4 juhud “156–171” ja “184–187” “vanus ± 1”)

Ühendamise tulemusena ühendus mõlemast andmebaasist 157 620 vaatlust. Tervise infosüsteemist jäi ühendamata 28 663 vaatlust ja Haigekassa andmebaasist jäi vasteta 137 124 vaatlust.

3.3 Näiteid ühendatud ridadest

Järgnevates näidetes kasutatud andmed on ühendamise protsessi kirjeldamise jaoks välja mõeldud ega vasta reaalsusele, kuna bakalaureusetöös on kasutatud delikaatseid isikuandmeid. Järgnevates näidetes on ühe objekti andmed jagunenud kahele reale.

Näide 1

Oletame, et Haigekassa vigastuse andmeid sisaldavas andmebaasis on vaatlus järgmine

ttokood_HK	ID_HK	tuup_HK	vanus_HK	sugu_HK	mk_HK	pohidgn_HK	valispohjus_HK
90001479	11436453	Statsionaarne	77	Mees	Järvamaa	S72.08	W10.01

algus_HK	lopp_HK	valtimatu_HK	ravikindl_HK	ravijuhte_HK	summa_HK	emo_HK
11.11.2016	12.11.2016	Jah	Jah	1	4003	ei

ja vastavalt tervise infosüsteemi andmestikus on objekt

ttokood_TIS	ID_TIS	tuup_TIS	vanus_TIS	sugu_TIS	mk_TIS	pohidgn_TIS	valispohjus1_TIS
90001479	476756325	Statsionaarne	77	Mees	Järvamaa	S72.08	W10.01

algus_TIS	lopp_TIS	valtimatu_TIS	ravikindl_TIS	doknr_TIS
11.11.2016	12.11.2016	Jah	Jah	663395869

Läbiproovitud kombinatsioonide põhjal oleksid vastavad vaatlused ühendunud 1.0 juhu põhjal (vt Lisa 4).

Näide 2

Kui meil on antud Haigekassa andmebaasis vaatlus

ttokood_HK	ID_HK	tuup_HK	vanus_HK	sugu_HK	mk_HK	pohidgn_HK	valispohjus_HK
90001498	10083912	Statsionaarne	0	Naine	Tartumaa	T84.0	Y83.1

algus_HK	lopp_HK	valtimatu_HK	ravikindl_HK	ravijuhte_HK	summa_HK	emo_HK
15.02.2016	15.02.2016	Jah	Jah	1	339	ei

siis see oleks ühendunud TIS andmebaasi vaatlusega

ttokood_TIS	ID_TIS	tuup_TIS	vanus_TIS	sugu_TIS	mk_TIS	pohidgn_TIS	valispohjus3_TIS
90001498	454357379	Statsionaarne	0	Naine	Tartumaa	T84.0	Y83.19

algus_TIS	lopp_TIS	valtimatu_TIS	ravikindl_TIS	doknr_TIS
15.02.2016	15.02.2016	.	Jah	623539687

juhu 5.2 põhjal (vt Lisa 4).

Näide 3

Kui meil on antud ravikindlustuseta patsiendi vaatlus Haigekassa andmebaasis

ttokood_HK	ID_HK	tuup_HK	vanus_HK	sugu_HK	mk_HK	pohidgn_HK	valispohjus_HK
90004587	11352739	Statsionaarne	.	Naine	Viljandimaa	S06.6	W19.48

algus_HK	lopp_HK	valtimatu_HK	ravikindl_HK	ravijuhte_HK	summa_HK	emo_HK
11.09.2016	17.09.2016	Ei	Ei	1	747	jah

siis see oleks ühendunud TIS andmebaasi vaatlusega

ttokood_TIS	ID_TIS	tuup_TIS	vanus_TIS	sugu_TIS	mk_TIS	pohidgn_TIS	valispohjus2_TIS
90004587	1990122244	Statsionaarne	34	Naine	Viljandimaa	S06.6	W19.49

algus_TIS	lopp_TIS	valtimatu_TIS	ravikindl_TIS	doknr_TIS
11.09.2016	17.09.2016	.	Ei	630630987

juhu 6.2 põhjal (vt Lisa 4).

Näide 4

Haigekassa andmebaasi vaatlus

ttokood_HK	ID_HK	tuup_HK	vanus_HK	sugu_HK	mk_HK	pohidgn_HK	valispohjus_HK
90001478	11471609	Statsionaarne	16	Mees	Tartumaa	S53.49	W51.01

algus_HK	lopp_HK	valtimatu_HK	ravikindl_HK	ravijuhte_HK	summa_HK	emo_HK
24.03.2016	24.03.2016	Jah	Jah	1	100	ei

oleks ühendunud TIS andmebaasi vaatlusega

ttokood_TIS	ID_TIS	tuup_TIS	vanus_TIS	sugu_TIS	mk_TIS	pohidgn_TIS	valispohjus1_TIS
90001478	1729907619	Statsionaarne	15	Mees	Tartumaa	S53.40	W51.01

algus_TIS	lopp_TIS	valtimatu_TIS	ravikindl_TIS	doknr_TIS
17.03.2016	24.03.2016	Jah	Jah	541166758

juhu 156.1 põhjal (vt Lisa 4).

4 IMPUTEERIMINE

Haigekassa ravikindlustuseta patsientide raviarvetele omistati tervise infosüsteemist saadud vanuse väärtused. Ühendamise protsessis suudeti leida vanused 5633 ravikindlustuseta patsiendi raviarvele ja vanuseta jäid 3515 raviarvet.

Imputeerimiseks kasutati kolme erinevat viisi: üldist juhuslikku *Hot-Deck* omistust, lähima naabri imputeerimist ning juhuslikku *Hot-Deck* omistust klassis kombineerituna lähima naabri imputeerimisega. Viimasena mainitud meetodi korral esines probleeme, kuna objektide grupeerimisel tekkis selliseid doonorgruppe, kus ühelgi vaatlusel ei leidunud vanuse väärtust. Rakendustarkvara R "VIM"-pakett omistas sellistele vaatlustele vanuse 1, mis polnud Tervise Arengu Instituudi analüütikute arvates õige tegutsemisviis. Tühjadesse doonorgruppidesse sattunud vaatluste jaoks otsustati seetõttu kasutada lähima naabri imputeerimismeetodit üle kogu andmestiku. (vt Lisa 7)

Tervise Arengu Instituudi analüütikute ja vanemanalüütikute meeskond otsustas eksperthinnanguna imputeerimise protsessi kaasata algupärasena järgnevaid tunnuseid:

- *tuup* – tervishoiuteenuse tüüp (1-"ambulatoorne", 2-"statsionaarne");
- *sugu* – patsiendi sugu (1-"Mees", 2-"Naine");
- *vanus* – patsiendi vanus (täisarvuline väärtus).

Lisaks otsustati imputeerimiseks luua järgnevad tunnused (vt Lisa 5):

- *perearst* – kas tegemist oli perearsti vastuvõtuga või suurema raviasutuse vastuvõtuga (0-"suurem raviasutus", 1-"perearst");
- *summa_ryhm* – patsiendi raviarve summa vahemik ("...-100", "101-200", "201-...");
- *pohidgn_grupp* – ravijuhu põhidiagnoosi koodi grupp RHK-10 järgi (vt Lisa 1);
- *valispohjus_grupp* – ravijuhu välispõhjuse koodi grupp RHK-10 järgi (vt Lisa 2);
- *kp_vahe_ryhm* – patsiendi ravijuhu kestus päevades vahemikuna ("<=0", "0-5", "6-10", ">10");
- *algus_kuu* – raviarve alguskuupäeva kuu (1-" jaanuar", ..., 12-"detsember").

Imputeerimisest jäeti välja järgmised tunnused:

- *valtimatu* – kas tegemist oli vältimatu abi ostamisega või mitte (0-"Ei", 1-"Jah");
- *emo* – kas tegemist oli erakorralise meditsiini osakonna abi ostamisega või mitte (0-"Ei", 1-"Jah");

- *mk* – patsiendi elukoht raviteenuse saamise ajal;
- *lopp* – raviarve lõpu kuupäev.

Tunnused *valtimatu* ja *emo* jäeti imputeerimisest välja seetõttu, et vanemanalüütikute eksperthinnangu põhjal tervishoiuteenuse osutajad ei sisesta antud tunnuste väärtusi hoolikalt. Tunnus *mk* ei võetud imputeerimise protsessi, sest eksperthinnangu põhjal otsustati, et inimese elukohast ei sõltu vigastuse välispõhjustaja liik, patsiendi sugu ega vanus. Lisaks sellele vaadeldava tunnuse väärtusi oli Haigekassa andmebaasi vähe sisestatud. Tunnust *lopp* ei kasutatud imputeerimisel, sest protsessi oli juba kaasatud tunnused *algus_kuu* ja *kp_vahe_ryhm*.

Kokkuvõttes esines ravikindlustuseta patsientide andmetes puuduvaid väärtusi välispõhjuste koodides 9, soo väärtustes 37 ja vanuse väärtustes 3515. Imputeerimise protsessis otsustati esmalt imputeerida välispõhjuste grupi tunnused, siis soo väärtused ja lõpuks vanuse väärtused. Iga järgneva tunnuse imputeerimisel võeti kasutusele ka eelmise tunnuse imputeerimisel saadud tulemused.

4.1 Andmete eelnev analüüs

Tunnuse imputeerimiseks otsustati kasutada ainult neid abitunnuseid, millel esineks olemasolevate andmete põhjal imputeeritava tunnusega statistiline seos. Käesolevas bakalaureusetöös on kasutatud statistiliste seoste uurimiseks χ^2-teste ja T-teste ning need on läbiviidud kasutades rakendustarkvara R. Olulisuse nivooks on võetud $\alpha = 0.05$.

Tunnuse *valispohjus_grupp* seoseid abitunnustega uuriti ainult kasutades χ^2-teste, sest tegemist on nominaaltunnusega (vt Lisa 6.a). Lõpptulemuseks saadi vaadeldava tunnuse imputeerimiseks järgmised tunnused:

- *tuup*
- *kp_vahe_ryhm*
- *summa_ryhm*
- *algus_kuu*
- *pohidgn_grupp*
- *perearst*

Tunnuse *sugu* seoseid abitunnustega uuriti ainult kasutades χ^2-teste, sest tegemist on binaarse tunnusega (vt Lisa 6.b). Lõpptulemuseks saadi vaadeldava tunnuse imputeerimiseks järgmised tunnused:

- *summa_ryhm*
- *pohidgn_grupp*

Pideva tunnuse *vanus* seoseid nominaalsete abitunnustega, millel esineb kaks väärtust, uuriti kasutades T-teste ja suurema arvu võimalike väärtustega tunnuste korral kasutades χ^2-teste. Viimasena mainitud testi läbiviimseks moodustati tunnus *vanusryhm* (vt Lisa 6.c). Lõpptulemuseks saadi tunnuse *vanus* imputeerimiseks järgmised tunnused:

- *tuup*
- *kp_vahe_ryhm*
- *summa_ryhm*
- *algus_kuu*
- *sugu*
- *pohidgn_grupp*
- *valispohjus_grupp*

4.2 Imputeerimise tulemused

Imputeerimine on läbi viidud kasutades rakendustarkvara R kasutades paketti "VIM" (vt Lisa 7).

4.2.1 Vigastuste välispõhjuse grupitunnuse imputeerimise tulemused

Tunnusel *valispohjus_grupp* oli puuduvaid väärtusi kokku 9. Imputeeritud gruppide kirjeldusi on võimalik vaadata Lisast 2.

Tabel 1. Vigastuste välispõhjuse RHK-10 koodigruppide osakaalud ja imputeeritud väärtuste arv (vt Lisa 8).

	Gr 1	Gr 2	Gr 3	Gr 4	SUM
Üldine juhuslik *Hot-Deck* meetod	5.14% (0)	72.60% (9)	7.37% (0)	14.90% (0)	100.01%
Lähima naabri meetod	5.14% (0)	72.58% (8)	7.38% (1)	14.92% (0)	100.00%
Juhuslik *Hot-Deck* omistus klassis ja lähima naabri meetod	5.14% (0)	72.55% (5)	7.39% (2)	14.90% (2)	100.00%
Olemasolevad andmed	5.14%	72.57%	7.37%	14.91%	99.99%

Tabelist 1 on näha, et imputeerimise tulemused ei anna osakaaludes erinevate meetodite korral väga erinevaid tulemusi. Võib välja tuua, et füüsikaliste faktorite poolt põhjustatud välispõhjuste (*Gr 2*) osakaal on suurenenud üldise juhusliku *Hot-Deck* meetodi kui ka lähima naabri meetodi korral – vastavalt 0.03% ja 0.01% võrra, kus imputeeritud sai juurde vastavalt 9 ja 8 väärtust. Sõidukiõnnetuses saadud vigastuse välispõhjuste (*Gr 1*) osakaal pole ühegi meetodi korral muutunud, sest väärtuseid juurde ei imputeeritud. Inimese ja muude faktorite poolt põhjustatud vigastuste välispõhjuste (*Gr 4*) osakaal on suurenenud ainult juhusliku *Hot-Deck* omistus klassis korral (antud tunnuse imputeerimisel lähima naabri meetodit kombinatsioonis ei kasutatud, vt Lisa 8) ja seda 0.01% võrra – imputeeriti juurde 2 väärtust. Teiste meetodite korral on selle väärtuse osakaal vähenenud 0.01% võrra, sest ei lisandunud ühtegi väärtust. Loodusnähtustest ja muudest teguritest põhjustatud vigastuste välispõhjuste (*Gr 3*) osakaal on suurenenud lähima naabri meetodi korral 0.01% võrra (imputeeriti juurde 1 väärtus) ja juhusliku *Hot-Deck* omistus klassis korral 0.02% võrra (lisandus 2 väärtust).

4.2.2 Sootunnuse imputeerimise tulemused

Tunnusel *sugu* oli puuduvaid väärtusi kokku 37.

Tabel 2. Sugude osakaalud (vt Lisa 8).

	Osakaalud	
	Naised	Mehed
Üldine juhuslik *Hot-Deck* meetod	16.48% (5)	83.52% (32)
Lähima naabri meetod	16.43% (0)	83.57% (37)
Juhuslik *Hot-Deck* omistus klassis ja lähima naabri meetod	16.46% (3)	83.54% (34)
Olemasolevad andmed	16.50%	83.50%

Tabelist 2 on võimalik näha, et suuri erinevusi sootunnuse osakaaludes andmete imputeerimisel erinevate meetoditega ei esine. Olemasolevate andmete hulgas oli 16.50% ravijuhtude korral tegemist naissoost patsientidega ja 83.50% juhtudest meessoost patsientidega. Kõikide meetodite korral suurenes meeste osakaal andmestikus. Lähima naabri meetodi korral imputeeriti tunnuse *sugu* väärtuseks "Mees" kõigile 37-le vaatlusele ja meessoost patsientide raviarvete osakaal and-

mestikus suurenes vastavalt sellele 0.07% võrra. Üldise juhusliku *Hot-Deck* meetodi korral imputeeriti juurde 5 naise ja 32 mehe soo väärtust – vastavalt sellele suurenes meeste raviarvete osakaal ja vähenes naiste raviarvete osakaal andmestikus 0.02% võrra. Juhusliku *Hot-Deck* omistuse klassis korral imputeeriti tunnusele *sugu* 3 väärtust "Naine" ja 34 väärtust "Mees". Viimasena mainitud väärtuse osakaal andmestikus suurenes 0.04% võrra (antud tunnuse imputeerimisel lähima naabri meetodit ei kasutatud, vt Lisa 8).

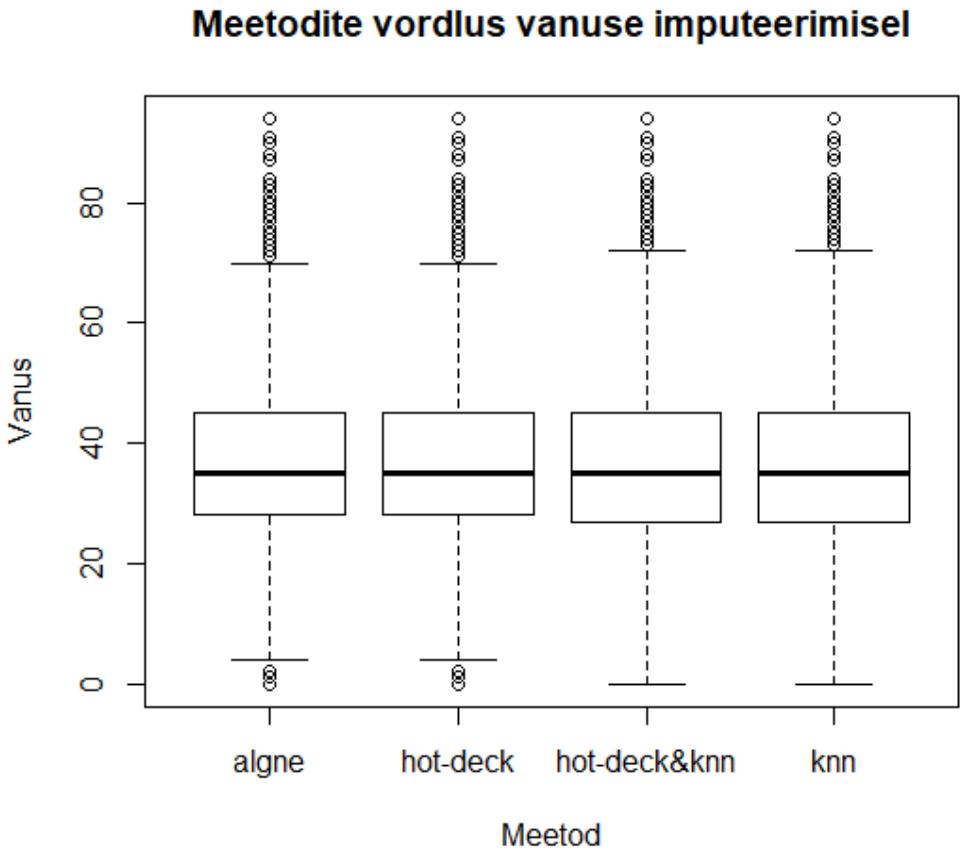
4.2.3 Patsiendi vanuse imputeerimise tulemused

Vaadeldavates andmetes eksisteerisid patsiendi vanused 5633 raviarvel ja vanuseta olid 3515 raviarvet.

Tabel 3. Vanuse karakteristikud kõikide andmete lõikes (vt Lisa 8).

