TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Kristina Mumm

Terviseandmete tabeli veergude automaatne tüübituvastus

Bakalaureusetöö (9 EAP)

Juhendaja: Sulev Reisberg

Tartu 2022

Terviseandmete tabeli veergude automaatne tüübituvastus

Lühikokkuvõte:

Käesoleva töö eesmärgiks on luua automatne veergude tüübi tuvastamise komponent projekti Health Sense raames loodavale terviseandmete anonüümimise tarkvarale. Täpsemalt keskendub komponent terviseandmetele, mis sisaldavad sageli kategoorilisi andmeid. Tüübi tuvastamise protsess jagati kolmeks osas. Regulaaravaldiste abil leitakse, milliste andmetüüpide struktuuriga veeru väärtused sobivad. Seejärel kontrollitakse veeru väärtuste vastavust leitud andmetüüpide lubatud väärtuste loendiga. Viimases sammus valitakse sobivatest andmetüüpidest kõige sobilikum. Lisaks analüüsitakse töös Eesti terviseandmeid, et aru saada, kas terviseandmetes esinevad vead võivad olla probleemiks tüübi tuvastamise juures.

Võtmesõnad:

Terviseandmed, tüübituvastus, projekt Health Sense

CERCS:

B110 Bioinformatics, medical informatics, biomathematics, biometrics
P160 Statistics, operation research, programming, actuarial mathematics
P170 Computer science, numerical analysis, systems, control

Automatic Type Detection of Columns in Table of Health Records

Abstract:

The Bachelor’s thesis proposes an automatic method for detecting the type of columns in a data table. Specifically, it focuses on health data where data columns often represent categorical data. The type detection process is divided into three parts. Regular expressions are used to find out which data type structures are appropriate for column values. Then, the values in the column are checked against the list of allowed values for these data types found. In the last step, the best fit of the suitable data types is found. Additionally, the thesis analyses Estonian health data to understand whether errors in health data might cause problems for type detection.

Keywords:

Health records, type detection, project Health Sense

CERCS:

B110 Bioinformatics, medical informatics, biomathematics, biometrics
P160 Statistics, operation research, programming, actuarial mathematics
P170 Computer science, numerical analysis, systems, control

Sisukord

Sissejuhatus

Terviseandmete maht on praeguseks kasvanud väga suureks ja andmeid tuleb igapeävaselt aina rohkem juurde. Digitaalsel kujul andmed pakuvad teadlastele võimalust teha uuringuid, mis võiksid elanikkonna tervist pikemas perspektiivis parandada. Näiteks on võimalik terviseandmete põhjal luua 2. tüüpi diabeeti ennustamismudel [1] või teha analüüse nakkushaiguste puhangute õigeaegseteks avastamiseks ning ennetemiseks [2].

Inimuringute jaoks kõige kasulikumad andmed on Julia Lane [3] sõnul mikroandmed, mis kujutavad endast ükiskisikute, leibkondade või asutuste tunnuseid [4]. Sellisteks tunnusteks võivad olla näiteks inimese haiguslood ja tarvitatavad retseptiravimid. Vaatamata sellele, et uuringute jaoks vajalike andmete maht on kõvasti kasvanud, on nende kättesaadavus konfidentsiaalsuse reeglite tõttu piiratud. Isegi kui andmetest likvideerida isiklik informatsioon nagu täisnimi, aadress ja telefoninumber, siis on võimalik andmeid ja inimest ikka kokku viia, kui tal on näiteks ainukesena teada mõni harudane haigus.