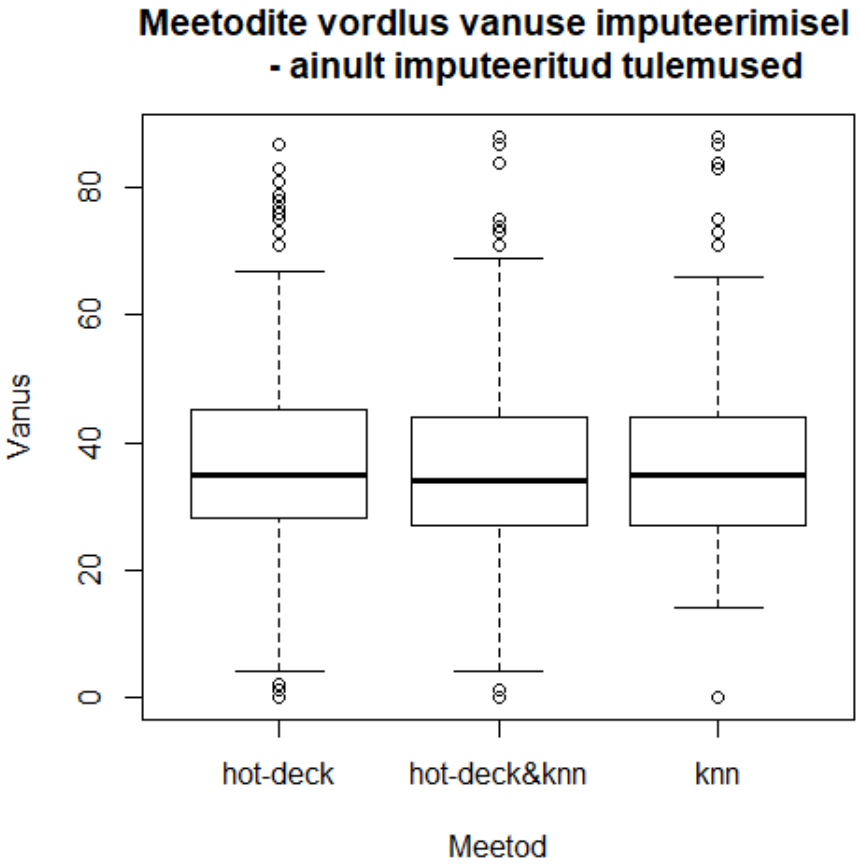
	Keskmine	Standardh.	Min	Max	Med
Üldine juhuslik *Hot-Deck* meetod	36.969	11.856	0	94	35
Lähima naabri meetod	36.715	11.862	0	94	35
Juhuslik *Hot-Deck* omistus klassis ja lähima naabri meetod	36.519	11.837	0	94	35
Olemasolevad andmed	36.890	11.858	0	94	35

Erinevate meetodite korral muutusi minimaalse ja maksimaalse vanuse korral ei esinenud – minimaalseks vanuseks jäi 0, maksimaalseks 94, mis on ka mõistetav, sest imputeeritakse olemaolevate andmete hulgast (vt ptk 1.2). Vanuste mediaaniks jäi kõikide meetodite korral 35. Olemasolevate andmete põhjal oli algselt keskmiseks vanuseks 36.890 aastat ja tunnuse hajuvuseks 11.858. Tabelist 3 on näha, et suuri erinevusi imputeeritud andmete vahel ei esine. Imputeeritava tunnuse hajuvus on suurenenud ainult lähima naabri meetodi korral, kuid seda kõigest 0.004 võrra. Keskmine vanus on tõusnud üldise juhusliku *Hot-Deck* meetodi korral kuni 36.969 aastani ning vähenenud ülejäänud meetodite korral – lähima naabri meetodi korral 36.715 aastani ja juhusliku *Hot-Deck* omistus klassis korral kombineerituna lähima naabri meetodiga 36.519 aastani.



Joonis 1. Vanuste karpdiagramm algses ja lõplikes andmestikes peale imputeerimist (vt Lisa 8).

Karpdiagrammil esitatakse arvkarakteristikutest kvartiilid ja mediaan horisontaaljoontega, mille otspunktid ühendatakse vertikaaljoontega. Jooniselt 1 on näha, et lõplikes andmestikes on tunnuse *vanus* vaadeldavad arvkarakteristikud samad kõikide tulemuste korral ja need ei erine algselt olemasolevatest andmetest. Vurrude tippudes asuvad valimi vaadeldava tunnuse maksimaalsed ja minimaalse väärtused, mis antud juhul on erinevate andmestike korral samuti üsna sarnased. Välja võib tuua, et algsete väärtustega on natukene sarnasemad eelnimetatud karakteristikud just üldise juhusliku *Hot-Deck* meetodi korral (*hot-deck*), väikseid erinevusi esineb grupipõhise *Hot-Deck* meetodi ja lähima naabri meetodi kombinatsiooni (*hot-deck&knn*) ning lähima naabri meetodi korral (*knn*). Punktidega on märgitud vaatlused, mis on mediaanist kaugemal kui poolteist kvartiilide vahet. On näha, et need on jaotunud erinevate meetodite korral samuti üsna sarnaselt.



Joonis 2. Vanuse imputeerimistulemuste meetodite võrdlus lõplikus andmestikus (vt Lisa 8).

Joonise 2 koostamisel on arvesse võetud ainult imputeeritud väärtused. Sellelt jooniselt on näha, et kõige väiksema varieeruvusega on raviarvete patsiendi vanuseid imputeerimisel andnud lähima naabri meetod (*knn*). Suuremat tunnuse *vanus* väärtuste varieeruvust esineb imputeeritud andmetes üldise juhusliku *Hot-Deck* meetodi (*hot-deck*) ning juhusliku *Hot-Deck* omistuse klassis ja lähima naabri meetodi kombinatsiooni (*hot-deck&knn*) korral imputeeritud andmetes. Tervise Arengu Instituudi analüütikute prognoosi kohaselt peaksidki ravikindlustuseta patsiendid kuuluma just sinna vanuserühma, kuhu lähima naabri meetod vanuseid kõige rohkem imputeeris.

5 IMPUTEERIMISMEETODITE KVALITEET

Antud bakalaureusetöös sõltub Haigekassa andmebaasis ravikindlustuseta patsiendi vanuse olemasolu TIS ja Haigekassa andmete ühendamise kvaliteedist. Vaadeldavas protsessis võisid ravijuhud jääda ühendamata juhul, kui arst tegi andmete sisestamisel olulisi vigu või kui mõned ravijuhtud olid liiga sarnased ühendamiseks (tekkisid duplikaadid). Vaadeldavad olukorrad võisid aga juhtuda täiesti juhuslikult. Järelikult vanuse puudumist võib käsitleda kui täiesti juhuslikku mittevastamist (ingl k *Missing Completely at Random (MCAR)*). Selle korral ei sõltu uuritav tunnus ühestki abitunnusest ja olemasolevate väärtuste jaotus on sama, mis puuduvatel väärtustel [2].

Imputeerimismeetodite võrdlemiseks selekteeriti ühendatud andmete hulgast need ravikindlustuseta patsientide raviarved, millel oli olemas kõikide tunnuste väärtused (vigastuse välispõhjuse koodigrupp, sugu ja vanus). Vaadeldavas peatükis viiakse läbi kaks katset, kus jäetakse esimesel juhul vähemalt 70% ja teisel juhul vähemalt 50% vanuse väärtust alles täiesti juhuslikult (vt Lisa 9.a ja 10.a) ning viiakse läbi käesolevas bakalaureusetöös kasutatavatel meetoditel imputeerimised (vt Lisa 9.b ja 10.b). Saadud imputeerimistulemusi võrreldakse algsete väärtustega ja vastavalt tulemuste analüüsile valitakse välja parim imputeerimismeetod.

On vaadeldud ka ainult imputeeritud tulemuste väärtusi ehk kui suur oli vahe patsiendi tegeliku vanusega. Selleks on eraldatud ainult need read, millele vanuse väärtus imputeeriti ja ühendatud vaatlused algsete andmetega. Ühendatud andmetes loodi tunnus *vanusevahe*, mis näitab, kui suur oli imputeerimistulemuse vahe tegelikust väärtusest (vt Lisa 9.c ja 10.c).

Imputeerimismeetodite headuses veendumiseks viidi läbi kokku 100 simulatsiooni, kus esimesel juhul imputeeriti 100 korda meetoditega ravijuhtude patsientidele vanused ühtedele ja samadele andmetele. Teisel juhul viidi enne igat simulatsiooni sammu läbi vastavalt 70% ja 50% andmete allesjätmine, nii et igal sammul kasutati imputeerimismeetodeid erinevatel algandmetel. (vt Lisa 9.d ja 9.d) Saadud tulemusi analüüsiti karpdiagrammide baasil (vt Lisa 9.e ja 10.e).

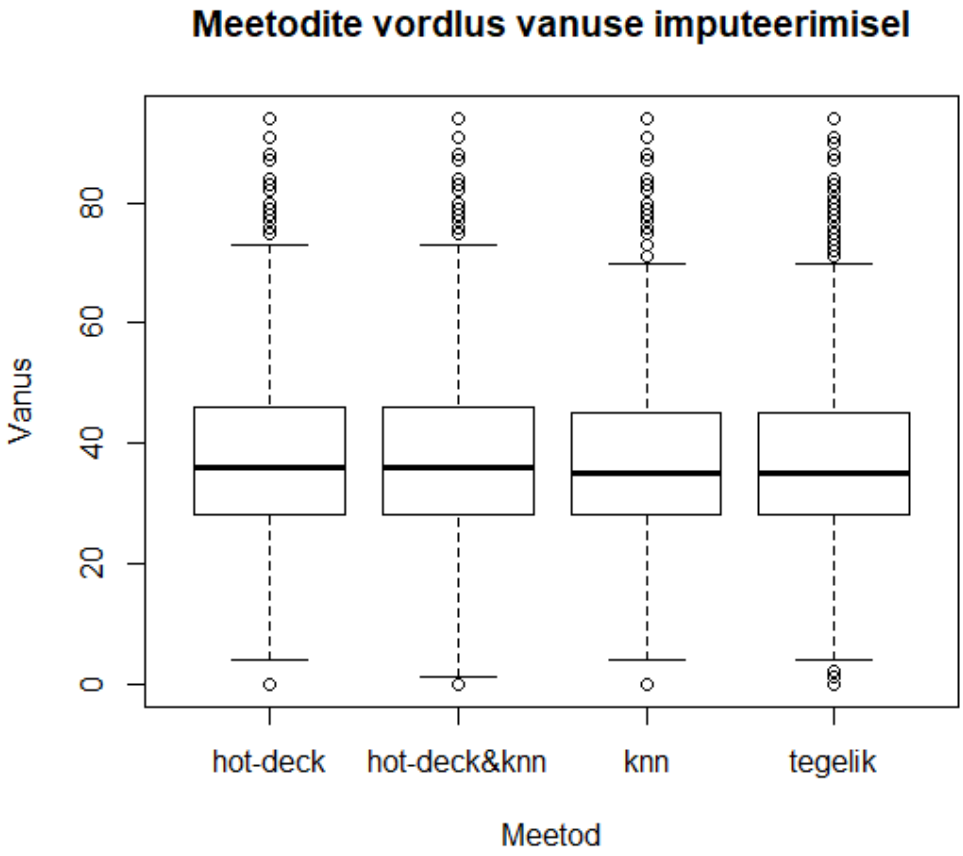
5.1 Imputeerimise kvaliteet 70% info olemasolu korral

Tabelis 4 on välja toodud 30% puuduvate andmete imputeerimisel saadud tulemuste karakteristikuid.

Tabel 4. Vanuse karakteristikud kõikide andmete lõikes (vt Lisa 9.c).

	Keskmine	Standardh.	Min	Max	Med
Üldine juhuslik *Hot-Deck* meetod	37.200	12.009	0	94	36
Lähima naabri meetod	36.942	12.020	0	94	35
Juhuslik *Hot-Deck* omistus klassis ja lähima naabri meetod	37.126	12.002	0	94	36
Tegelik	36.897	11.858	0	94	35

Tabelist 4 on näha, et üldise juhusliku *Hot-Deck* ning doonorgrupipõhise *Hot-Deck* meetodi ja lähima naabri meetodi kombinatsiooni korral on vanuse mediaan erinev tegelikkusest – 35 asemel on see 36. Vanuse miinimumid ja maksimumid on jäänud muutumatuks, sest väärtuseid imputeeritakse olemasolevatest andmetest (vt ptk 1.2) ning kui maksimaalseid ja minimaalseid väärtuseid ei kustutatud, siis need jäävad samaks ka imputeeritud andmetes. Tabelist on veel näha, et iga imputeerimismeetodi korral on vaadeldava tunnuse *vanus* hajuvus suurenenud. Selline olukord on ilmselt tingitud sellest, et doonoripõhiste imputeerimismeetoditega võib kaasneda imputeerimisnihe (vt ptk 1.3), mis suurendab omakorda standardhälvet. Tegelik keskmine vanus oli 36.897 aastat ja suuri muutusi imputeerimisel vaadeldava karakteristiku väärtuses ei esinenud. Kõige täpsema keskmise andsid lähima naabri meetodiga imputeeritud andmed – 36.942. Üldise juhusliku *Hot-Deck* meetodi ning juhusliku doonorgrupipõhise *Hot-Deck* ja lähima naabri meetodi kombinatsiooni korral tuli andmete keskmine vanus üle 37 aasta – vastavalt 37.200 ja 37.126.



Joonis 3. Vanuse imputeerimistulemuste karpdiagramm kõikide andmete lõikes (vt Lisa 9.c).

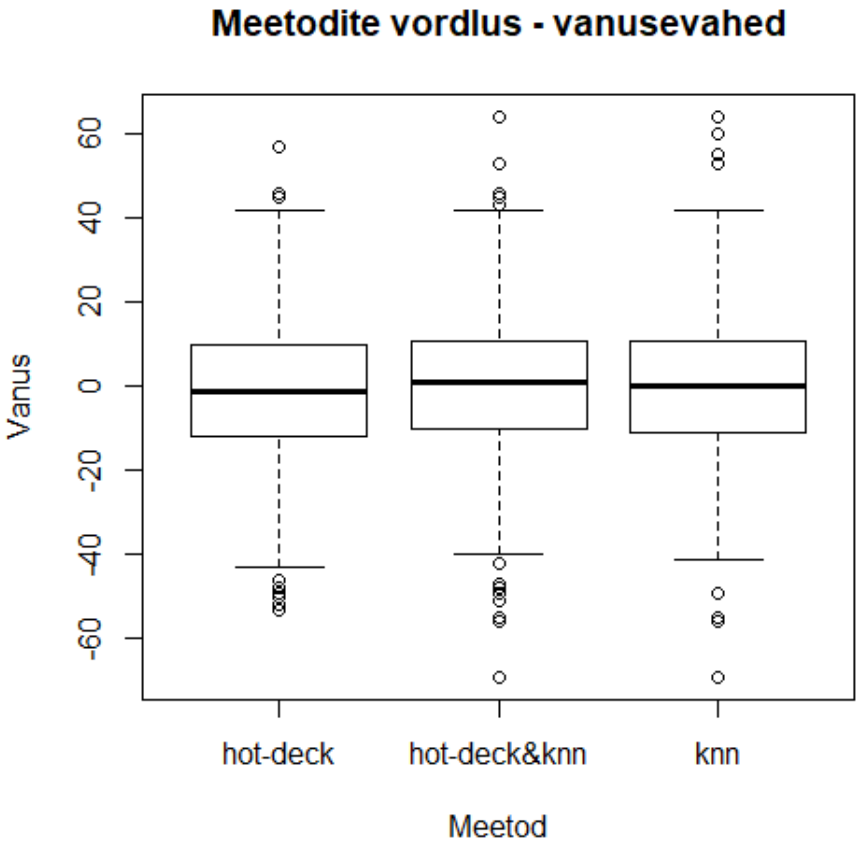
Joonise 3 karpdiagrammilt on näha, et erilisi erinevusi imputeerimismeetodite tulemuste vahel ei esine. Erinevate meetodite korral on tunnuse *vanus* kvartiilid ja mediaan tulnud küll samad, kuid väikeseid erinevusi esineb eelkõige valimite minimaalsetetes ja maksimaalsetes väärtustes, mille korral on kõige parema tulemuse andnud lähima naabri meetod (*knn*). Üldise *Hot-Deck* meetodi (*hot-deck*) korral on valimi maksimaalne väärtus veidi suurenenud. Grupipõhise *Hot-Deck* meetodi ja lähima naabri meetodi kombinatsiooni (*hot-deck&knn*) korral on nii valimi tunnuse *vanus* maksimaalne väärtus suurenenud kui ka minimaalne väärtus vähenenud. Punktidega märgitud vaatlused, mis on mediaanist kaugemal kui poolteist kvartiilide vahet, on jaotunud erinevate meetodite korral üsna sarnaselt.

Tabelis 5 on välja toodud 30% puuduvate andmete imputeerimisel saadud vanuste erinevuste karakteristikuid. Tulemusi on vaadeldud ainult imputeeritud andmete lõikes.

Tabel 5. Imputeerimistulemuste ja tegelike vanuste erinevuse karakteristikud imputeeritud andmete lõikes (vt Lisa 9.c).

	Keskmine	Standardh.	Min	Max	Med
Üldine juhuslik *Hot-Deck* meetod	-0.709	16.566	-53	57	-1
Lähima naabri meetod	0.215	16.666	-69	64	0
Juhuslik *Hot-Deck* omistus klassis ja lähima naabri meetod	-0.11	16.426	-69	64	1

Uurides tabelit 5 on võimalik näha, et kõige väiksema keskmise vanuse hinnangu nihe tuli juhusliku *Hot-Deck* omistuse korral klassis kombineerituna lähima naabri meetodiga: -0.11. Kõige suurem nihe -0.709 tuli aga üldise juhusliku *Hot-Deck* meetodi korral ja keskmise tulemuse 0.215 andis lähima naabri meetod. Kõige suurema vanuse erinevuse varieeruvuse andis aga lähima naabri meetod ja seda 16.666. Natuke parema tulemuse andsid üldine juhuslik *Hot-Deck* meetod tulemusega 16.566 ja grupipõhine juhuslik *Hot-Deck* omistus kombineerituna lähima naabri meetodiga tulemusega 16.426. Seejuures tasub mainida, et üldise juhusliku *Hot-Deck* omistuse korral minimaalne ja maksimaalne erinevus on absoluutväärtuselt kõige väiksemad – vastavalt -53 ja 57. Ülejäänud kahe meetodi korral oli minimaalseks vanuste erinevuseks -69 ja maksimaalseks 64. Kõige parema mediaani andis lähima naabri meetod, milleks oli 0, üldise juhusliku *Hot-Deck* meetodi korral tuli karakteristiku väärtuseks 1 ja kolmandana vaadeldud meetodi korral tuli selle väärtuseks -1.

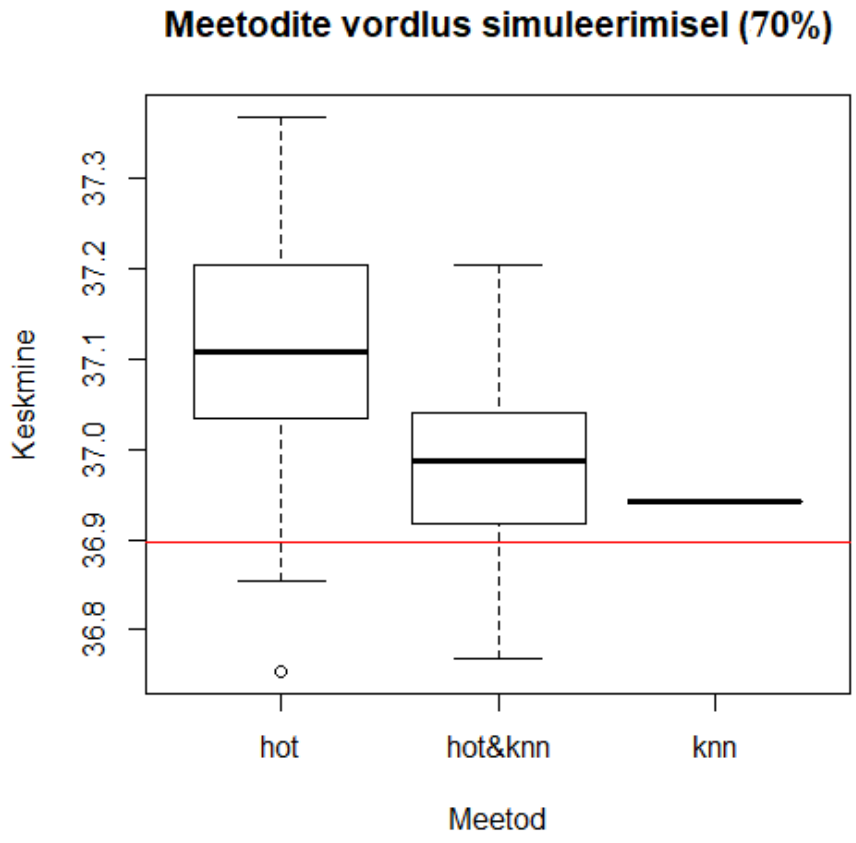


Joonis 4. Imputeerimistulemuste ja tegelike vanuste erinevuse karpdiagramm imputeeritud andmete lõikes (vt Lisa 9.c).