Konfidentsiaalsete andmete kätte saamiseks ja mistahes inimuuringu, kus uuritakse inimesi läbi viimiseks on vaja eetika komitee luba. Eetika komitee hindab, et vastav uuring oleks eetiline ega kahjustaks uuritavaid. Sealhulgas hindab ka seda, kas uurimiseks kasutatavad andmed ei riiva liigselt uuritavate privaatsust. Privaatsuse riive tagajärjel võivad konfidentsiaalsed andmed teatavaks saada kõrvalistele isikutele, kes võivad neid omakasu eesmärgil ära kasutada [3]. Uuritava jaoks, kelle andmed lekkisid, võib see tähendada näiteks tööpakkumiste arvu vähenemist või kindlustuse negatiivset otsust, sest kõrvalistel isikutel on rohkem informatsiooni, mille alusel saavad otsuseid langetada [3]. Eetikakomitee loa taotlemine on üsna ajakulukas ja keeruline protsess ning selle lihtsustamiseks Joseph Ficek [5] sõnul võiks enne andmete väljastamist rakendada andmetele meetodeid, mis vähendaksid võimalust viia kokku andmeid ja uuritavaid. Üheks meetodiks, mille ta välja tõi, on andmete anonüümimine. Projekti Health Sense, mille partneriks on TEHIK, raames luuakse andmete teisendamise keskkonda, kus kasutatakse andmete anonüümimise meetodit[1].

Täpsemalt uuritakse projektis Health Sense tabelikujul olevaid terviseandmeid ja nende anonüümimisvõimalusi. Konkreetsemalt käsitletakse üldistamismeetodil anonüümimist, kus andmetabeli igas väljas võidakse väärtus jätta samaks või üldistada suunas konkreetsest-üldisele. Selleks, et seda automaatselt teha, on tarvis esimese sammuna tuvastada etteantud

[1] Valik TEHIK-u töödest ja teenustest on leitavad järgneval veebileheküljel: https://www.tehik.ee/projektid.

andmetabeli veergude tüübid, sest arvulisi, tekstilisi ja kuupäevalisi väärtusi tuleb käsitleda erinevalt. Näiteks arvude anonüümise puhul on võimalik konkreetne arv muuta vahemikuks, kuid seda sama ei saa teha arstipoolse kommentaariga. Tekstiliste, kuid kindlast loendist pärit väärtuste puhul (näiteks diagnoos, ravimi toimeaine), mida antud töös nimetatakse kategoorilisteks andmetüüpideks, tuleb õigeks anonüümimiseks esmalt tuvastada konkreetne loend, millest väärtused pärinevad.

Autori teada ei ole varasemalt loodud tüübituvastajaid, mis võimaldaksid määrata andmeveeru tüüpi tulenevalt etteantud lubatud loendite hulgast. Käesolevas töös pakutakse välja metoodika tabeli kujul terviseandmete veergude tüübi tuvastamiseks ning luuakse selle implementatsioon projekti Health Sense.

Käesoleva töö esimeses peatükis antakse ülevaade terviseandmetest ning Eestis levinumatest kategoorilistest andmetüüpidest. Samuti tutvustatakse Health Sense projekti ja tuuakse välja tüübituvastaja olulisus ning olemus. Teises peatükis kirjeldatakse terviseandmete analüüsimise ning tüübituvastaja metoodikat . Kolmandas peatükis kirjeldatakse terviseandmete analüüsi tulemusi, tüübituvastuse implementatsiooni ning selle testimise tulemusi.

1. Taust

Sellest peatükis antakse ülevaade terviseadnmetest ning Eestis levinumatest kategoorilistest andmetüüpidest.

1.1 Terviseandmed

Terviseandmete maht Eestis on praeguseks kasvanud väga suureks, sest neid kogutakse inimese kohta kogu tema elu jooksul alates sündimise hetkest [6]. Andmeid võib säilitada näiteks paberil, digitaalsetes toimikutes või andmebaasides [7]. Terviseandmete hoidmine elektroonsel kujul aitab paremini hallata olemasolevaid andmeid ning võimaldab pakkuda paremat tervishoiu teenust, sest andmed on pidevalt ajakohased, jagatavad erinevate organisatsioonide vahel ning kiiresti kättesaadavad ka patsiendile endale. Ei ole vaja tegeleda aeganõudva paberimajandusega, sest andmeid sisestatakse arvutisse. Võimalus andmeid jagada tähendab, et neid on kergem ka koguda ning hiljem kasutada uuringute jaoks.