Vaadates joonise 4 karpdiagrammi on näha, et tegelike vanuste ja imputeeritud vanuste vahede karakteristikud erinevate meetodite korral on üsna sarnased. Punktidena märgitud vaatluste puhul on näha, et üldise *Hot-Deck* meetodi (*hot-deck*) korral on erindid mitte nii erinevad nagu teiste meetodite puhul (oli näha ka tabelist 5).

5.2 Simulatsioon vähemalt 70% info olemasolu korral

Esimese simuleerimise katse puhul püüti imputeerida andmeid erinevate meetoditega ühtedele ja samadele andmetele, st üks kord oli läbi viidud 70% andmete allesjätmine ja peale seda 99 korda imputeeritud andmed (vt Lisa 9.d). Ühe simulatsiooni sammuna oli arvesse võetud ka eelmise imputeerimise tulemusi – seega viidi läbi kokku 100 simulatsiooni.

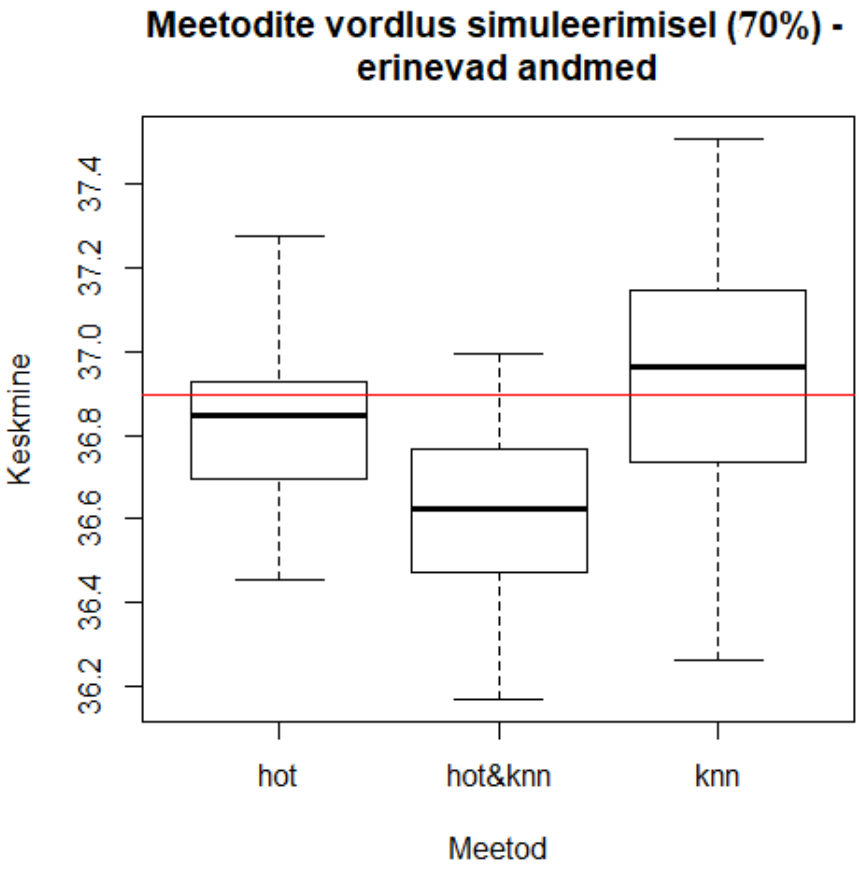


Joonis 5. Meetodite võrdlus simuleerimisel (70%) - samad algandmed (vt Lisa 9.d).

Vaadates joonist 5 on võimalik näha, et lähima naabri meetod (*knn*) annab igal simulatsiooni sammul sama tulemuse. Põhjus võib olla selles, et imputeerimisel omistatakse puuduvatele väärtustele alati samade naabrite väärtused (vt ptk 1.6). Vaatamata sellele on vaadeldava imputeerimismeetodiga saadud tulemuste keskmiste mediaan kõige lähemal tegelikule keskmisele (joonisel tähistatud punase

joonega). Paremuselt järgmise tulemuse andis grupipõhine juhuslik *Hot-Deck* omistus kombineerituna lähima naabri meetodiga (*hot&knn*), sest imputeerimise tulemuste keskmiste kvartiilid ja mediaan on lähemal kui üldise juhusliku *Hot-Deck* meetodi (*hot*) korral. Kahjuks antud joonise põhjal polnud võimalik veel mudelite paremuse osas otsuseid langetada.

Teise simulatsiooni puhul imputeeriti andmeid igal 99-l sammul erinevatele algandmetele, st igal imputeerimise katsel oli kustutatud ligikaudu 30% andmeid ja peale seda imputeeriti andmed erinevate meetoditega (vt Lisa 9.e). Ühe simulatsiooni sammuna oli ka siin arvesse võetud esimese imputeerimise tulemusi – järelikult tehti kokku 100 simulatsiooni. Antud katse viidi läbi eesmärgiga uurida lähemalt lähima naabri meetodit.



Joonis 6. Meetodite võrdlus simuleerimisel (70%) - erinevad algandmed (vt Lisa 9.e).

Jooniselt 6 on võimalik näha, et lähima naabri meetod (*knn*) ja üldine juhuslik *Hot-Deck* meetod (*hot*) annavad kõige paremaid tulemusi, sest vaadeldavate imputeerimismeetoditega saadud tulemuste mediaanid on kõige lähemal tegelikule keskmisele (joonisel tähistatud punase joonega). Üldise *Hot-Deck* meetodi korral on vaadeldava statistiku ülemine kvartiil ja lähima naabri meetodi korral alumine kvartiil lähemal tegelikule väärtusele. Tasub veel välja tuua, et lähima naabri meetodiga imputeeritud tulemuste keskmised on kõige suurema varieeruvusega, mis pani uurijaid kahtlema vaadeldava meetodi stabiilsuses. Seetõttu vaadeldavast katsest tehti järeldus, et kõige paremaid tulemusi andis just üldine juhuslik *Hot-Deck* omistus. Kõige ebatäpsemad tulemused andis grupipõhine juhuslik *Hot-Deck* omistus kombineerituna lähima naabri meetodiga (*hot&knn*).

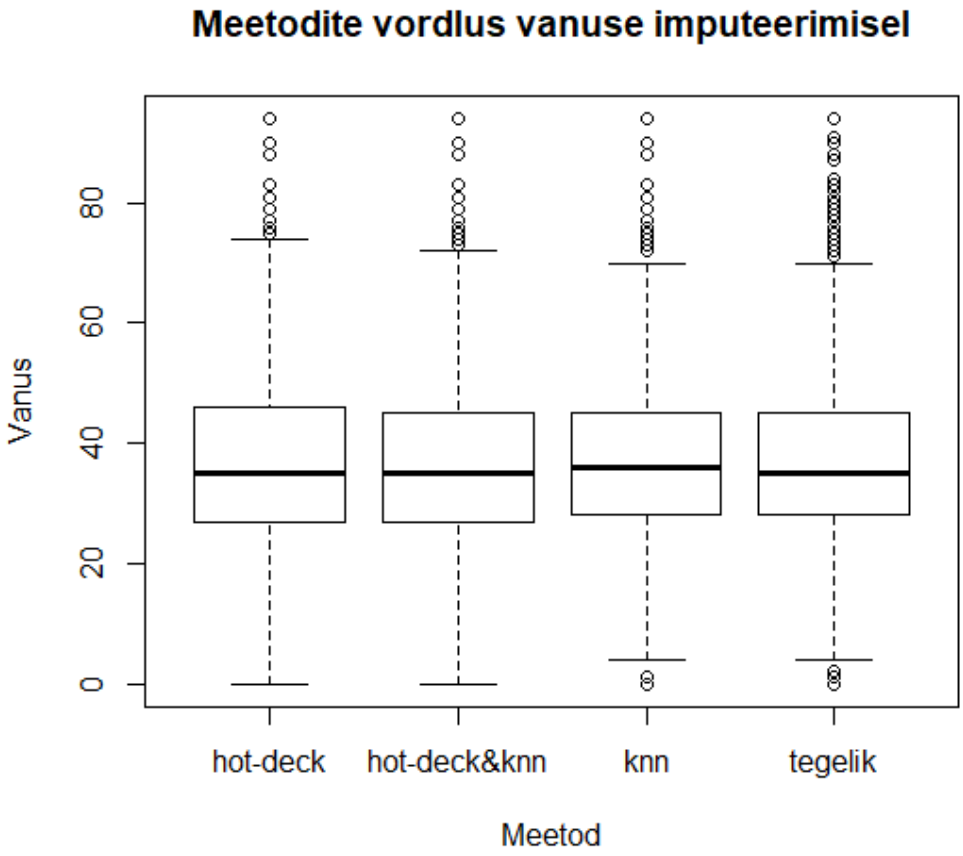
5.3 Imputeerimise kvaliteet vähemalt 50% info olemasolu korral

Tabelis 6 on välja toodud 50% puuduvate andmete imputeerimisel saadud tulemuste karakteristikuid.

Tabel 6. Vanuse karakteristikud kõikide andmete lõikes (vt Lisa 10.c).

	Keskmine	Standardh.	Min	Max	Med
Üldine juhuslik *Hot-Deck* meetod	36.909	11.949	0	94	35
Lähima naabri meetod	37.158	11.761	0	94	36
Juhuslik *Hot-Deck* omistus klassis ja lähima naabri meetod	36.488	12.075	0	94	35
Tegelik	36.897	11.858	0	94	35

Tabelist 6 on näha, et lähima naabri meetodi korral vanuse mediaan on erinev tegelikkusest – 35 asemel on see 36. Vanuse miinimumid ja maksimumid on jäänud muutumatuks. Tegelik tunnuse *vanus* hajuvus oli 11.858. Tabelist on näha, et lähima naabri imputeerimismeetodi korral on vaadeldava karakteristiku hajuvus isegi vähenenud ning seda 0.097 võrra. Teiste meetodite korral on see tõusnud – üldise *Hot-Deck* meetodi korral 0.091 võrra ning juhusliku doonorgrupipõhise *Hot-Deck* ja lähima naabri meetodi kombinatsiooni korral 0.217 võrra. Tegelik keskmine vanus oli 36.897 aastat ja suuri muutusi imputeerimisel vaadeldava karakteristiku väärtuses ei esinenud. Kõige täpsema keskmise andsid üldise juhusliku *Hot-Deck* meetodiga imputeeritud andmed – 36.909. Lähima naabri meetodi korral tuli keskmiseks vanuseks 37.158 ning juhusliku doonorgrupipõhise *Hot-Deck* ja lähima naabri meetodi kombinatsiooni korral tuli vaadeldava karakteristiku väärtuseks 36.488.



Joonis 7. Vanuse imputeerimistulemuste karpdiagramm kõikide andmete lõikes (vt Lisa 10.c).

Joonise 7 karpdiagrammilt on näha, et suuri erinevusi imputeerimistulemuste karakteristikutes ei esine. Kõikide meetodite korral on tunnuse *vanus* kvartiilid ja mediaan tulnud samad, mis tegelike andmete korral. Valimite minimaalsetetes ja maksimaalsetes tunnuse *vanus* väärtustes (karpdiagrammide vurrud) on veidi parema tulemuse andnud lähima naabri meetod (*knn*). Üldise *Hot-Deck* meetodi (*hot-deck*) korral on valimi maksimaalne väärtus suurenenud ja minimaalne väärtus vähenenud. Grupipõhise *Hot-Deck* meetodi ja lähima naabri meetodi (*hot-deck&knn*) korral on tulemused aga paremad kui eelmises katses – valimi vaadeldava tunnuse minimaalne väärtus on siiski vähenenud, kuid maksimaalne väärtus ei erine tegelikkusest enam nii palju. Põhjus võib olla selles, et kui puudu oli 50% andmetest, siis moodustus palju rohkem tühje gruppe, mistõttu rakendati suuremal hulgal andmetel siiski lähima naabri meetodit. See aga parandas saadavaid

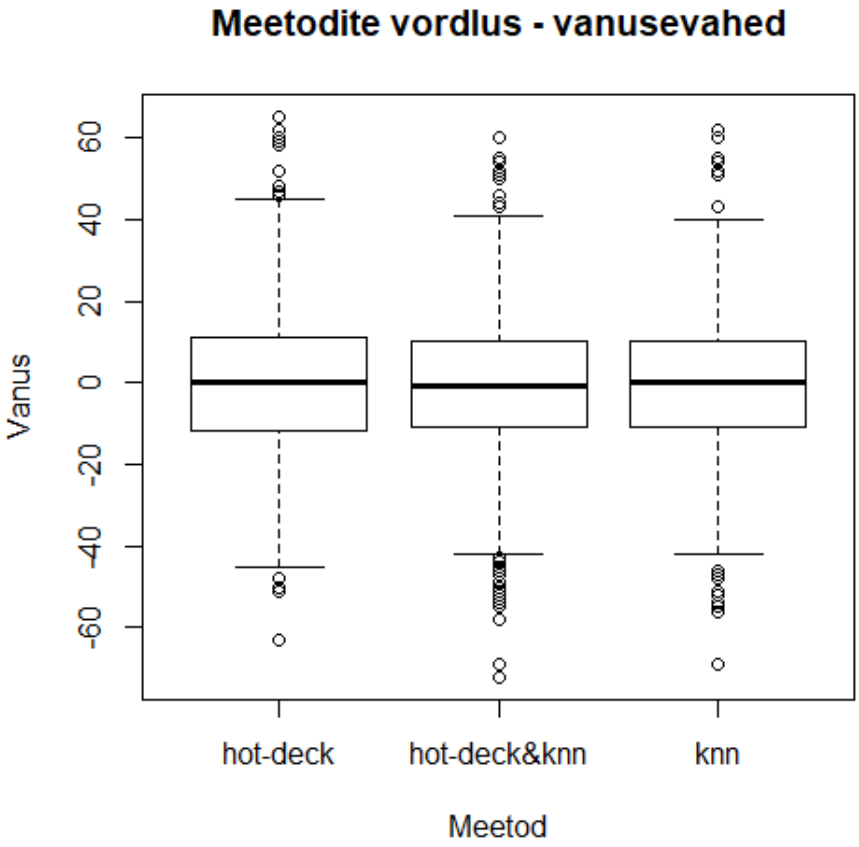
tulemusi. Punktidega märgitud vaatlused on jaotunud erinevate meetodite korral üsna sarnaselt.

Tabelis 7 on välja toodud 50% puuduvate andmete imputeerimisel saadud vanuste erinevuste karakteristikuid. Tulemusi on vaadeldud ainult imputeeritud andmete lõikes.

Tabel 7. Imputeerimistulemuste ja tegelike vanuste erinevuse karakteristikud imputeeritud andmete lõikes (vt Lisa 10.c).

	Keskmine	Standardh.	Min	Max	Med
Üldine juhuslik *Hot-Deck* meetod	-0.212	16.839	-63	65	0
Lähima naabri meetod	-0.619	16.128	-69	62	0
Juhuslik *Hot-Deck* omistus klassis ja lähima naabri meetod	-0.774	16.606	-72	60	-1

Vaadates tabelit 7 on võimalik näha, et kõige väiksema keskmise vanuse hinnangu nihe tuli käesolevas katses üldise juhusliku *Hot-Deck* omistuse korral: -0.212. Kõige suurem nihe -0.744 tuli aga grupipõhise juhusliku *Hot-Deck* meetodi korral kombineerituna lähima naabri meetodiga ning keskmise tulemuse -0.619 andis lähima naabri meetod. Kõige väiksema vanuse erinevuse varieeruvuse andis lähima naabri meetod ja seda 16.128. Halvemad tulemused andsid üldine juhuslik *Hot-Deck* meetod tulemusega 16.839 ja grupipõhine juhuslik *Hot-Deck* omistus kombineerituna lähima naabri meetodiga väärtusega 16.606. Üldise juhusliku *Hot-Deck* omistuse korral minimaalne ja maksimaalne vanuste erinevus on -63 ja 65. Lähima naabri meetodi korral vaadeldavate karakteristikute väärtusteks olid vastavalt -69 ja 62 ning kolmandana vaadeldud meetodi korral -72 ja 60. Kõige parema mediaani andsid üldine juhuslik *Hot-Deck* ja lähima naabri meetod, milleks oli 0. Grupipõhise juhusliku *Hot-Deck* meetodi kombineerituna lähima naabri meetodiga korral tuli karakteristiku väärtuseks taaskord -1.

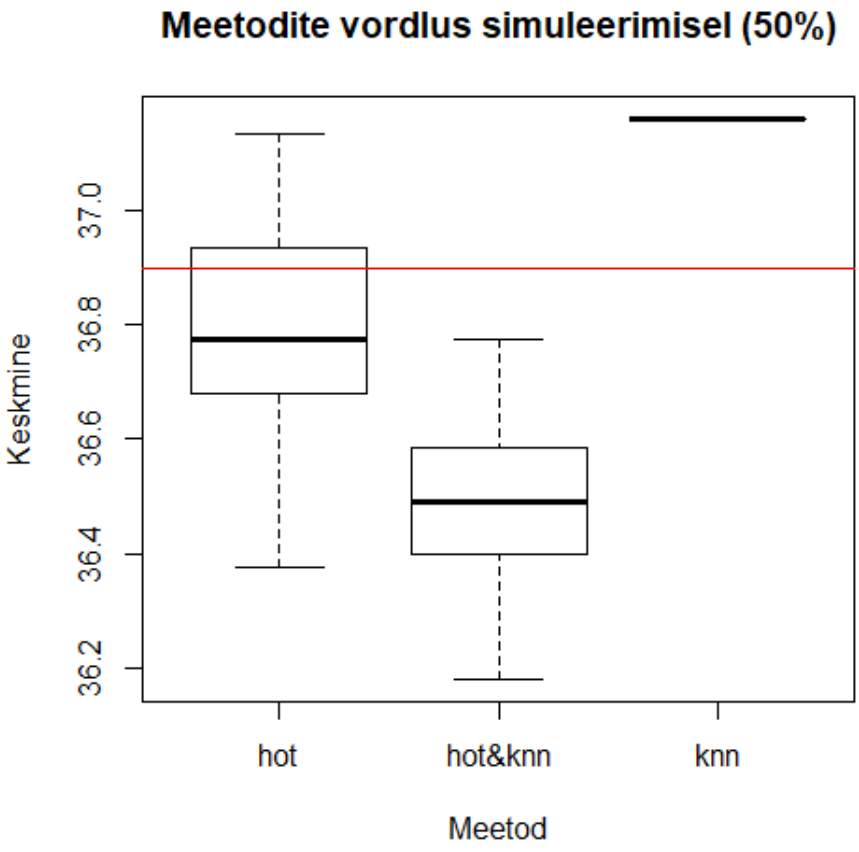


Joonis 8. Imputeerimistulemuste ja tegelike vanuste erinevuse karpdiagramm imputeeritud andmete lõikes (vt Lisa 10.c).

Uurides joonise 8 karpdiagrammi on näha, et tegelike ja imputeeritud vanuste vahede karakteristikud on erinevate meetodite korral taaskord üsna sarnased. Grupipõhise juhusliku *Hot-Deck* meetodi kombineerituna lähima naabri meetodiga (*hot-deck&knn*) ja lähima naabri meetod (*knn*) annavad ka sarnaseid valimi minimaalseid ja maksimaalseid väärtuseid. Põhjus võib olla ka selles, et esmalt mainitud meetodi korral moodustus palju tühje gruppe, mistõttu rakendati suuremal hulgal andmetel siiski lähima naabri meetodit. Üldise *Hot-Deck* meetodi (*hot-deck*) korral on valimi vaadeldava tunnuse maksimaalne väärtus mõnevõrra suurem ja minimaalne väärtus pisut väiksem kui teiste meetodite korral.