Andmed elektroonsel kujul võivad sisaldada mitmekülgset infot [8]:

- patsiendi isiklik info identifitseerimiseks, mis sisaldab endas täisnime, sünniaega, isikukoodi, kontaktandmeid, hädaolukorra kontaktandmeid, tööandja kontaktandmeid jms;
- demograafiline info, mis sisaldab endas näiteks sugu, vanust, rassi jms;
- diagnoosid;
- ravimid ning nende toimeained;
- protseduurid;
- uuringud;
- elulised näitajad nagu kehatemperatuur, pulsisagedus, hingamissagedus ja vererõhk;
- patsiendi enda poolt genereeritud andmed nagu näiteks magamismustrid, füüsiline aktiivsus, kodus mõõdetud veresuhkur.

Terviseandmetes peamiselt esinevaid väärtusi jaotatakse neljaks erinevaks tüübiks: kategoorilised, arvulised nagu kehakaal, kuupäevalised ja vabatekstilised väärtused [9]. Kategooriliste andmete all antud töös mõistetakse andmeid, mille väärtused on pärit kindlast loetelust, neid ei saa otseselt väljendada arvudena ega järjestada mingi loogika alusel Järgnevalt alampeatükkidena kirjeldatakse tüüpilisi Eesti terviseandmetes esinevaid kategoorilisi ja kuupäevalisi väärtusi.

1.1.1 Eesti haldus- ja asustusjaotuse klassifikaator

Eesti haldus- ja asustusjaotuse klassifikaator (EHAK) on klassifikaator territoraalise paiknemise tähistamiseks Eestis ning koosneb kahest osast: identifitseerivast ja klassifitseerivast. Igale klassifitseerimisobjektile on antud unikaalne neljakohaline identifitseeriv kood. Need koodid on ära jaotatud järgmiselt:

- 0030-0089 – maakonnad;
- 0100-0999 – vallad, linnad (haldusüksused) ja linnaosad;
- 1000-9999 – linnad (asustusüksused), alevid, alevikud ja külad

Näiteks tähistab EHAK kood „4471“ Loksa küla. Kuigi koodid on neljakohalised, siis andmestikes võivad need esineda ka vasakpoolsete nullideta.

Klassifitseerimisobjekte on võimalik esitada ka pikemal kujul koos klassifitseeriva osaga, mis koosneb iga objekti kolmest tunnusest:

1. maakonna-kuuluvuse tunnus;
2. vallakuuluvuse tunnus;
3. objekti tüübi tunnus.

Loksa küla pikk kood oleks seega „003703534471M8“, kus kood „0037“ tähistab Harju maakonda, „0353“ Kuusalu valda, „4471“ Loksa küla, „M“ maalist asustustuspiirkonda ning „8“ küla.

Kokkuvõttes on EHAK-ul kolm esitamisviisi: lühike kood ilma nullideta („37“), lühike kood nullidega („0037“) ja pikk kood („003703534471M8“). Hetkel kehtivaid koode koos täpsemate selgitusega on võimalik leida Statistikaameti klassifikaatorite portaalist, mis asub veebilehekülje https://klassifikaatorid.stat.ee/.