5.4 Simulatsioon vähemalt 50% info olemasolu korral

Esimese katse puhul püüti imputeerida andmeid erinevate meetoditega samadele andmetele, st üks kord oli läbi viidud 50% andmete allesjätmine ja peale seda imputeeritud 100 korda andmeid (vt Lisa 10.d). Ühe simulatsiooni sammuna oli arvesse võetud ka eelmise imputeerimise tulemusi – seega viidi läbi kokku 100 simulatsiooni.

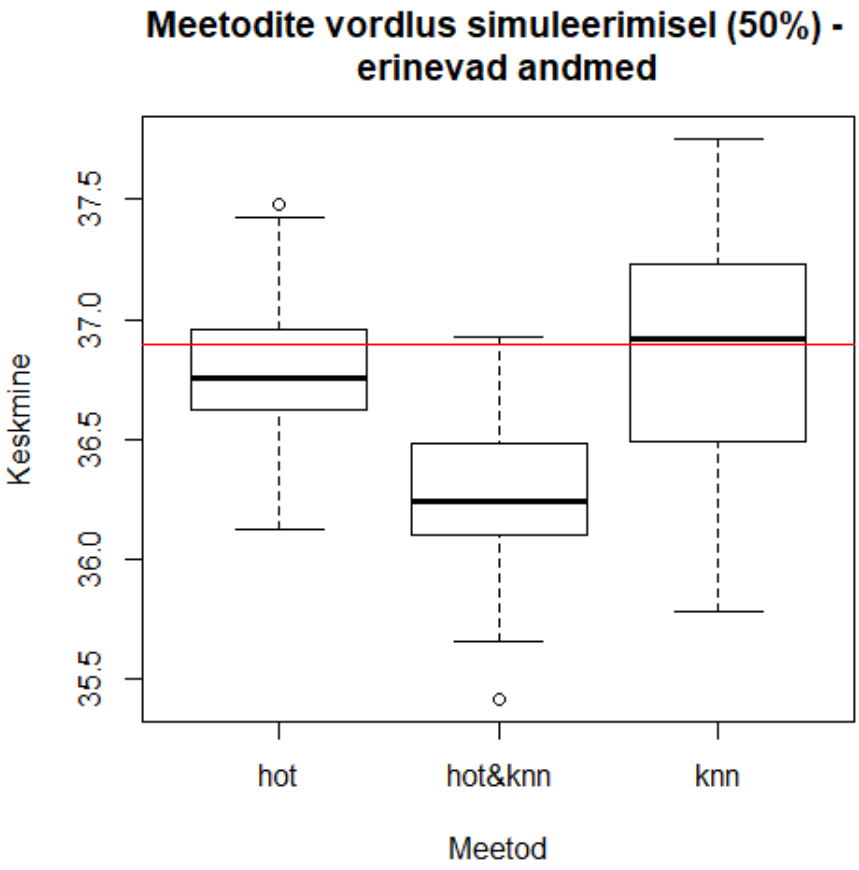


Joonis 9. Meetodite võrdlus simuleerimisel (50%) *- samad algandmed (vt Lisa 10.d).*

Vaadates joonist 9 on võimalik näha, et lähima naabri meetod (*knn*) annab jällegi igal simulatsiooni sammul sama tulemuse, kuna igal sammul omistatakse puuduvatele väärtustele samu vanuseid (vt ptk 1.6). Käesoleva karpdiagrammi põhjal saab väita, et kõige täpsemaid imputeerimise tulemusi annab üldine juhuslik *Hot-Deck* omistus (*hot*), mille korral vanuste keskmiste mediaan ja kvartiilid on kõige

lähemal tegelikule keskmisele vanusele (märgitud punase joonega). Grupipõhine juhuslik *Hot-Deck* omistus kombineerituna lähima naabri meetodiga (*hot&knn*) annab antud juhul väga ebatäpseid tulemusi. Kahjuks polnud vaadeldava joonise põhjal võimalik mudelite headuse osas otsuseid langetada, kuna puudus ka seekord hea ülevaade lähima naabri meetodi töökindluses.

Teise katse puhul püüti imputeerida andmeid igal sammul erinevatele algandmetele, st igal imputeerimise katsel oli kustutatud ligikaudu 50% andmeid ja peale seda imputeeriti andmeid erinevate meetoditega (vt Lisa 10.e). Ühe simulatsiooni sammuna oli ka siin arvesse võetud esimese imputeerimise tulemusi – järelikult viidi läbi kokku 100 simulatsiooni. Vaadeldav protsess tehti taaskord läbi selleks, et lähemalt uurida lähima naabri meetodi headust.



Joonis 10. Meetodite võrdlus simuleerimisel (50%) - erinevad algandmed (vt Lisa 10.e).

Jooniselt 10 on võimalik näha, et lähima naabri meetod (*knn*) on simuleerimisel andnud kõige täpsema vanuste keskmiste mediaani, sest see statistik on kõige lähemal tegelikule keskmisele (joonisel tähistatud punase joonega). Paremuselt järgmise tulemuse on mediaani puhul andnud üldine *Hot-Deck* meetod ja kõige halvemaid tulemusi grupipõhine juhuslik *Hot-Deck* omistus kombineerituna lähima naabri meetodiga (*hot&knn*). Veel tasub mainida, et üldise *Hot-Deck* meetodi korral on keskmiste kvartiilid tegelikule vanuste keskmisele väärtusele lähemal kui lähima naabri meetodi korral. Selle põhjal otsustati, et lähima naabri meetod annab ebastabiilsemaid tulemusi ja mõttekam oleks ka siin rakendada just üldist juhuslikku *Hot-Deck* omistust.

KOKKUVÕTE

Eesti tervishoiuteenuste osutajad esitavad ravijuhtude kohta dokumentatsiooni mitmesse erinevasse süsteemi. On otsustatud statistika tegemisel üle minna ühele andmeesitussüsteemile, milleks on tervise infosüsteem, kuid vaadeldavasse andmebaasi saadetakse kahjuks liiga puudulikke andmeid. Kõige põhjalikumalt esitatakse teenuseosutajate poolt dokumentatsiooni Eesti Haigekassale, sest saadetud andmete põhjal makstakse neile institutsiooni poolt raha. Seetõttu on otsustatud antud analüüsis kasutada just viimasena mainitud andmebaasi andmeid.

Töö eesmärgiks oli Haigekassa andmebaasi andmete täiustamine kasutades lisainformatsioonina tervise infosüsteemi andmebaasi esitatud andmeid. Töö põhiline probleem seisnes selles, et Haigekassa andmete põhjal pole kahjuks võimalik teha täielikku demograafilist statistikat, kuna ravikindlustuseta patsientide raviarvetel pole märgitud ravitava inimese vanust.

Analüüsis kasutatavale andmebaasi ravijuhtude dokumentidele patsiendi vanuse leidmiseks otsustati ühendada omavahel Haigekassale ja tervise infosüsteemi esitatud andmed. Ülesannet raskendas asjaolu, et kahe andmestiku piires oli ühele ja samale patsiendile genereeritud erinev ID-kood, mistõttu vaadeldavat tunnust ei saanud ühendamisel kasutada. Tervise Arengu Instituudi (TAI) analüütikute meeskonna poolt võeti vastu otsus kasutada andmete ühendamiseks teisi andmestikes leiduvaid tunnuseid. Protsessi käigus avastati, et tervishoiuteenuste osutajad esitavad mõlemasse andmebaasi ühe ja sama ravijuhu andmeid erinevalt. Selleks, et ühendamist siiski läbi viia, valiti TAI analüütikute eksperthinnangu põhjal välja 187 erinevat ühendamisel kaasatavate tunnuse kombinatsiooni ja reeglit. Kui andmetes esines duplikaate ehk ühendamiseks liiga sarnaseid ravijuhte, siis käesolevas analüüsis võeti TAI analüütikute poolt vastu otsus selliseid vaatlusi mitte ühendada. Ühendamise protsessis suudeti leida Haigekassa andmebaasis 5633-le ravijuhu patsiendile vanus. Patsiendi vanuseta jäi 3515 ravijuhu raviarvet.

Nendele andmetele, millele kahe andmebaasi ühendamise protsessis patsiendi vanust ei suudetud leida, otsustati vanused olemasolevate andmete põhjal imputeerida. Selleks kasutati kolme erinevat viisi: üldine juhuslik *Hot-Deck* meetod, lähima naabri meetod ja doonorgrupipõhine juhuslik *Hot-Deck* omistus kombineerituna lähima naabri meetodiga. Viimasena mainitud meetodi korral rakendati lähima naabri meetodit ainult nende vaatluste korral, mis sattusid sellistesse doonorgruppidesse, kus ühegi ravijuhu korral patsiendi vanust ei leidunud. Kokku esines puuduvaid väärtusi vigastuse välispõhjuse diagnoosi liigis (9), soo tunnuses (37) ja vanuses (3515). Väärtusi imputeeriti alustades sellest tunnusest, milles esines kõi-

ge vähem puuduvaid väärtuseid ja iga järgneva tunnuse imputeerimisel kasutati eelmise imputeerimise tulemusi.

Kahjuks polnud vaadeldavate meetodite paremust võimalik peale imputeerimist kontrollida. Seetõttu otsustati läbi viia kaks katset, kus esimesel juhul olemasolevate andmete hulgast jäeti alles 70% ja teisel juhul 50% tunnuse *vanus* väärtustest. Puuduolevatele andmetele viidi algselt vaadeldavate meetoditega läbi üks imputeerimine. Seejärel otsustati mõlemal juhul läbi viia kaks erinevat simulatsiooni, kus esimesel juhul imputeeriti andmeid igal simulatsiooni sammul samadele andmetele ja teisel juhul viidi läbi 70% või 50% andmete selekteerimine iga simulatsiooni sammu alguses. Esimese variandi puhul ei olnud kahjuks võimalik uurida lähima naabri meetodi paremust. Küll aga teise variandi põhjal saadi teada, et kõige stabiilsemaid tulemusi vanuste imputeerimisel annab üldine juhuslik *Hot-Deck* meetod. Seetõttu otsustati ka käesolevas töös rakendada praktikas just selle meetodiga imputeeritud andmeid.

Edaspidi on Tervise Arengu Instituudil plaanis teha sarnane analüüs, kuid tahetakse Haigekassast ja tervise infosüsteemist taotleda selliseid andmeid, kus mitme andmestiku piires oleks patsientidele genereeritud ID-koodid ühesuguselt. Järgmises analüüsis soovitakse ühendada omavahel veel ka andmestikes esinevaid duplikaate, mis antud töös jäi ekspertide otsuse tõttu tegemata.

Viited

[1] Andridge, R. R., Little , R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *Int Stat Rev.* 2010 aprill, 78(1), 40–64. doi: 10.1111/j.1751-5823.2010.00103.x

[2] Bhaskaran, K., Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *Int J Epidemiol.* 2014 Aug; 43(4): 1336–1339. doi: 10.1093/ije/dyu080

[3] Bogovsi, P. (1996). *RHK-10: Rahvusvaheline haiguste ja nendega seotud terviseprobleemide statistiline klassifikatsioon.* Tallinn: Tallinna Raamatutrükikoda.

[4] Eesti Keele Sihtasutus: *Eesti õigekeelsussõnaraamat ÕS 2013* (2013). Kasutatud 20.02.2018. http://www.eki.ee/dict/qs/

[5] Kirpu, V., Eigo, N. (2018). *Andmekadu ja vead ning nendega kaasnevate takistuste lahendamine.* Kasutatud 10.04.2018. http://www.tai.ee/images/Andmekadu_artikkel.pdf

[6] Kowarik, A., Templ , M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software.* 2016 oktoober, 74(7). doi: 10.18637/jss.v074.i07

[7] Lepik, N., Traat, I. (2016). *Tõenäosuslik valikuuring I* . Loengukonspekt. Tartu Ülikool, matemaatilise statistika instituut. Kasutatud 08.02.2018. https://courses.ms.ut.ee/MTMS.01.003/2016_fall/uploads/Main/loengud 2016.pdf

[8] Prostakova, J. (2007). *Mittevastamine ja selle kompenseerimine.* Bakalaureusetöö. Tartu Ülikool, matemaatilise statistika instituut.

[9] Toompere, K. (2009). *Imputeerimis- ja kaalumismeetodite mõju hinnangute nihkele.* Magistritöö. Tartu Ülikool, matemaatilise statistika instituut.

LISAD

Lisa 1. Vigastuste põhidiagnoosi koodide selgitused RHK-10 järgi.

Nr	Diagnoosigrupi nimetus	Koodide vahemik	Tähendus
1	Pea- ja kehapiirkonna vigastused	*S00–S09*	Peavigastused
		S10–S19	Kaelavigastused
		S20–S29	Rindkerevigastused
		S30–S39	Kõhu, selja alaosa, lülisamba nimmeosa ja vaagna vigastused
2	Kätepiirkonna vigastused	*S40–S49*	Õla- ja õlavarrevigastused
		S50–S59	Küünarpiirkonna- ja küünarvarrevigastused
		S60–S69	Randme- ja käevigastused
3	Jalapiirkonna vigastused	*S70–S79*	Puusa- ja reievigastused
		S80–S89	Põlve- ja säärevigastused
		S90–S99	Kanna- ja jalavigastused
4	Muud täpsustamata piirkonna ja muud liiki vigastused või tüsistused	*T00–T07*	Mitut kehapiirkonda haaravad vigastused
		T08–T14	Kere, jäsemete või keha täpsustamata piirkonna vigastused
		T15–T19	Loomuliku kehaava kaudu sisenenud võõrkeha toime
		T20–T25	Keha välispinna täpsustatud paikme põletused ja söövitused
		T20–T32	Põletused ja söövitused
		T26–T28	Silma ning siseelunditega piirdunud põletused ja söövitused
		T29–T32	Mitme ning täpsustamata kehapiirkonna põletused ja söövitused
		T33–T35	Külmumused
		T36–T50	Mürgistused rohtude, ravimite ja bioloogiliste ainetega
		T51–T65	Peamiselt mittemeditsiinilise päritoluga ainete toksiline toime
		T66–T78	Muude ja täpsustamata välispõhjuste toime
		T79	Trauma teatavad varajased tüsistused
		T80–T88	Mujal klassifitseerimata kirurgilise ja muu meditsiiniabi tüsistused
		T90–T98	Vigastuste, mürgistuste ja välispõhjuste toime muude tagajärgede jääknähud

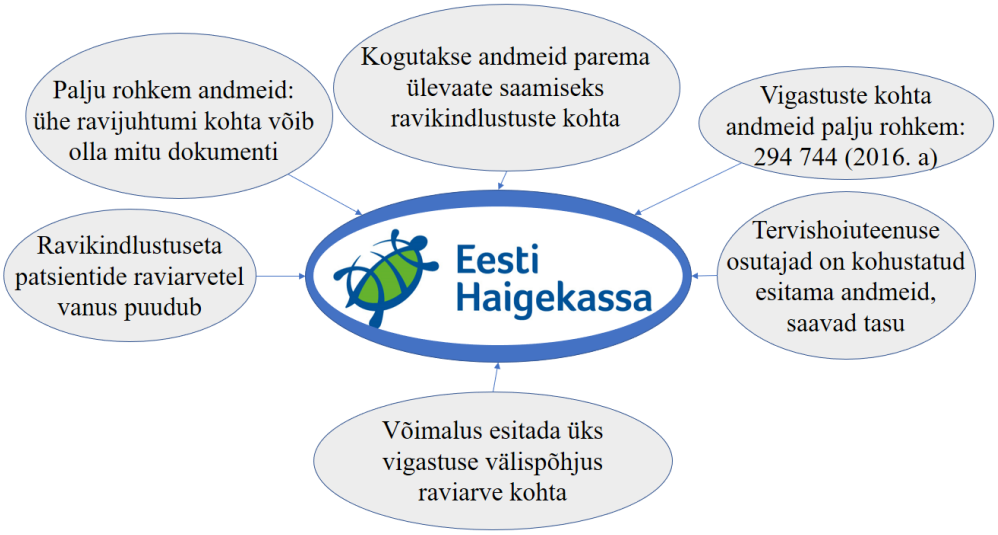
Lisa 2. Vigastuste välispõhjuse koodide tähendused RHK-10 järgi.

Nr	Diagnoosigrupi nimetus	Koodide vahemik	Tähendus
1	Sõidukiõnnetuses saadud vigastuste välispõhjused	*V01–V09*	Sõidukiõnnetuses vigastatud jalakäija
		V10–V19	Sõidukiõnnetuses vigastatud jalgrattur
		V20–V29	Sõidukiõnnetuses vigastatud mootorrattur
		V30–V39	Sõidukiõnnetuses vigastatud kolmerattalisel mootorsõidukil sõitja
		V40–V49	Sõidukiõnnetuses vigastatud sõiduautosõitja
		V50–V59	Sõidukiõnnetuses vigastatud pikapi- või veoautosõitja
		V60–V69	Sõidukiõnnetuses vigastatud raskeveoautosõitja
		V70–V79	Sõidukiõnnetuses vigastatud bussisõitja
		V80–V89	Muud maismaaliiklus-ja sõidukiõnnetused
		V90–V94	Veesõidukiõnnetused
		V95–V97	Kosmose- ja õhusõidukiõnnetused
		V98–V99	Muud ja täpsustamata sõidukiõnnetused
2	Füüsikaliste faktorite poolt põhjustatud vigastuste välispõhjused (kukkumised, elekter jms)	*W00–W19*	Kukkumised
		W20–W49	Eluta mehhaanilise jõu toime
		W50–W64	Elusolendi mehhaanilise jõu toime
		W65–W74	Juhuslik uppumine ja vee alla vajumine
		W75–W84	Muu juhuslik hingamisohustus
		W85–W99	Elektrivoolu, kiirguse ja ümbritseva õhu äärmuslike temperatuuride ning rõhu toime
3	Loodusnähtustest ja muudest teguritest põhjustatud vigastuste välispõhjused (tuli, põletused, mürgitused, ülepingutus jms)	*X00–X09*	Suitsu, tule ja leekide toime
		X10–X19	Kokkupuude kuumuse ja tuliste esemetega
		X20–X29	Kokkupuude mürgiste loomade ja taimedega
		X30–X39	Loodusjõudude toime
		X40–X49	Juhuslik mürgistus kahjulike ainetega ja nende toime
		X50–X57	Ülepingutus, reisimine ja puudusseisundid
		X58–X59	Muude ja täpsustamata tegurite juhuslik toime

4	Inimese ja muude faktorite poolt põhjustatud vigastuste välispõhjused	*X60–X84*	Tahtlik enesekahjustus
		X85–Y09	Rünne
		Y10–Y34	Ebaselge tahtlusega sündmus
		Y35–Y36	Seaduslik sekkumine ja sõjategevus
		Y40–Y59	Ravimisel ebasoodsat toimet avaldavad rohud, ravimid ja bioloogilised ained
		Y40–Y84	Kirurgilise või muu meditsiiniabi tüsistused
		Y60–Y69 *Y70–Y82*	Äpardused kirurgilise või muu meditsiiniabi korral Diagnoosimisel ja ravimisel kasutatud meditsiiniliste seadistega seotud äpardused
		Y83–Y84	Kirurgilised või muud meditsiinilised menetlused patsiendi ebaloomuliku reaktsiooni või hilisema tüsistuse põhjusena menetluseaegset äpardust mainimata
		Y85–Y89	Haigestumise ja surma välispõhjuste toime hilisnähud
		Y90–Y98	Haigestumise ja surma mujal klassifitseeritud põhjustega seotud lisategurid

Lisa 3. Tervise infosüsteemi ja Haigekassa andmete eripärad.

Lisa 3.a. Haigekassa andmete eripärad.



Lisa 3.b. Tervise infosüsteemi (Digilugu) andmete eripärad.

Lisa 4. Ühendamise kombinatsioonid.