1.1.2 Diagnoos

Rahvusvaheline haiguste ja nendega seotud terviseprobleemide statistiline klassifikatsioon (RHK) on süsteem, mis võimaldab eri riikides ja eri aegadel haiguste kohta kogutud andmeid süstemaatiliselt töödelda. Hetkel on Eestis kohustuslik kasutada viimast ehk 10. versiooni. [10]

RHK-10 kood koosneb 3 kuni 6 sümbolist. Iga kood algab tähega, mis viitab enamasti suuremale peatükile nagu näiteks „närvisüsteemi haigused“ või „haigestumise ja surma välispõhjused“. Ülejäänud koodiosa viitab alampeatükkidele ja jaotistele, mis on iga peatüki

puhul erinevalt ära nummerdatud. Järgmiselt on välja toodud RHK-10 võimalikud esinemiskujud koos näidetega alates kõige üldisemast:

- Peatükk – F00-F99 – psüühika- ja käitumishäired;
- Alampeatükk – F30-F39 – meeleoluhäired;
- Jaotis – F33 – korduv depressiivne häire ehk korduv depressioon;
- Alamjaotis – F33.3 – raske korduv depressioon psühhootiliste sümptomitega;
- 5.koha alajaotis – F33.30 – meeleoluga ühtuvad psühhootilised sümptomid. [10]

Hetkel Eestis kehtivaid diagnoosikoode on võimalik leida Sotsiaalministeeriumi RHK-10 pühendatud veebileheküljel https://rhk.sm.ee/.

1.1.3 Ravimi toimeainete klassifikatsioon

Raamatus „Farmaatsiaterminoloogia“ [11] on selgitatud ATC-koodi ehk anatoomilist-terapeutilist-keemilist koodi kui inimestel kasutatavate raviainete klassifiaktsioonisüsteemi. Toimeained on jaotatud erinevatesse rühmadesse vastavalt sellele, millisele elundile või elundsüsteemile need toimivad ning nende toime ja keemiliste omadustele. Iga rühm koosneb viiest tasandist:

- B – veri ja vereloomeorganid (1. tasand – anatoomiline põhirühm)
- B01 – tromboosivastased ained (2. tasand – terapeutiline alamrühm)
- B01A – tromboosivastased ained (3.tasand – toimeline alamrühm)
- B01AD – ensüümid (4.tasand – keemiline alamrühm)
- B01AD12 – proteiin C (5.tasand - toimeaine)

Lisaks on olemas ka veterinaar-ATC süsteem, mis sarnaneb eeltoodud ATC-koodi süsteemile, kuid struktuuri poolest erineb ainult ühe tähe poolest: Q-täht lisatakse koodi ette [11].

Hetkel Eestis kehtivaid ATC-koode on võimalik leida Ravimiregistri koduleheküljelt.

1.1.4 Sugu

Soo andmetüüp valdavas enamuses koosneb väärtustest „mees“ ja „naine“. Neid saab erinevat moodi märkida, näiteks „M“ ja „N“, „M“ ja „F“, „poiss“ ja „tüdruk“. Käesoleva töö raames piirdutakse vaid väärtustega „M“ ja „N“.

1.1.5 Kuupäev

Maailmas kasutatakse mitmeid erinevaid formaate kuupäevade esitamiseks. Eesti infosüsteemides kasutatavate kuupäeva formaatide kohta ei ole teadaolevalt statistikat avaldatud, seega käesolevas töös kasutatakse vaid arvulisi kuupäeva formaate kujul „YYYY-MM-DD“, „MM-DD-YYYY“ ja „DD-MM-YYYY“ kus „YYYY“ on neljakohaline aastaarv, „MM“ kahekohaline kuu, „DD“ kahekohaline päev ja „-“ on üks järgmistest kuupäeva osade eraldajatest: punkt, mõttekriips, kaldkriips või siis pole eraldajat.