Juht	Nr	Tunnused, mis võivad erineda	Palju ühendas
1	0	**Kõik kattub** (vanusega)	11 361
	1	Põhidiagnoosi koodi 5 esimest sümbolit	2
	2	Välispõhjuse koodi 5 esimest sümbolit	23
	3	Mõlema koodi 5 esimest sümbolit	0
2	0	***vanus* puudub**	382
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	2
	3	Mõlema koodi 5 esimest sümbolit	0
3	0	***ravikindl*** (vanusega)	148
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
4	0	***ravikindl*; *vanus* puudub**	9
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
5	0	***valtimatu*** (vanusega)	119 245
	1	Põhidiagnoosi koodi 5 esimest sümbolit	5
	2	Välispõhjuse koodi 5 esimest sümbolit	452
	3	Mõlema koodi 5 esimest sümbolit	0
6	0	***valtimatu*; *vanus* puudub**	3 459
	1	Põhidiagnoosi koodi 5 esimest sümbolit	1
	2	Välispõhjuse koodi 5 esimest sümbolit	35
	3	Mõlema koodi 5 esimest sümbolit	0
7	0	***algus*** (vanusega)	10
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
8	0	***algus*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

9	0	***lopp*** (vanusega)	61
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
10	0	***lopp*****; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
11	0	***mk*** (vanusega)	131
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	1
	3	Mõlema koodi 5 esimest sümbolit	0
12	0	***mk*****; *vanus* puudub**	37
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
13	0	***tuup*** (vanusega)	14
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
14	0	***tuup*****; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
15	0	***ravikindl*****, *valtimatu*** (vanusega)	372
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
16	0	***ravikindl*****, *valtimatu*****; *vanus* puudub**	34
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	2
	3	Mõlema koodi 5 esimest sümbolit	0
17	0	***ravikindl*****, *algus*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
18	0	***ravikindl*****, *algus*****; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

19	0	***ravikindl*, *lopp*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
20	0	***ravikindl*, *lopp*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
21	0	***ravikindl*, *mk*** (vanusega)	2
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
22	0	***ravikindl*, *mk*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
23	0	***ravikindl*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
24	0	***ravikindl*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
25	0	***valtimatu*, *algus*** (vanusega)	4 148
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
26	0	***valtimatu*, *algus*; *vanus* puudub**	113
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
27	0	***valtimatu*, *lopp*** (vanusega)	3 784
	1	Põhidiagnoosi koodi 5 esimest sümbolit	1
	2	Välispõhjuse koodi 5 esimest sümbolit	3
	3	Mõlema koodi 5 esimest sümbolit	0
28	0	***valtimatu*, *lopp*; *vanus* puudub**	101
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	1
	3	Mõlema koodi 5 esimest sümbolit	0

29	0	***valtimatu*, *mk*** (vanusega)	933
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	5
	3	Mõlema koodi 5 esimest sümbolit	0
30	0	***valtimatu*, *mk*; *vanus* puudub**	167
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	4
	3	Mõlema koodi 5 esimest sümbolit	0
31	0	***valtimatu*, *tuup*** (vanusega)	9
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
32	0	***valtimatu*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
33	0	***algus*, *lopp* (max vahe 30 päeva)** (vanusega)	3
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	1
34	0	***algus*, *lopp*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
35	0	***algus*, *mk*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
36	0	***algus*, *mk*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
37	0	***algus*, *tuup*** (vanusega)	1
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
38	0	***algus*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

39	0	***lopp*, *mk*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
40	0	***lopp*, *mk*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
41	0	***lopp*, *tuup*** (vanusega)	4
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
42	0	***lopp*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
43	0	***mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
44	0	***mk*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
45	0	***ravikindl*, *valtimatu*, *algus*** (vanusega)	11
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
46	0	***ravikindl*, *valtimatu*, *algus*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
47	0	***ravikindl*, *valtimatu*, *lopp*** (vanusega)	18
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
48	0	***ravikindl*, *valtimatu*, *lopp*; *vanus* puudub**	2
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

49	0	***ravikindl*, *valtimatu*, *mk*** (vanusega)	26
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
50	0	***ravikindl*, *valtimatu*, *mk*; *vanus* puudub**	2
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
51	0	***ravikindl*, *valtimatu*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
52	0	***ravikindl*, *valtimatu*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
53	0	***ravikindl*, *algus*, *lopp*** (vanusega)	1
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
54	0	***ravikindl*, *algus*, *lopp*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
55	0	***ravikindl*, *algus*, *mk*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
56	0	***ravikindl*, *algus*, *mk*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
57	0	***ravikindl*, *algus*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
58	0	***ravikindl*, *algus*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

59	0	***ravikindl*, *lopp*, *mk*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
60	0	***ravikindl*, *lopp*, *mk*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
61	0	***ravikindl*, *lopp*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
62	0	***ravikindl*, *lopp*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
63	0	***ravikindl*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
64	0	***ravikindl*, *mk*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
65	0	***valtimatu*, *algus*, *lopp*** (vanusega)	375
	1	Põhidiagnoosi koodi 5 esimest sümbolit	6
	2	Välispõhjuse koodi 5 esimest sümbolit	18
	3	Mõlema koodi 5 esimest sümbolit	0
66	0	***valtimatu*, *algus*, *lopp*; *vanus* puudub**	32
	1	Põhidiagnoosi koodi 5 esimest sümbolit	1
	2	Välispõhjuse koodi 5 esimest sümbolit	4
	3	Mõlema koodi 5 esimest sümbolit	1
67	0	***valtimatu*, *algus*, *mk*** (vanusega)	33
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
68	0	***valtimatu*, *algus*, *mk*; *vanus* puudub**	8
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

69	0	***valtimatu*, *algus*, *tuup*** (vanusega)	1
	1	Põhidiagnoosi koodi 5 esimest sümbolit	1
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
70	0	***valtimatu*, *algus*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
71	0	***valtimatu*, *lopp*, *mk*** (vanusega)	19
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
72	0	***valtimatu*, *lopp*, *mk*; *vanus* puudub**	6
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
73	0	***valtimatu*, *lopp*, *tuup*** (vanusega)	8
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
74	0	***valtimatu*, *lopp*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
75	0	***valtimatu*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
76	0	***valtimatu*, *mk*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
77	0	***algus*, *lopp*, *mk*** (vanusega)	1
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
78	0	***algus*, *lopp*, *mk*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

79	0	***algus*, *lopp*, *tuup*** (vanusega)	11
	1	Põhidiagnoosi koodi 5 esimest sümbolit	2
	2	Välispõhjuse koodi 5 esimest sümbolit	3
	3	Mõlema koodi 5 esimest sümbolit	0
80	0	***algus*, *lopp*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
81	0	***algus*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
82	0	***algus*, *mk*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
83	0	***lopp*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
84	0	***lopp*, *mk*, *tuup*; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
85	0	***ravikindl*, *valtimatu*, *algus*, *lopp*** (vanusega)	13
	1	Põhidiagnoosi koodi 5 esimest sümbolit	1
	2	Välispõhjuse koodi 5 esimest sümbolit	2
	3	Mõlema koodi 5 esimest sümbolit	0
86	0	***ravikindl*, *valtimatu*, *algus*, *mk*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
87	0	***ravikindl*, *valtimatu*, *algus*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
88	0	***ravikindl*, *valtimatu*, *lopp*, *mk*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

89	0	***ravikindl*, *valtimatu*, *lopp*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
90	0	***algus*, *lopp*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
91	0	***ravikindl*, *algus*, *lopp*, *mk*** (vanusega)	13
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	2
	3	Mõlema koodi 5 esimest sümbolit	0
92	0	***ravikindl*, *algus*, *lopp*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
93	0	***ravikindl*, *algus*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
94	0	***ravikindl*, *lopp*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
95	0	***valtimatu*, *algus*, *lopp*, *mk*** (vanusega)	15
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	3
	3	Mõlema koodi 5 esimest sümbolit	0
96	0	***valtimatu*, *algus*, *lopp*, *tuup*** (vanusega)	23
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	1
	3	Mõlema koodi 5 esimest sümbolit	1
97	0	***valtimatu*, *algus*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
98	0	***valtimatu*, *lopp*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

99	0	***algus*, *lopp*, *mk*, *tuup*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
100	0	***valispohjus*** (vanusega)	255
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
101	0	***valispohjus*; *vanus* puudub**	17
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
102	0	***pohidgn*** (vanusega)	1
	2	Välispõhjuse koodi 5 esimest sümbolit	0
103	0	***pohidgn*; *vanus* puudub**	4
	2	Välispõhjuse koodi 5 esimest sümbolit	0
104	0	***pohidgn*, *valtimatu*** (vanusega)	19
	2	Välispõhjuse koodi 5 esimest sümbolit	0
105	0	***pohidgn*, *valtimatu*; *vanus* puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
106	0	***valispohjus*, *valtimatu*** (vanusega)	9 648
	1	Põhidiagnoosi koodi 5 esimest sümbolit	1
107	0	***valispohjus*, *valtimatu*; *vanus* puudub**	251
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
108	0	***pohidgn* 3 esimest sümbolit** (vanusega)	20
	2	Välispõhjuse koodi 5 esimest sümbolit	0
109	0	***pohidgn* 3 esimest sümbolit; *vanus* puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
110	0	***valispohjus* 3 esimest sümbolit** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
111	0	***valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
112	0	***pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	2
113	0	***pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
114	0	***valtimatu*, *pohidgn* 3 esimest sümbolit** (vanusega)	357
	2	Välispõhjuse koodi 5 esimest sümbolit	50
115	0	***valtimatu*, *pohidgn* 3 esimest sümbolit; *vanus* puudub**	25
	2	Välispõhjuse koodi 5 esimest sümbolit	2
116	0	***valtimatu*, *valispohjus* 3 esimest sümbolit** (vanusega)	2
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
117	0	***valtimatu*, *valispohjus* 3 esimest sümbolit; *vanus* puudub**	3
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
118	0	***valtimatu*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	127

119	0	***valtimatu*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
120	0	***algus*, *pohidgn* 3 esimest sümbolit** (vanusega)	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
121	0	***algus*, *pohidgn* 3 esimest sümbolit; *vanus* puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
122	0	***lopp*, *pohidgn* 3 esimest sümbolit** (vanusega)	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
123	0	***lopp*, *pohidgn* 3 esimest sümbolit; *vanus* puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
124	0	***algus*, *lopp*, *pohidgn* 3 esimest sümbolit** (vanusega)	0
	2	Välispõhjuse koodi 5 esimest sümbolit	1
125	0	***algus*, *lopp*, *pohidgn* 3 esimest sümbolit; *vanus* puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
126	0	***algus*, *valispohjus* 3 esimest sümbolit** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
127	0	***algus*, *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
128	0	***lopp*, *valispohjus* 3 esimest sümbolit** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
129	0	***lopp*, *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
130	0	***algus*, *lopp*, *valispohjus* 3 esimest sümbolit** (vanusega)	1
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
131	0	***algus*, *lopp*, *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
132	0	***algus*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	1
133	0	***algus*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
134	0	***lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	0
135	0	***lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
136	0	***algus*, *lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	1
137	0	***algus*, *lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
138	0	***valtimatu*, *algus*, *pohidgn* 3 esimest sümbolit** (vanusega)	4
	2	Välispõhjuse koodi 5 esimest sümbolit	0
139	0	***valtimatu*, *algus*, *pohidgn* 3 esimest sümbolit; *vanus* puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
140	0	***valtimatu*, *lopp*, *pohidgn* 3 esimest sümbolit** (vanusega)	28
	2	Välispõhjuse koodi 5 esimest sümbolit	0
141	0	***valtimatu*, *lopp*, *pohidgn* 3 esimest sümbolit; *vanus* puudub**	1
	2	Välispõhjuse koodi 5 esimest sümbolit	0

142	0	***valtimatu*, *algus*, *lopp*, *pohidgn* 3 esimest sümbolit** (vanusega)	33
	2	Välispõhjuse koodi 5 esimest sümbolit	8
143	0	***valtimatu*, *algus*, *lopp*, *pohidgn* 3 esimest sümbolit; *vanus* puudub**	2
	2	Välispõhjuse koodi 5 esimest sümbolit	0
144	0	***valtimatu*, *algus*, *valispohjus* 3 esimest sümbolit** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
145	0	***valtimatu*, *algus*, *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
146	0	***valtimatu*, *lopp*, *valispohjus* 3 esimest sümbolit** (vanusega)	30
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
147	0	***valtimatu*, *lopp*, *valispohjus* 3 esimest sümbolit; *vanus* puudub**	4
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
148	0	***valtimatu*, *algus*, *lopp*, *valispohjus* 3 esimest sümbolit** (vanusega)	50
	1	Põhidiagnoosi koodi 5 esimest sümbolit	1
149	0	***valtimatu*, *algus*, *lopp*, *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
150	0	***valtimatu*, *algus*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	1
151	0	***valtimatu*, *algus*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
152	0	***valtimatu*, *lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	4
153	0	***valtimatu*, *lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	0
154	0	***valtimatu*, *algus*, *lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	35
155	0	***valtimatu*, *algus*, *lopp*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	1
156	0	***vanus* ± 1, *algus*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
157	0	***vanus* ± 1, *algus*, *lopp*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
158	0	***vanus* ± 1, *algus*, *pohidgn* 3 esimest sümbolit** (vanusega)	25
	2	Välispõhjuse koodi 5 esimest sümbolit	0
159	0	***vanus* ± 1, *algus*, *lopp*, *pohidgn* 3 esimest sümbolit** (vanusega)	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
160	0	***vanus* ± 1, *algus*, *valispohjus* 3 esimest sümbolit** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
161	0	***vanus* ± 1, *algus*, *lopp*, *valispohjus* 3 esimest sümbolit** (vanusega)	1
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
162	0	***vanus* ± 1, *algus*, *pohidgn* ja *valispohjus* 3 esimest sümbolit** (vanusega)	0

163	0	***vanus*** ± 1**,** ***algus*****,** ***lopp*****,** ***pohidgn*** **ja** ***valispohjus*** **3 esimest sümbolit** (vanusega)	0
164	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*** (vanusega)	484
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	2
	3	Mõlema koodi 5 esimest sümbolit	0
165	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*****,** ***lopp*** (vanusega)	7
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
166	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*****,** ***pohidgn*** **3 esimest sümbolit** (vanusega)	4
	2	Välispõhjuse koodi 5 esimest sümbolit	0
167	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*****,** ***lopp*****,** ***pohidgn*** **3 esimest sümbolit** (vanusega)	237
	2	Välispõhjuse koodi 5 esimest sümbolit	9
168	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*****,** ***valispohjus*** **3 esimest sümbolit** (vanusega)	3
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
169	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*****,** ***lopp*****,** ***valispohjus*** **3 esimest sümbolit** (vanusega)	63
	1	Põhidiagnoosi koodi 5 esimest sümbolit	18
170	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*****,** ***pohidgn*** **ja** ***valispohjus*** **3 esimest sümbolit** (vanusega)	0
171	0	***vanus*** ± 1**,** ***valtimatu*****,** ***algus*****,** ***lopp*****,** ***pohidgn*** **ja** ***valispohjus*** **3 esimest sümbolit** (vanusega)	9
172	0	***mk*****,** ***pohidgn*** **3 esimest sümbolit** (vanusega)	1
	2	Välispõhjuse koodi 5 esimest sümbolit	0
173	0	***mk*****,** ***pohidgn*** **3 esimest sümbolit;** ***vanus*** **puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
174	0	***mk*****,** ***valispohjus*** **3 esimest sümbolit** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
175	0	***mk*****,** ***valispohjus*** **3 esimest sümbolit;** ***vanus*** **puudub**	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
176	0	***mk*****,** ***pohidgn*** **ja** ***valispohjus*** **3 esimest sümbolit** (vanusega)	0
177	0	***mk*****,** ***pohidgn*** **ja** ***valispohjus*** **3 esimest sümbolit;** ***vanus*** **puudub**	0
178	0	***valtimatu*****,** ***mk*****,** ***pohidgn*** **3 esimest sümbolit** (vanusega)	4
	2	Välispõhjuse koodi 5 esimest sümbolit	1
179	0	***valtimatu*****,** ***mk*****,** ***pohidgn*** **3 esimest sümbolit;** ***vanus*** **puudub**	0
	2	Välispõhjuse koodi 5 esimest sümbolit	1
180	0	***valtimatu*****,** ***mk*****,** ***valispohjus*** **3 esimest sümbolit** (vanusega)	14
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
181	0	***valtimatu*****,** ***mk*****,** ***valispohjus*** **3 esimest sümbolit;** ***vanus*** **puudub**	5
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
182	0	***valtimatu*****,** ***mk*****,** ***pohidgn*** **ja** ***valispohjus*** **3 esimest sümbolit** (vanusega)	1

183	0	***valtimatu*, *mk*, *pohidgn* ja *valispohjus* 3 esimest sümbolit; *vanus* puudub**	1
184	0	***vanus* $\pm$1, *mk*, *algus*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
185	0	***vanus* $\pm$1, *mk*, *algus*, *lopp*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
186	0	***vanus* $\pm$1, *valtimatu*, *mk*, *algus*** (vanusega)	8
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0
187	0	***vanus* $\pm$1, *valtimatu*, *mk*, *algus*, *lopp*** (vanusega)	0
	1	Põhidiagnoosi koodi 5 esimest sümbolit	0
	2	Välispõhjuse koodi 5 esimest sümbolit	0
	3	Mõlema koodi 5 esimest sümbolit	0

Lisa 5. Uute tunnuste loomine imputeerimiseks (rakendustarkvara R).

```
library(readstata13)
dat <- read.dta13("C://KINDLUSTAMATA.dta")

# Loome alguskuu faktortunnuse:
dat$algus_kuu <- format(dat$algus, "%m")
dat$algus_kuu=factor(as.numeric(dat$algus_kuu))

# Loome summaryhma faktortunnuse:
dat$summa_ryhm <- cut(dat$summa,
    breaks = c(-Inf, 100, 200, Inf),
    labels=c("...-100", "101-200", "201-..."))
table(dat$summa_ryhm)

dat$kp_vahe_ryhm <- cut(as.numeric(dat$lopp-dat$algus),
    breaks = c(-Inf, 0, 5, 10, Inf),
    labels=c("<=0", "0-5", "6-10", ">10"))
table(dat$kp_vahe_ryhm)

# Loome p6hidiagnooside koodide ryhmade faktortunnuse:
dat$pohidgn_grupp <- 4
dat$pohidgn_grupp[substr(dat$pohidgn,0,1) == "S" &
    as.numeric(substr(dat$pohidgn,2,3)) < 40] <- 1
dat$pohidgn_grupp[substr(dat$pohidgn,0,1) == "S" &
    as.numeric(substr(dat$pohidgn,2,3)) >= 40 &
    as.numeric(substr(dat$pohidgn,2,3)) < 70] <- 2
dat$pohidgn_grupp[substr(dat$pohidgn,0,1) == "S" &
    as.numeric(substr(dat$pohidgn,2,3)) >= 70] <- 3
dat$pohidgn_grupp=factor(as.numeric(dat$pohidgn_grupp))
table(dat$pohidgn_grupp)


# Loome v2lisp6hjuste koodide ryhmade faktortunnuse:
dat$valispohjus_grupp <- NA
dat$valispohjus_grupp[substr(dat$valispohjus,0,1) == "V"] <- 1
dat$valispohjus_grupp[substr(dat$valispohjus,0,1) == "W"] <- 2
dat$valispohjus_grupp[substr(dat$valispohjus,0,1) == "Y"] <- 4
dat$valispohjus_grupp[substr(dat$valispohjus,0,1) == "X" &
```

```
    as.numeric(substr(dat$valispohjus,2,3)) < 60] <- 3
dat$valispohjus_grupp[substr(dat$valispohjus,0,1) == "X" &
    as.numeric(substr(dat$valispohjus,2,3)) >= 60] <- 4
dat$valispohjus_grupp=
    factor(as.numeric(dat$valispohjus_grupp))
table(dat$valispohjus_grupp)

# Loome perearstivisiite ja haiglavisiite
# eristava faktortunnuse:
dat$perearst <- 0
dat$perearst[substr(dat$ttokood,0,1) == "1"] <- 1
table(dat$perearst)

# Muudame muud tunnused ka faktortunnusteks:
dat$tuup=factor(dat$tuup)
table(dat$tuup)
dat$sugu=factor(dat$sugu)
table(dat$sugu)
```

Lisa 6. Imputeeritavate tunnuste eelnev analüüs (rakendustarkvara R).

Lisa 6.a. Tunnuse *valispohjus_grupp* seos teiste tunnustega.

```
sum(is.na(dat$valispohjus_grupp)) # 9 puuduvat tunnus

#-----------------------------

# valispohjus_grupp ja tuup:
tbl <- table(dat$valispohjus_grupp, dat$tuup)
(tbl <- tbl[complete.cases(tbl),])
#      1    2
# 1  381   89
# 2 6207  425
# 3  594   80
# 4 1228  135
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 119.94, df = 3, p-value < 2.2e-16

# valispohjus_grupp ja kp_vahe_ryhm:
tbl <- table(dat$valispohjus_grupp, dat$kp_vahe_ryhm)
(tbl <- tbl[complete.cases(tbl),])
#    <=0  0-5 6-10  >10
# 1  371   57   25   17
# 2 6029  418  104   81
# 3  576   58   15   25
# 4 1178  136   25   24
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 119.45, df = 9, p-value < 2.2e-16

# valispohjus_grupp ja summa_ryhm:
tbl <- table(dat$valispohjus_grupp, dat$summa_ryhm)
(tbl <- tbl[complete.cases(tbl),])
#  ...-100 101-200 201-...
#1     286     51     133
```

```
#2     5520      503      609
#3      508       64      102
#4      777      272      314
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 570.48, df = 6, p-value < 2.2e-16

# valispohjus_grupp ja algus_kuu:
tbl <- table(dat$valispohjus_grupp, dat$algus_kuu)
(tbl <- tbl[complete.cases(tbl),])
#     1   2   3   4   5   6   7   8   9  10  11  12
# 1  18  18  22  39  58  63  80  64  41  25  21  21
# 2 520 520 505 457 659 665 751 682 487 478 432 476
# 3  68  46  41  56  60  58  95  63  48  52  47  40
# 4 123 100 118 111 107 123 140 124 116 108  89 104
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 98.359, df = 33, p-value = 2.033e-08

# valispohjus_grupp ja perearst:
tbl <- table(dat$valispohjus_grupp, dat$perearst)
(tbl <- tbl[complete.cases(tbl),])
#      0    1
# 1  341  129
# 2 4315 2317
# 3  440  234
# 4  862  501
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 13.556, df = 3, p-value = 0.003576

# valispohjus_grupp ja pohidgn_grupp:
tbl <- table(dat$valispohjus_grupp, dat$pohidgn_grupp)
(tbl <- tbl[complete.cases(tbl),])
#      1    2    3    4
# 1  163  155  118   34
# 2 1609 2232 1836  955
# 3   73   97  112  392
```

```
# 4  747  126   45  445
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 1791.5, df = 9, p-value < 2.2e-16
```

Lisa 6.b. Tunnuse *sugu* seos teiste tunnustega.

```
sum(is.na(dat$sugu)) # 37 puuduvat tunnust

#-------------------------

# sugu ja tuup:
tbl <- table(dat$sugu, dat$tuup)
(tbl <- tbl[complete.cases(tbl),])
#      1    2
# 1 6995  613
# 2 1393  110
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 0.83883, df = 1, p-value = 0.3597

# sugu ja kp_vahe_ryhm:
tbl <- table(dat$sugu, dat$kp_vahe_ryhm)
(tbl <- tbl[complete.cases(tbl),])
#    <=0  0-5 6-10  >10
# 1 6788  545  146  129
# 2 1343  121   22   17
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 5.2603, df = 3, p-value = 0.1537

# sugu ja summa_ryhm:
tbl <- table(dat$sugu, dat$summa_ryhm)
(tbl <- tbl[complete.cases(tbl),])
#   ...-100 101-200 201-...
# 1    5867     763     978
# 2    1212     119     172
chisq.test(tbl)
```

```
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 9.764, df = 2, p-value = 0.007582

# sugu ja algus_kuu:
tbl <- table(dat$sugu, dat$algus_kuu)
(tbl <- tbl[complete.cases(tbl),])
#     1   2   3   4   5   6   7   8   9  10  11  12
# 1 599 572 573 559 756 743 861 775 591 565 487 527
# 2 130 110 110 103 125 163 203 152  98  97  98 114
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 16.562, df = 11, p-value = 0.1215

# sugu ja perearst:
tbl <- table(dat$sugu, dat$perearst)
(tbl <- tbl[complete.cases(tbl),])
#      0    1
# 1 4944 2664
# 2  984  519
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 0.10934, df = 1, p-value = 0.7409

# sugu ja pohidgn_grupp:
tbl <- table(dat$sugu, dat$pohidgn_grupp)
(tbl <- tbl[complete.cases(tbl),])
#      1    2    3    4
# 1 2208 2127 1674 1599
# 2  369  476  429  229
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 59.087, df = 3, p-value = 9.21e-13

# sugu ja valispohjus_grupp:
tbl <- table(dat$sugu, dat$valispohjus_grupp)
(tbl <- tbl[complete.cases(tbl),])
#      1    2    3    4
```

```
# 1   406 5504   552 1139
# 2    61 1101   118  222
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 4.7798, df = 3, p-value = 0.1887
```

Lisa 6.c. Tunnuse *vanus* seos teiste tunnustega.

```
sum(is.na(dat$vanus)) # 3515 puuduvat tunnust

#-------------------------

# T-testid:

# vanus ja tuup:
tbl <- cbind(dat$vanus, dat$tuup)
tbl <- tbl[complete.cases(tbl),]
colnames(tbl) <- c("vanus", "tuup")
tbl <- as.data.frame(tbl)
t.test(tbl$vanus[tbl$tuup==1],
    tbl$vanus[tbl$tuup==2])
# Welch Two Sample t-test
# data:  tbl$vanus[tbl$tuup == 1] and
#   tbl$vanus[tbl$tuup == 2]
# t = -8.5011, df = 828.1, p-value < 2.2e-16
# alternative hypothesis:
#   true difference in means is not equal to 0
# 95 percent confidence interval:
#   -5.617629 -3.510119
# sample estimates:
#   mean of x mean of y
#    36.33657  40.90044

# vanus ja perearst:
tbl <- cbind(dat$vanus, dat$perearst)
tbl <- tbl[complete.cases(tbl),]
colnames(tbl) <- c("vanus", "perearst")
tbl <- as.data.frame(tbl)
t.test(tbl$vanus[tbl$perearst==0],
    tbl$vanus[tbl$perearst==1])
```

```r
# Welch Two Sample t-test
# data:  tbl$vanus[tbl$perearst == 0] and
#   tbl$vanus[tbl$perearst == 1]
# t = 0.17912, df = 4923.1, p-value = 0.8579
# alternative hypothesis:
#   true difference in means is not equal to 0
# 95 percent confidence interval:
#    -0.5696324  0.6841880
# sample estimates:
#   mean of x mean of y
#    36.91256  36.85528

# vanus ja sugu:
tbl <- cbind(dat$vanus, dat$sugu)
tbl <- tbl[complete.cases(tbl),]
colnames(tbl) <- c("vanus", "sugu")
tbl <- as.data.frame(tbl)
t.test(tbl$vanus[tbl$sugu==1],
    tbl$vanus[tbl$sugu==2])
# Welch Two Sample t-test
# data:  tbl$vanus[tbl$sugu == 1] and
    tbl$vanus[tbl$sugu == 2]
# t = -2.9694, df = 1162.6, p-value = 0.003045
# alternative hypothesis:
#   true difference in means is not equal to 0
# 95 percent confidence interval:
#    -2.4412552 -0.4986938
# sample estimates:
#   mean of x mean of y
#    36.64933  38.11931

#-------------------------

# Hii-ruut testid:

# loome selleks vanusegrupid:
summary(dat$vanus)
dat$vanusryhm <- cut(dat$vanus,
    breaks = c(-Inf, 20, 40, 60, Inf),
    labels=c("0-20", "21-40", "41-60", "61-..."))
summary(dat$vanusryhm)
```

```
# vanusryhm ja kp_vahe_ryhm:
tbl <- table(dat$vanusryhm, dat$kp_vahe_ryhm)
(tbl <- tbl[complete.cases(tbl),])
#           <=0  0-5 6-10  >10
#   0-20    228   17    5    3
#  21-40   2978  278   50   38
#  41-60   1561  190   86   71
# 61-...    93   16    9   10
chisq.test(tbl, simulate.p.value = TRUE)
# Pearson's Chi-squared test with simulated p-value
    (based on 2000 replicates)
# data:  tbl
# X-squared = 126.25, df = NA, p-value = 0.0004998

# vanusryhm ja summa_ryhm:
tbl <- table(dat$vanusryhm, dat$summa_ryhm)
(tbl <- tbl[complete.cases(tbl),])
#         ...-100 101-200 201-...
#   0-20     199      23      31
#  21-40    2592     279     473
#  41-60    1278     207     423
# 61-...      67      17      44
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 108.38, df = 6, p-value < 2.2e-16

# vanusryhm ja algus_kuu:
tbl <- table(dat$vanusryhm, dat$algus_kuu)
(tbl <- tbl[complete.cases(tbl),])
#           1   2   3   4   5   6   7   8   9  10  11
12
#   0-20   20  15  27  16  25  20  48  21  13  24  13
11
#  21-40  275 247 273 234 332 338 380 351 240 249 216 209
#  41-60  165 135 128 134 164 174 206 190 157 162 138 155
# 61-...   14  19  10   6  15   6  14   8   7  17   7
5
chisq.test(tbl)
# Pearson's Chi-squared test
```

```
# data:  tbl
# X-squared = 67.922, df = 33, p-value = 0.0003267

# vanusryhm ja pohidgn_grupp:
tbl <- table(dat$vanusryhm, dat$pohidgn_grupp)
(tbl <- tbl[complete.cases(tbl),])
#           1    2    3    4
#   0-20   63   90   64   36
#  21-40  985  904  742  713
#  41-60  557  450  435  466
# 61-...   43   39   21   25
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 33.93, df = 9, p-value = 9.19e-05

# vanus ja valispohjus_grupp:
tbl <- table(dat$vanusryhm, dat$valispohjus_grupp)
(tbl <- tbl[complete.cases(tbl),])
#           1    2    3    4
#  0-20    26  176   24   27
#  21-40  176 2361  211  592
#  41-60  102 1366  165  275
# 61-...    4   99   11   14
chisq.test(tbl)
# Pearson's Chi-squared test
# data:  tbl
# X-squared = 54.69, df = 9, p-value = 1.395e-08

dat$vanusryhm <- NULL
```

Lisa 7. Imputeerimine (rakendustarkvara R).

```
library("VIM")
library(dplyr)

#-----------------------------
#-----------------------------
# YLDINE HOT–DECK MEETOD:
#-----------------------------
# valispohjus_grupp imputeerimine:
sum(is.na(dat$valispohjus_grupp)) # 9 puuduvat tunnust
hot_yld <- hotdeck(dat, variable=c("valispohjus_grupp"))
sum(is.na(hot_yld$valispohjus_grupp)) # 0 puuduvat tunnust
#-----------------------------
# sugu imputeerimine:
sum(is.na(hot_yld$sugu))  # 37 puuduvat tunnust
hot_yld <- hotdeck(hot_yld, variable=c("sugu"))
sum(is.na(hot_yld$sugu)) # 0 puuduvat tunnust
#-----------------------------
# vanus imputeerimine:
sum(is.na(hot_yld$vanus))  # 3515 puuduvat tunnust
hot_yld <- hotdeck(hot_yld, variable=c("vanus"))
sum(is.na(hot_yld$vanus)) # 0 puuduvat tunnust
#-----------------------------
#-----------------------------
# L2HIMA NAABRI MEETOD MEETOD:
#-----------------------------
# valispohjus_grupp imputeerimine:
sum(is.na(dat$valispohjus_grupp)) # 9 puuduvat tunnust
knn<- kNN(dat, variable=c("valispohjus_grupp"),
          dist_var = c("summa_ryhm", "algus_kuu",
                       "perearst", "kp_vahe_ryhm",
                       "pohidgn_grupp", "tuup"), k=1)
sum(is.na(knn$valispohjus_grupp)) # 0 puuduvat tunnust
#-----------------------------
# sugu imputeerimine:
sum(is.na(knn$sugu))  # 37 puuduvat tunnust
knn <- kNN(knn, variable=c("sugu"),
           dist_var = c("summa_ryhm",
                        "pohidgn_grupp"), k=1)
sum(is.na(knn$sugu)) # 0 puuduvat tunnust
```

```
#-----------------------------
# vanus imputeerimine:
sum(is.na(knn$vanus)) # 3515 puuduvat tunnust
knn <- kNN(knn, variable=c("vanus"),
           dist_var = c("summa_ryhm", "algus_kuu",
                        "sugu","kp_vahe_ryhm",
                        "pohidgn_grupp", "tuup",
                        "valispohjus_grupp"), k=1)
sum(is.na(knn$vanus)) # 0 puuduvat tunnust
#-----------------------------
#-----------------------------
# JUHUSLIK HOT-DECK OMISTUS KLASSIS
# JA L2HIMA NAABRI IMPUTEERIMINE:
#-----------------------------
# valispohjus_grupp imputeerimine
sum(is.na(dat$valispohjus_grupp)) # 9 puuduvat tunnust

# Uurime kas esineb gruppe, kus ainult puuduvad v22rtused:
data <- dat
data$puuduv <- 0
data$puuduv[is.na(data$valispohjus_grupp)] <- 1
data <- transform(data, group = as.numeric(factor(
  paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
        algus_kuu, tuup, perearst, sep=""))))
max(data$group) # 692 gruppi moodustus
# M22rame grupeeriva tunnuse:
data <- group_by(data, group)
data <- mutate(data, puuduvaid = sum(puuduv),
               yldse = length(puuduv))
data$tyhjad <- 0
data$tyhjad[data$puuduvaid == data$yldse] <- 1
sum(data$tyhjad) # tyhje gruppe ei esine

# seega ainult Hot-Deck omistus klassis:
hot_knn <- hotdeck(data, variable=c("valispohjus_grupp"),
                   domain_var = c("summa_ryhm", "algus_kuu",
                                  "perearst","kp_vahe_ryhm",
                                  "pohidgn_grupp", "tuup"))
sum(is.na(hot_knn$valispohjus_grupp)) # 0 puuduvat tunnust
hot_knn$tyhjad <- NULL
hot_knn$yldse <- NULL
```

```
hot_knn$puuduvaid <- NULL
hot_knn$group <- NULL
hot_knn$puuduv <- NULL
#-------------------------
# sugu imputeerimine:
sum(is.na(hot_knn$sugu))  # 37 puuduvat tunnust

# Uurime kas esineb gruppe, kus ainult puuduvad v22rtused:
data <- hot_knn
data$puuduv <- 0
data$puuduv[is.na(data$sugu)] <- 1
data <- transform(data, group = as.numeric(factor(
  paste(summa_ryhm, pohidgn_grupp, sep=""))))
max(data$group) # 12 gruppi moodustus
# M22rame grupeeriva tunnuse:
data <- group_by(data, group)
data <- mutate(data, puuduvaid = sum(puuduv),
               yldse = length(puuduv))
data$tyhjad <- 0
data$tyhjad[data$puuduvaid == data$yldse] <- 1
sum(data$tyhjad) # tyhje gruppe ei esine

# seega ainult Hot-Deck omistus klassis:
hot_knn <- hotdeck(data, variable=c("sugu"),
                   domain_var = c("summa_ryhm",
                                  "pohidgn_grupp"))
sum(is.na(hot_knn$sugu))  # 0 puuduvat tunnust
hot_knn$tyhjad <- NULL
hot_knn$yldse <- NULL
hot_knn$puuduvaid <- NULL
hot_knn$group <- NULL
hot_knn$puuduv <- NULL
#-------------------------
# vanus imputeerimine
sum(is.na(hot_knn$vanus)) # 3515 puuduvat tunnust

# Uurime kas esineb gruppe, kus ainult puuduvad v22rtused:
data <- hot_knn
data$puuduv <- 0
data$puuduv[is.na(data$vanus)] <- 1
data <- transform(data, group = as.numeric(factor(
```

```
   paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
         algus_kuu, tuup, sugu, valispohjus_grupp,
         sep="")))))
max(data$group) # 1226 gruppi moodustus
# M22rame grupeeriva tunnuse:
data <- group_by(data, group)
data <- mutate(data, puuduvaid = sum(puuduv),
                yldse = length(puuduv))
data$tyhjad <- 0
data$tyhjad[data$puuduvaid == data$yldse] <- 1
sum(data$tyhjad) # esineb 236 tyhja gruppi kuuluvat vaatlust

hot_knn <- hotdeck(data, variable=c("vanus"),
                   domain_var = c("summa_ryhm", "algus_kuu",
                                  "sugu","kp_vahe_ryhm",
                                  "pohidgn_grupp", "tuup",
                                  "valispohjus_grupp"))
sum(is.na(hot_knn$vanus)) # 0 puuduvat tunnust
# vaatame, mis v22rtus omistati tyhjas grupis:
hot_knn$vanus[data$tyhjad==1] # omistatakse alati 1
hot_knn$vanus_imputeeritud <- hot_knn$vanus_imp
hot_knn$vanus_imp <- NULL

# tyhjades gruppides imputeerime l2hima naabri meetodiga:
hot_knn$vanus[data$tyhjad==1] <- NA
sum(is.na(hot_knn$vanus))
hot_knn<- kNN(hot_knn, variable=c("vanus"),
              dist_var = c("summa_ryhm", "algus_kuu",
                           "sugu","kp_vahe_ryhm",
                           "pohidgn_grupp", "tuup",
                           "valispohjus_grupp"), k=1)
sum(is.na(knn$vanus)) # 0 puuduvat tunnust
# vaatame, mis v22rtus omistati tyhjas grupis:
hot_knn$vanus[data$tyhjad==1] # omistati erinevaid vanuseid
hot_knn$vanus_imp[hot_knn$vanus_imputeeritud==TRUE]<-TRUE
hot_knn$tyhjad <- NULL
hot_knn$yldse <- NULL
hot_knn$puuduvaid <- NULL
hot_knn$group <- NULL
hot_knn$puuduv <- NULL
hot_knn$vanus_imputeeritud <- NULL
```

Lisa 8. Tulemuste analüüs (rakendustarkvara R).

```
# Loome meetodeid eristavad tunnused:
hot_knn$meetod <- "hot-deck&knn"
hot_yld$meetod <- "hot-deck"
knn$meetod <- "knn"
dat$meetod <- "algne"
dat$vanus_imp <- NA
dat$sugu_imp <- NA
dat$valispohjus_grupp_imp <- NA
#-----------------------
# VALISPOHJUS_GRUPP
#-----------------------
# Yhendame v6rdlemiseks andmestikud
v6rdluseks <- rbind(hot_yld, knn, hot_knn,
    subset(dat, !is.na(valispohjus_grupp)))
tbl <- table(v6rdluseks$meetod,
    v6rdluseks$valispohjus_grupp)
names(dimnames(tbl)) <- c("Meetod", "Valispohjuse grupp")
summake <- apply(tbl,1,sum)
addmargins(round(sweep(tbl,1,summake,"/"),4)*100)
v6rdluseks <- rbind(
    subset(hot_yld, valispohjus_grupp_imp==TRUE),
    subset(knn, valispohjus_grupp_imp==TRUE),
    subset(hot_knn, valispohjus_grupp_imp==TRUE))
tbl <- table(v6rdluseks$meetod,
    v6rdluseks$valispohjus_grupp)
names(dimnames(tbl)) <- c("Meetod", "Valispohjuse grupp")
tbl
#-----------------------
# SUGU
#-----------------------
# Yhendame v6rdlemiseks andmestikud
v6rdluseks <- rbind(hot_yld, knn, hot_knn,
    subset(dat, !is.na(sugu)))
# kontrollime osakaalusid:
round(table(v6rdluseks$sugu[v6rdluseks$meetod==
    "hot-deck"])/
  (0.01*length(v6rdluseks$sugu[v6rdluseks$meetod==
    "hot-deck"])), 2)
round(table(v6rdluseks$sugu[v6rdluseks$meetod=="knn"])/
```

```
  (0.01*length(v6rdluseks$sugu[v6rdluseks$meetod==
    "knn"])), 2)
round(table(v6rdluseks$sugu[v6rdluseks$meetod==
    "hot-deck&knn"])/
  (0.01*length(v6rdluseks$sugu[v6rdluseks$meetod==
    "hot-deck&knn"])), 2)
round(table(v6rdluseks$sugu[v6rdluseks$meetod=="algne"])/
  (0.01*length(v6rdluseks$sugu[v6rdluseks$meetod==
    "algne"])), 2)
v6rdluseks <- rbind(
    subset(hot_yld, sugu_imp==TRUE),
    subset(knn, sugu_imp==TRUE),
    subset(hot_knn, sugu_imp==TRUE))
tbl <- table(v6rdluseks$meetod, v6rdluseks$sugu)
names(dimnames(tbl)) <- c("Meetod", "Sugu")
tbl
#-------------------------
# VANUS
#-------------------------
# Yhendame v6rdlemiseks andmestikud
v6rdluseks <- rbind(hot_yld, knn, hot_knn,
    subset(dat, !is.na(vanus)))

round(mean(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])

round(mean(v6rdluseks$vanus[v6rdluseks$meetod=="knn"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod=="knn"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])

round(mean(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck&knn"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck&knn"]), 3)
```

```
min(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])

round(mean(v6rdluseks$vanus[v6rdluseks$meetod=="algne"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod=="algne"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="algne"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="algne"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="algne"])

boxplot(vanus~meetod,data=v6rdluseks,
    main="Meetodite vordlus vanuse imputeerimisel",
    xlab="Meetod", ylab="Vanus")

v6rdluseks <- rbind(subset(hot_yld, vanus_imp==TRUE),
    subset(knn, vanus_imp==TRUE),
    subset(hot_knn, vanus_imp==TRUE))
boxplot(vanus~meetod,data=v6rdluseks,
    main="Meetodite vordlus vanuse imputeerimisel
-ainult imputeeritud tulemused",
    xlab="Meetod", ylab="Vanus")
```

Lisa 9. 70% olemasolevate andmete täiustamine (rakendustarkvara R).

Lisa 9.a. Puuduvate andmete tekitamine.

```
# Votame katses kasutusele ainult taielikud andmed:
data_proov <- subset(dat, !is.na(valispohjus_grupp))
data_proov <- subset(data_proov, !is.na(sugu))
data_proov <- subset(data_proov, !is.na(vanus))
data_proov$vanus_imp <- NULL
data_proov$sugu_imp <- NULL
data_proov$valispohjus_grupp_imp <- NULL
data_proov$meetod <- NULL

data_proov$ID <- seq.int(nrow(data_proov))
# Kui olemas on v2hemalt 70% andmeid:
kustutatavaid_kolmkymmend <- floor(nrow(data_proov)*0.3)
kustutatavad_read_kolmkymmend <- sample(data_proov$ID,
    kustutatavaid_kolmkymmend)
data_kolmkymmend <- data_proov
data_kolmkymmend$vanus[is.element(data_kolmkymmend$ID,
    kustutatavad_read_kolmkymmend)] <- NA
sum(!is.na(data_kolmkymmend$vanus)) # on olemas 3941 v22rtust
```

Lisa 9.b. Puuduvate andmete imputeerimine.

```
# vanus imputeerimine:
sum(is.na(data_kolmkymmend$vanus)) # on puudu 1688 v22rtust
#-------------------------
#-------------------------
# YLDINE HOT-DECK MEETOD:
hot_kolmkymmend <- hotdeck(data_kolmkymmend,
    variable=c("vanus"))
sum(is.na(hot_kolmkymmend$vanus)) # 0 puuduvat tunnust
#-------------------------
#-------------------------
# L2HIMA NAABRI MEETOD:
knn_kolmkymmend <- kNN(data_kolmkymmend,
    variable=c("vanus"),
    dist_var = c("summa_ryhm", "algus_kuu",
```

```
        "sugu","kp_vahe_ryhm",
        "pohidgn_grupp", "tuup",
        "valispohjus_grupp"), k=1)
sum(is.na(knn_kolmkymmend$vanus)) # 0 puuduvat tunnust
#-------------------------
#-------------------------
# JUHUSLIK HOT-DECK OMISTUS KLASSIS
# JA L2HIMA NAABRI IMPUTEERIMINE:
# Uurime kas esineb gruppe, kus ainult puuduvad v22rtused:
data <- data_kolmkymmend
data$puuduv <- 0
data$puuduv[is.na(data$vanus)] <- 1
data <- transform(data, group = as.numeric(factor(
    paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
        algus_kuu, tuup, sugu, valispohjus_grupp, sep=""))))
max(data$group) # 1041 gruppi moodustus
# M22rame grupeeriva tunnuse:
data <- group_by(data, group)
data <- mutate(data, puuduvaid = sum(puuduv),
    yldse = length(puuduv))
data$tyhjad <- 0
data$tyhjad[data$puuduvaid == data$yldse] <- 1

hot_knn_kolmkymmend <- hotdeck(data, variable=c("vanus"),
    domain_var = c("summa_ryhm", "algus_kuu", "sugu",
        "kp_vahe_ryhm", "pohidgn_grupp", "tuup",
        "valispohjus_grupp"))
sum(is.na(hot_knn_kolmkymmend$vanus)) # 0 puuduvat tunnust
# vaatame, mis v22rtus omistati tyhjas grupis:
hot_knn_kolmkymmend$vanus[data$tyhjad==1] # omistatakse alati 1
hot_knn_kolmkymmend$vanus_imputeeritud <-
    hot_knn_kolmkymmend$vanus_imp
hot_knn_kolmkymmend$vanus_imp <- NULL

# tyhjades gruppides imputeerime l2hima naabri meetodiga:
hot_knn_kolmkymmend$vanus[data$tyhjad==1] <- NA
sum(is.na(hot_knn_kolmkymmend$vanus))
hot_knn_kolmkymmend<- kNN(hot_knn_kolmkymmend,
    variable=c("vanus"),
    dist_var = c("summa_ryhm", "algus_kuu", "sugu",
        "kp_vahe_ryhm", "pohidgn_grupp", "tuup",
```

```
        "valispohjus_grupp"), k=1)
sum(is.na(hot_knn_kolmkymmend$vanus)) # 0 puuduvat tunnust
# vaatame, mis v22rtus omistati tyhjas grupis:
hot_knn_kolmkymmend$vanus[data$tyhjad==1]
# omistati erinevaid v22rtuseid
hot_knn_kolmkymmend$vanus_imp[
    hot_knn_kolmkymmend$vanus_imputeeritud==TRUE]<-TRUE
hot_knn_kolmkymmend$tyhjad <- NULL
hot_knn_kolmkymmend$yldse <- NULL
hot_knn_kolmkymmend$puuduvaid <- NULL
hot_knn_kolmkymmend$group <- NULL
hot_knn_kolmkymmend$puuduv <- NULL
hot_knn_kolmkymmend$vanus_imputeeritud <- NULL
```

Lisa 9.c. Imputeeritud andmete analüüs.

```
# Loome meetodeid eristavad tunnused:
hot_knn_kolmkymmend$meetod <- "hot-deck&knn"
hot_kolmkymmend$meetod <- "hot-deck"
knn_kolmkymmend$meetod <- "knn"
data_pr <- data_proov
data_pr$meetod <- "tegelik"
data_pr$vanus_imp <- NA

# Yhendame v6rdlemiseks andmestikud
v6rdluseks <- rbind(hot_kolmkymmend, knn_kolmkymmend,
    hot_knn_kolmkymmend, data_pr)
round(mean(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
round(mean(v6rdluseks$vanus[v6rdluseks$meetod=="knn"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod=="knn"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
round(mean(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck&knn"]), 3)
```

```
round(sd(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck&knn"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
round(mean(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"])

boxplot(vanus~meetod,data=v6rdluseks,
        main="Meetodite␣vordlus␣vanuse␣imputeerimisel",
        xlab="Meetod", ylab="Vanus")

# Uurime vanusevahesid:
data_proovike <- subset(data_pr,
    is.element(data_kolmkymmend$ID,
    kustutatavad_read_kolmkymmend))
hot_kolm_v6rdlus <- merge(subset(hot_kolmkymmend,
    vanus_imp==TRUE), data_proovike,by="ID")
hot_kolm_v6rdlus$vanusevahe <-
    hot_kolm_v6rdlus$vanus.x-hot_kolm_v6rdlus$vanus.y
round(mean(hot_kolm_v6rdlus$vanusevahe), 3)
round(sd(hot_kolm_v6rdlus$vanusevahe), 3)
round(min(hot_kolm_v6rdlus$vanusevahe), 3)
round(max(hot_kolm_v6rdlus$vanusevahe), 3)
round(median(hot_kolm_v6rdlus$vanusevahe), 3)
knn_kolm_v6rdlus <- merge(subset(knn_kolmkymmend,
    vanus_imp==TRUE), data_proovike,by="ID")
knn_kolm_v6rdlus$vanusevahe <-
    knn_kolm_v6rdlus$vanus.x-knn_kolm_v6rdlus$vanus.y
round(mean(knn_kolm_v6rdlus$vanusevahe), 3)
round(sd(knn_kolm_v6rdlus$vanusevahe), 3)
round(min(knn_kolm_v6rdlus$vanusevahe), 3)
round(max(knn_kolm_v6rdlus$vanusevahe), 3)
round(median(knn_kolm_v6rdlus$vanusevahe), 3)
hot_knn_kolm_v6rdlus <- merge(subset(hot_knn_kolmkymmend,
    vanus_imp==TRUE), data_proovike, by="ID")
hot_knn_kolm_v6rdlus$vanusevahe <-
  hot_knn_kolm_v6rdlus$vanus.x-hot_knn_kolm_v6rdlus$vanus.y
```

```
round(mean(hot_knn_kolm_v6rdlus$vanusevahe), 3)
round(sd(hot_knn_kolm_v6rdlus$vanusevahe), 3)
round(min(hot_knn_kolm_v6rdlus$vanusevahe), 3)
round(max(hot_knn_kolm_v6rdlus$vanusevahe), 3)
round(median(hot_knn_kolm_v6rdlus$vanusevahe), 3)

v6rdluseks <- rbind(hot_kolm_v6rdlus, knn_kolm_v6rdlus,
    hot_knn_kolm_v6rdlus)
boxplot(vanusevahe~meetod.x,data=v6rdluseks,
    main="Meetodite vordlus - vanusevahed",
    xlab="Meetod", ylab="Vanus")
```

Lisa 9.d. Imputeerimise simulatsioon ja selle analüüs (ühtedele ja samadele andmetele imputeerimine).

```
hot_vec_keskmised <- c(mean(hot_kolmkymmend$vanus))
knn_vec_keskmised <- c(mean(knn_kolmkymmend$vanus))
hot_knn_vec_keskmised <- c(mean(hot_knn_kolmkymmend$vanus))

# Viime l2bi imputeerimist 99 korda veel:
for (i in 1:99){
    # YLDINE HOT-DECK MEETOD:
    hot_kolmkymmend_sim <- hotdeck(data_kolmkymmend,
        variable=c("vanus"))
    hot_vec_keskmised <- append(hot_vec_keskmised,
        mean(hot_kolmkymmend_sim$vanus))
    #-------------------------
    # L2HIMA NAABRI MEETOD:
    knn_kolmkymmend_sim <- kNN(data_kolmkymmend,
        variable=c("vanus"),
        dist_var = c("summa_ryhm", "algus_kuu",
            "sugu", "kp_vahe_ryhm", "pohidgn_grupp",
            "tuup", "valispohjus_grupp"), k=1)
    knn_vec_keskmised <- append(knn_vec_keskmised,
        mean(knn_kolmkymmend_sim$vanus))
    #-------------------------
    # JUHUSLIK HOT-DECK OMISTUS KLASSIS
    # JA L2HIMA NAABRI IMPUTEERIMINE:
    data <- data_kolmkymmend
    data$puuduv <- 0
    data$puuduv[is.na(data$vanus)] <- 1
```

```
    data <- transform(data, group = as.numeric(factor(
        paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
            algus_kuu, tuup, sugu, valispohjus_grupp,
            sep=""))))
    # M22rame grupeeriva tunnuse:
    data <- group_by(data, group)
    data <- mutate(data, puuduvaid = sum(puuduv),
        yldse = length(puuduv))
    data$tyhjad <- 0
    data$tyhjad[data$puuduvaid == data$yldse] <- 1

    hot_knn_kolmkymmend_sim <- hotdeck(data,
        variable=c("vanus"),
        domain_var = c("summa_ryhm",
            "algus_kuu", "sugu", "kp_vahe_ryhm",
            "pohidgn_grupp", "tuup", "valispohjus_grupp"))
    hot_knn_kolmkymmend_sim$vanus_imputeeritud <-
    hot_knn_kolmkymmend_sim$vanus_imp
    hot_knn_kolmkymmend_sim$vanus_imp <- NULL

    # tyhjades gruppides imputeerime l2hima naabri meetodiga:
    hot_knn_kolmkymmend_sim$vanus[data$tyhjad==1] <- NA
    hot_knn_kolmkymmend_sim<- kNN(hot_knn_kolmkymmend_sim,
        variable=c("vanus"),
        dist_var = c("summa_ryhm",
            "algus_kuu", "sugu", "kp_vahe_ryhm",
            "pohidgn_grupp", "tuup",
            "valispohjus_grupp"), k=1)
    hot_knn_kolmkymmend_sim$vanus_imp[
        hot_knn_kolmkymmend_sim$vanus_imputeeritud==TRUE]<-TRUE
    hot_knn_kolmkymmend_sim$tyhjad <- NULL
    hot_knn_kolmkymmend_sim$yldse <- NULL
    hot_knn_kolmkymmend_sim$puuduvaid <- NULL
    hot_knn_kolmkymmend_sim$group <- NULL
    hot_knn_kolmkymmend_sim$puuduv <- NULL
    hot_knn_kolmkymmend_sim$vanus_imputeeritud <- NULL
    hot_knn_vec_keskmised <- append(hot_knn_vec_keskmised,
        mean(hot_knn_kolmkymmend_sim$vanus))
}

vahepealne_30 <- data.frame(hot_vec_keskmised,
```

```
    knn_vec_keskmised, hot_knn_vec_keskmised)
keskmised_hot <- data.frame(vahepealne_30$hot_vec_keskmised)
colnames(keskmised_hot) <- "keskmine"
keskmised_hot$meetod <- "hot"
keskmised_knn <- data.frame(vahepealne_30$knn_vec_keskmised)
colnames(keskmised_knn) <- "keskmine"
keskmised_knn$meetod <- "knn"
keskmised_hot_knn <- data.frame(
    vahepealne_30$hot_knn_vec_keskmised)
colnames(keskmised_hot_knn) <- "keskmine"
keskmised_hot_knn$meetod <- "hot&knn"
v6rdluseks <- rbind(keskmised_hot, keskmised_knn,
    keskmised_hot_knn)
boxplot(keskmine~meetod,data=v6rdluseks,
    main="Meetodite vordlus simuleerimisel (70%)",
    xlab="Meetod", ylab="Keskmine")
abline(h = mean(data_proov$vanus), col = "red")
```

Lisa 9.e. Imputeerimise simulatsioon ja selle analüüs (erinevatele andmetele imputeerimine).

```
hot_vec_erinev_keskmised <- c(mean(hot_kolmkymmend$vanus))
knn_vec_erinev_keskmised <- c(mean(knn_kolmkymmend$vanus))
hot_knn_vec_erinev_keskmised <- c(mean(hot_knn_kolmkymmend$vanus))

# Viime l2bi imputeerimist 99 korda veel:
for (i in 1:99){
    # Viime l2bi kustutamise igal sammul:
    kustutatavaid_kolmkymmend <- floor(nrow(data_proov)*0.3)
    kustutatavad_read_kolmkymmend <- sample(data_proov$ID,
        kustutatavaid_kolmkymmend)
    data_kolmkymmend <- data_proov
    data_kolmkymmend$vanus[is.element(data_kolmkymmend$ID,
        kustutatavad_read_kolmkymmend)] <- NA
    sum(!is.na(data_kolmkymmend$vanus))
    #-------------------------
    # YLDINE HOT-DECK MEETOD:
    hot_kolmkymmend_sim <- hotdeck(data_kolmkymmend,
        variable=c("vanus"))
    hot_vec_erinev_keskmised <- append(hot_vec_erinev_keskmised,
        mean(hot_kolmkymmend_sim$vanus))
```

```
#-----------------------------
# L2HIMA NAABRI MEETOD:
knn_kolmkymmend_sim <- kNN(data_kolmkymmend,
     variable=c("vanus"),
     dist_var = c("summa_ryhm",
          "algus_kuu", "sugu","kp_vahe_ryhm",
          "pohidgn_grupp", "tuup",
          "valispohjus_grupp"), k=1)
knn_vec_erinev_keskmised <- append(knn_vec_erinev_keskmised,
     mean(knn_kolmkymmend_sim$vanus))
#-----------------------------
# JUHUSLIK HOT-DECK OMISTUS KLASSIS
# JA L2HIMA NAABRI IMPUTEERIMINE:
data <- data_kolmkymmend
data$puuduv <- 0
data$puuduv[is.na(data$vanus)] <- 1
data <- transform(data, group = as.numeric(factor(
     paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
          algus_kuu, tuup, sugu, valispohjus_grupp,
          sep=""))))
# M22rame grupeeriva tunnuse:
data <- group_by(data, group)
data <- mutate(data, puuduvaid = sum(puuduv),
     yldse = length(puuduv))
data$tyhjad <- 0
data$tyhjad[data$puuduvaid == data$yldse] <- 1

hot_knn_kolmkymmend_sim <- hotdeck(data,
     variable=c("vanus"),
     domain_var = c("summa_ryhm",
          "algus_kuu", "sugu", "kp_vahe_ryhm",
          "pohidgn_grupp", "tuup", "valispohjus_grupp"))
hot_knn_kolmkymmend_sim$vanus_imputeeritud <-
     hot_knn_kolmkymmend_sim$vanus_imp
hot_knn_kolmkymmend_sim$vanus_imp <- NULL

# tyhjades gruppides imputeerime l2hima naabri meetodiga:
hot_knn_kolmkymmend_sim$vanus[data$tyhjad==1] <- NA
hot_knn_kolmkymmend_sim<- kNN(hot_knn_kolmkymmend_sim,
     variable=c("vanus"),
     dist_var = c("summa_ryhm", "algus_kuu", "sugu",
```

```
            "kp_vahe_ryhm", "pohidgn_grupp", "tuup",
            "valispohjus_grupp"), k=1)
    hot_knn_kolmkymmend_sim$vanus_imp[
        hot_knn_kolmkymmend_sim$vanus_imputeeritud==TRUE]<-TRUE
    hot_knn_kolmkymmend_sim$tyhjad <- NULL
    hot_knn_kolmkymmend_sim$yldse <- NULL
    hot_knn_kolmkymmend_sim$puuduvaid <- NULL
    hot_knn_kolmkymmend_sim$group <- NULL
    hot_knn_kolmkymmend_sim$puuduv <- NULL
    hot_knn_kolmkymmend_sim$vanus_imputeeritud <- NULL
    hot_knn_vec_erinev_keskmised <-
        append(hot_knn_vec_erinev_keskmised,
        mean(hot_knn_kolmkymmend_sim$vanus))
}

vahepealne_30_erinev <- data.frame(hot_vec_erinev_keskmised,
    knn_vec_erinev_keskmised, hot_knn_vec_erinev_keskmised)

keskmised_hot <- data.frame(
    vahepealne_30_erinev$hot_vec_erinev_keskmised)
colnames(keskmised_hot) <- "keskmine"
keskmised_hot$meetod <- "hot"
keskmised_knn <- data.frame(
    vahepealne_30_erinev$knn_vec_erinev_keskmised)
colnames(keskmised_knn) <- "keskmine"
keskmised_knn$meetod <- "knn"
keskmised_hot_knn <- data.frame(
    vahepealne_30_erinev$hot_knn_vec_erinev_keskmised)
colnames(keskmised_hot_knn) <- "keskmine"
keskmised_hot_knn$meetod <- "hot&knn"
v6rdluseks <- rbind(keskmised_hot, keskmised_knn,
    keskmised_hot_knn)
boxplot(keskmine~meetod,data=v6rdluseks,
    main="Meetodite␣vordlus␣simuleerimisel␣(70%)
␣␣␣␣-␣erinevad␣andmed",
    xlab="Meetod", ylab="Keskmine")
abline(h = mean(data_proov$vanus), col = "red")
```

Lisa 10. 50% olemasolevate andmete täiustamine (rakendustarkvara R).

Lisa 10.a. Puuduvate andmete tekitamine.

```
# Votame proovi-imputeerimise alla ainult taielikud andmed:
data_proov <- subset(dat, !is.na(valispohjus_grupp))
data_proov <- subset(data_proov, !is.na(sugu))
data_proov <- subset(data_proov, !is.na(vanus))
data_proov$vanus_imp <- NULL
data_proov$sugu_imp <- NULL
data_proov$valispohjus_grupp_imp <- NULL
data_proov$meetod <- NULL

data_proov$ID <- seq.int(nrow(data_proov))
# Kui olemas on v2hemalt 50% andmeid:
kustutatavaid_viiskymmend <- floor(nrow(data_proov)*0.5)
kustutatavad_read_viiskymmend <- sample(data_proov$ID,
    kustutatavaid_viiskymmend)
data_viiskymmend <- data_proov
data_viiskymmend$vanus[is.element(data_viiskymmend$ID,
    kustutatavad_read_viiskymmend)]<-NA
sum(!is.na(data_viiskymmend$vanus)) # on olemas 2815 v22rtust
```

Lisa 10.b. Puuduvate andmete imputeerimine.

```
# vanus imputeerimine:
sum(is.na(data_viiskymmend$vanus)) # on puudu 2814 v22rtust
#-------------------------
#-------------------------
# YLDINE HOT-DECK MEETOD:
hot_viiskymmend <- hotdeck(data_viiskymmend,
                           variable=c("vanus"))
sum(is.na(hot_viiskymmend$vanus)) # 0 puuduvat tunnust
#-------------------------
#-------------------------
# L2HIMA NAABRI MEETOD:
knn_viiskymmend <- kNN(data_viiskymmend,
    variable=c("vanus"),
    dist_var = c("summa_ryhm", "algus_kuu",
```

```
        "sugu","kp_vahe_ryhm",
        "pohidgn_grupp", "tuup",
        "valispohjus_grupp"), k=1)
sum(is.na(knn_viiskymmend$vanus)) # 0 puuduvat tunnust
#-------------------------
#-------------------------
# JUHUSLIK HOT-DECK OMISTUS KLASSIS
# JA L2HIMA NAABRI IMPUTEERIMINE:
# Uurime kas esineb gruppe, kus ainult puuduvad v22rtused:
data <- data_viiskymmend
data$puuduv <- 0
data$puuduv[is.na(data$vanus)] <- 1
data <- transform(data, group = as.numeric(factor(
    paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
        algus_kuu, tuup, sugu, valispohjus_grupp, sep=""))))
max(data$group) # 1041 gruppi moodustus
# M22rame grupeeriva tunnuse:
data <- group_by(data, group)
data <- mutate(data, puuduvaid = sum(puuduv),
    yldse = length(puuduv))
data$tyhjad <- 0
data$tyhjad[data$puuduvaid == data$yldse] <- 1

hot_knn_viiskymmend <- hotdeck(data,
    variable=c("vanus"),
    domain_var = c("summa_ryhm", "algus_kuu", "sugu",
    "kp_vahe_ryhm", "pohidgn_grupp", "tuup",
    "valispohjus_grupp"))
sum(is.na(hot_knn$vanus)) # 0 puuduvat tunnust
# vaatame, mis v22rtus omistati tyhjas grupis:
hot_knn_viiskymmend$vanus[data$tyhjad==1] # omistatakse alati 1
hot_knn_viiskymmend$vanus_imputeeritud <-
    hot_knn_viiskymmend$vanus_imp
hot_knn_viiskymmend$vanus_imp <- NULL

# tyhjades gruppides imputeerime l2hima naabri meetodiga:
hot_knn_viiskymmend$vanus[data$tyhjad==1] <- NA
sum(is.na(hot_knn_viiskymmend$vanus))
hot_knn_viiskymmend<- kNN(hot_knn_viiskymmend,
    variable=c("vanus"),
    dist_var = c("summa_ryhm", "algus_kuu","sugu",
```

```
        "kp_vahe_ryhm", "pohidgn_grupp", "tuup",
        "valispohjus_grupp"), k=1)
sum(is.na(hot_knn_viiskymmend$vanus)) # 0 puuduvat tunnust
# vaatame, mis v22rtus omistati tyhjas grupis:
hot_knn_viiskymmend$vanus[data$tyhjad==1]
# omistati erinevaid v22rtuseid
hot_knn_viiskymmend$vanus_imp[
    hot_knn_viiskymmend$vanus_imputeeritud==TRUE]<-TRUE
hot_knn_viiskymmend$tyhjad <- NULL
hot_knn_viiskymmend$yldse <- NULL
hot_knn_viiskymmend$puuduvaid <- NULL
hot_knn_viiskymmend$group <- NULL
hot_knn_viiskymmend$puuduv <- NULL
hot_knn_viiskymmend$vanus_imputeeritud <- NULL
```

Lisa 10.c. Imputeeritud andmete analüüs.

```
# Loome meetodeid eristavad tunnused:
hot_knn_viiskymmend$meetod <- "hot-deck&knn"
hot_viiskymmend$meetod <- "hot-deck"
knn_viiskymmend$meetod <- "knn"
data_pr <- data_proov
data_pr$meetod <- "tegelik"
data_pr$vanus_imp <- NA

# Yhendame v6rdlemiseks andmestikud
v6rdluseks <- rbind(hot_viiskymmend, knn_viiskymmend,
    hot_knn_viiskymmend, data_pr)
round(mean(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck"])
round(mean(v6rdluseks$vanus[v6rdluseks$meetod=="knn"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod=="knn"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="knn"])
round(mean(v6rdluseks$vanus[v6rdluseks$meetod==
```

```
    "hot-deck&knn"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod==
    "hot-deck&knn"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="hot-deck&knn"])
round(mean(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"]), 3)
round(sd(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"]), 3)
min(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"])
max(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"])
median(v6rdluseks$vanus[v6rdluseks$meetod=="tegelik"])

boxplot(vanus~meetod,data=v6rdluseks,
    main="Meetodite␣vordlus␣vanuse␣imputeerimisel",
    xlab="Meetod", ylab="Vanus")

# Uurime vanusevahesid:
data_proovike <- subset(data_pr,
    is.element(data_viiskymmend$ID,
    kustutatavad_read_viiskymmend))
hot_viis_v6rdlus <- merge(subset(hot_viiskymmend,
    vanus_imp==TRUE), data_proovike,by="ID")
hot_viis_v6rdlus$vanusevahe <-
    hot_viis_v6rdlus$vanus.x-hot_viis_v6rdlus$vanus.y
round(mean(hot_viis_v6rdlus$vanusevahe), 3)
round(sd(hot_viis_v6rdlus$vanusevahe), 3)
round(min(hot_viis_v6rdlus$vanusevahe), 3)
round(max(hot_viis_v6rdlus$vanusevahe), 3)
round(median(hot_viis_v6rdlus$vanusevahe), 3)
knn_viis_v6rdlus <- merge(subset(knn_viiskymmend,
    vanus_imp==TRUE), data_proovike,by="ID")
knn_viis_v6rdlus$vanusevahe <-
    knn_viis_v6rdlus$vanus.x-knn_viis_v6rdlus$vanus.y
round(mean(knn_viis_v6rdlus$vanusevahe), 3)
round(sd(knn_viis_v6rdlus$vanusevahe), 3)
round(min(knn_viis_v6rdlus$vanusevahe), 3)
round(max(knn_viis_v6rdlus$vanusevahe), 3)
round(median(knn_viis_v6rdlus$vanusevahe), 3)
hot_knn_viis_v6rdlus <- merge(subset(hot_knn_viiskymmend,
    vanus_imp==TRUE), data_proovike,by="ID")
hot_knn_viis_v6rdlus$vanusevahe <-
```

```
    hot_knn_viis_v6rdlus$vanus.x-hot_knn_viis_v6rdlus$vanus.y
round(mean(hot_knn_viis_v6rdlus$vanusevahe), 3)
round(sd(hot_knn_viis_v6rdlus$vanusevahe), 3)
round(min(hot_knn_viis_v6rdlus$vanusevahe), 3)
round(max(hot_knn_viis_v6rdlus$vanusevahe), 3)
round(median(hot_knn_viis_v6rdlus$vanusevahe), 3)

v6rdluseks <- rbind(hot_viis_v6rdlus, knn_viis_v6rdlus,
    hot_knn_viis_v6rdlus)
boxplot(vanusevahe~meetod.x,data=v6rdluseks,
    main="Meetodite vordlus - vanusevahed",
    xlab="Meetod", ylab="Vanus")
```

Lisa 10.d. Imputeerimise simulatsioon ja selle analüüs (ühtedele ja samadele andmetele imputeerimine).

```
hot_vec_keskmised <- c(mean(hot_viiskymmend$vanus))
knn_vec_keskmised <- c(mean(knn_viiskymmend$vanus))
hot_knn_vec_keskmised <- c(mean(hot_knn_viiskymmend$vanus))

# Viime l2bi imputeerimist 99 korda veel:
for (i in 1:99){
    # YLDINE HOT-DECK MEETOD:
    hot_viiskymmend_sim <- hotdeck(data_viiskymmend,
        variable=c("vanus"))
    hot_vec_keskmised <- append(hot_vec_keskmised,
        mean(hot_viiskymmend_sim$vanus))
    #-------------------------
    # L2HIMA NAABRI MEETOD:
    knn_viiskymmend_sim <- kNN(data_viiskymmend,
        variable=c("vanus"),
        dist_var = c("summa_ryhm", "algus_kuu",
            "sugu","kp_vahe_ryhm", "pohidgn_grupp",
            "tuup", "valispohjus_grupp"), k=1)
    knn_vec_keskmised <- append(knn_vec_keskmised,
        mean(knn_viiskymmend_sim$vanus))
    #-------------------------
    # JUHUSLIK HOT-DECK OMISTUS KLASSIS
    # JA L2HIMA NAABRI IMPUTEERIMINE:
    data <- data_viiskymmend
    data$puuduv <- 0
```

```
    data$puuduv[is.na(data$vanus)] <- 1
    data <- transform(data, group = as.numeric(factor(
        paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
            algus_kuu, tuup, sugu, valispohjus_grupp,
            sep=""))))
    # M22rame grupeeriva tunnuse:
    data <- group_by(data, group)
    data <- mutate(data, puuduvaid = sum(puuduv),
        yldse = length(puuduv))
    data$tyhjad <- 0
    data$tyhjad[data$puuduvaid == data$yldse] <- 1

    hot_knn_viiskymmend_sim <- hotdeck(data,
        variable=c("vanus"),
        domain_var = c("summa_ryhm",
            "algus_kuu", "sugu", "kp_vahe_ryhm",
            "pohidgn_grupp", "tuup",
            "valispohjus_grupp"))
    hot_knn_viiskymmend_sim$vanus_imputeeritud <-
        hot_knn_viiskymmend_sim$vanus_imp
    hot_knn_viiskymmend_sim$vanus_imp <- NULL

    # tyhjades gruppides imputeerime l2hima naabri meetodiga:
    hot_knn_viiskymmend_sim$vanus[data$tyhjad==1] <- NA
    hot_knn_viiskymmend_sim<- kNN(hot_knn_viiskymmend_sim,
        variable=c("vanus"),
        dist_var = c("summa_ryhm",
            "algus_kuu", "sugu", "kp_vahe_ryhm",
            "pohidgn_grupp", "tuup",
            "valispohjus_grupp"), k=1)
    hot_knn_viiskymmend_sim$vanus_imp[
        hot_knn_viiskymmend_sim$vanus_imputeeritud==TRUE]<-TRUE
    hot_knn_viiskymmend_sim$tyhjad <- NULL
    hot_knn_viiskymmend_sim$yldse <- NULL
    hot_knn_viiskymmend_sim$puuduvaid <- NULL
    hot_knn_viiskymmend_sim$group <- NULL
    hot_knn_viiskymmend_sim$puuduv <- NULL
    hot_knn_viiskymmend_sim$vanus_imputeeritud <- NULL
    hot_knn_vec_keskmised <- append(hot_knn_vec_keskmised,
        mean(hot_knn_viiskymmend_sim$vanus))
}
```

```
vahepealne_50 <- data.frame(hot_vec_keskmised,
    knn_vec_keskmised, hot_knn_vec_keskmised)

keskmised_hot <- data.frame(vahepealne_50$hot_vec_keskmised)
colnames(keskmised_hot) <- "keskmine"
keskmised_hot$meetod <- "hot"
keskmised_knn <- data.frame(vahepealne_50$knn_vec_keskmised)
colnames(keskmised_knn) <- "keskmine"
keskmised_knn$meetod <- "knn"
keskmised_hot_knn <- data.frame(
    vahepealne_50$hot_knn_vec_keskmised)
colnames(keskmised_hot_knn) <- "keskmine"
keskmised_hot_knn$meetod <- "hot&knn"
v6rdluseks <- rbind(keskmised_hot, keskmised_knn,
    keskmised_hot_knn)
boxplot(keskmine~meetod,data=v6rdluseks,
    main="Meetodite␣vordlus␣simuleerimisel␣(50%)",
    xlab="Meetod", ylab="Keskmine")
abline(h = mean(data_proov$vanus), col = "red")
```

Lisa 10.e. Imputeerimise simulatsioon ja selle analüüs (erinevatele andmetele imputeerimine).

```
hot_vec_erinev_keskmised <- c(mean(hot_viiskymmend$vanus))
knn_vec_erinev_keskmised <- c(mean(knn_viiskymmend$vanus))
hot_knn_vec_erinev_keskmised <- c(mean(hot_knn_viiskymmend$vanus))

# Viime l2bi imputeerimist 99 korda veel:
for (i in 1:99){
    # Viime l2bi kustutamise igal sammul:
    kustutatavaid_viiskymmend <- floor(nrow(data_proov)*0.5)
    kustutatavad_read_viiskymmend <- sample(data_proov$ID,
        kustutatavaid_viiskymmend)
    data_viiskymmend <- data_proov
    data_viiskymmend$vanus[is.element(data_viiskymmend$ID,
        kustutatavad_read_viiskymmend)]<-NA
    #-----------------------
    # YLDINE HOT-DECK MEETOD:
    hot_viiskymmend_sim <- hotdeck(data_viiskymmend,
        variable=c("vanus"))
```

```
hot_vec_erinev_keskmised <- append(hot_vec_erinev_keskmised,
    mean(hot_viiskymmend_sim$vanus))
#-------------------------
# L2HIMA NAABRI MEETOD:
knn_viiskymmend_sim <- kNN(data_viiskymmend,
    variable=c("vanus"),
    dist_var = c("summa_ryhm", "algus_kuu",
        "sugu","kp_vahe_ryhm", "pohidgn_grupp", "tuup",
        "valispohjus_grupp"), k=1)
knn_vec_erinev_keskmised <- append(knn_vec_erinev_keskmised,
    mean(knn_viiskymmend_sim$vanus))
#-------------------------
# JUHUSLIK HOT-DECK OMISTUS KLASSIS
# JA L2HIMA NAABRI IMPUTEERIMINE:
data <- data_viiskymmend
data$puuduv <- 0
data$puuduv[is.na(data$vanus)] <- 1
data <- transform(data, group = as.numeric(factor(
    paste(summa_ryhm, kp_vahe_ryhm, pohidgn_grupp,
        algus_kuu, tuup, sugu, valispohjus_grupp,
        sep=""))))
# M22rame grupeeriva tunnuse:
data <- group_by(data, group)
data <- mutate(data, puuduvaid = sum(puuduv),
    yldse = length(puuduv))
data$tyhjad <- 0
data$tyhjad[data$puuduvaid == data$yldse] <- 1

hot_knn_viiskymmend_sim <- hotdeck(data,
    variable=c("vanus"),
    domain_var = c("summa_ryhm",
        "algus_kuu", "sugu", "kp_vahe_ryhm",
        "pohidgn_grupp", "tuup", "valispohjus_grupp"))
hot_knn_viiskymmend_sim$vanus_imputeeritud <-
    hot_knn_viiskymmend_sim$vanus_imp
hot_knn_viiskymmend_sim$vanus_imp <- NULL

# tyhjades gruppides imputeerime l2hima naabri meetodiga:
hot_knn_viiskymmend_sim$vanus[data$tyhjad==1] <- NA
hot_knn_viiskymmend_sim<- kNN(hot_knn_viiskymmend_sim,
    variable=c("vanus"),
```

```
        dist_var = c("summa_ryhm", "algus_kuu","sugu",
            "kp_vahe_ryhm", "pohidgn_grupp", "tuup",
            "valispohjus_grupp"), k=1)
    hot_knn_viiskymmend_sim$vanus_imp[
        hot_knn_viiskymmend_sim$vanus_imputeeritud==TRUE]<-TRUE
    hot_knn_viiskymmend_sim$tyhjad <- NULL
    hot_knn_viiskymmend_sim$yldse <- NULL
    hot_knn_viiskymmend_sim$puuduvaid <- NULL
    hot_knn_viiskymmend_sim$group <- NULL
    hot_knn_viiskymmend_sim$puuduv <- NULL
    hot_knn_viiskymmend_sim$vanus_imputeeritud <- NULL
    hot_knn_vec_erinev_keskmised <-
    append(hot_knn_vec_erinev_keskmised,
        mean(hot_knn_viiskymmend_sim$vanus))
}

vahepealne_50_erinev <- data.frame(hot_vec_erinev_keskmised,
    knn_vec_erinev_keskmised, hot_knn_vec_erinev_keskmised)

keskmised_hot <- data.frame(
    vahepealne_50_erinev$hot_vec_erinev_keskmised)
colnames(keskmised_hot) <- "keskmine"
keskmised_hot$meetod <- "hot"
keskmised_knn <- data.frame(
    vahepealne_50_erinev$knn_vec_erinev_keskmised)
colnames(keskmised_knn) <- "keskmine"
keskmised_knn$meetod <- "knn"
keskmised_hot_knn <- data.frame(
    vahepealne_50_erinev$hot_knn_vec_erinev_keskmised)
colnames(keskmised_hot_knn) <- "keskmine"
keskmised_hot_knn$meetod <- "hot&knn"
v6rdluseks <- rbind(keskmised_hot, keskmised_knn,
    keskmised_hot_knn)
boxplot(keskmine~meetod,data=v6rdluseks,
    main="Meetodite␣vordlus␣simuleerimisel␣(50%)
␣␣␣␣-␣erinevad␣andmed",
    xlab="Meetod", ylab="Keskmine")
abline(h = mean(data_proov$vanus), col = "red")
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Viktoria Kirpu,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

“Haigekassa kindlustamata patsientide vigastuste andmete imputeerimine”,

mille juhendajad on Natalja Lepik ja Natalja Eigo,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni; 1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace’i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **08.05.2018**