1.2 Health Sense projekt

Health Sense on projekt, mille peamiseks kasusaajaks on TEHIK ning mille eesmärgiks on parandada terviseandmete kättesaadavust[2]. Tooreid terviseandmeid, kus on patsientide kohta konfidentsiaalset infot, ei saa Isikuandmete kaitse seaduse §16 järgi väljastada teadusuuringute eesmärkidel isikutele, kes ei ole seotud tervishoiu valdkonnaga ega oma vastavat luba [12]. TEHIK-u sõnul hetkel võtab spetsialistidel liiga kaua aega, et andmed oleksid töödeldud anonüümsemaks. Neil oleks vaja tööriista, mis automatiseeriks ja kiirendaks protsessi. Kuigi eelmainitud seaduse järgi on anonüümitud andmete kasutamiseks sellegipoolest vaja taotleda eetika komitee eriluba, mis tõendab, et uuritavate konfidentsuaalsus on tagatud, peaks protsess olema mõnevõrra lihtsam, kuna tegemist ei ole enam toorete terviseandmetega [12].

Projekt koosneb mitmest osast ja Tartu Ülikooli vastutada on andmete anonüümija arendus. Anonüümija tarkvara eesmärk on konkreetseid andmeid üldistada hierarhiliselt üles poole. Esialgu projektis anonüümitakse vaid kategoorilisi ja kuupäevalisi väärtusi ning lisaks ainukesena numbrilise väärtusena ka vanust. Anonüümimise jaoks on tarvis teada, millist andmetüüpi veerud terviseandmetes on, et igale veerule anda ette õige hierarhia. Seda soovitakse teha automaatselt, sest terviseandmed võivad sisaldada kümneid ridu. Nii kergendatakse tarkvara kasutaja tööd.

Hierarhiaks käesolevas töös loetakse andmefaili, kus veergude väärtused muutuvad suunas konkreetsest-üldisele. Hierarhiad on koostatud iga andmetüübi jaoks eraldi: Eestis kehtivatest EHAK-u, diagnoosi- ja ATC-koodidest, võimalikest kuupäevadest ning võimalikest inimese

[2] Projektist Health Sense saab lähemalt lugeda lingilt https://www.tehik.ee/projektid.

vanustest vahemikus 0-130[3]. Hierarhiate esimene veerg koosneb kõige konkreetsematest väärtustest (Tabel 1). Järgnevad veerud koosnevad väärtustest, mis on üldisemad võrreldes eelneva veeruga.

Tabel 1. Näide osa diagnoosikoodide hierarhiast, kus iga järgmine veerg koosneb üldisemast väärtusest võrreldes eelneva veeruga.

F33.10	F33.1	F33	F30-F39	F00-F99	*
F33.11	F33.1	F33	F30-F39	F00-F99	*
F33.30	F33.3	F33	F30-F39	F00-F99	*
F33.31	F33.3	F33	F30-F39	F00-F99	*
F38.00	F38.0	F38	F30-F39	F00-F99	*
F38.10	F38.1	F38	F30-F39	F00-F99	*
F40.00	F40.0	F40	F40-F49	F00-F99	*

Anonüümimise tarkvara oskab hägustada vaid väärtusi, mis asuvad hierarhiate esimeses veerus. Andmeid analüüsides tarkvara vaid otsustab, millise tasemeni on vaja konkreetset väärtust üldistada. See tähendab, et kui Tabelis 1 puuduks esimeses veerus väärtus „V12.59“, siis anonüümija ei oskaks selle väärtusega midagi teha vaatamata sellele, et väärtus võib esineda ka teistes veergudes. Seega edaspidi töös kutsutakse esimese veeru väärtusi lubatud väärtusteks ning tervet veergu lubatud väärtuste loendiks.

Genereeritud näidisandmetest (Tabel 2) on näha, et andmed on mitme isiku kohta, kes elavad EHAK koodi „130“ järgi Alutaguse vallas. On antud ka patsientide sünnipäevad ning diagnoosid. Alutaguse vallas elab seisuga 01.01.2022 vaid 4658 elanikku [13], mis tähendab, et tegelikult sünnipäevade järgi on võimalik üsna lihtsasti selgeks teha, kes on antud näidisandmete järgi millise diagnoosi saanud, sest ei ole tõenäoline, et elanike seas on palju tüdrukuid, kes sündisid kuupäeval 08.07.2019. Anonüümija eesmärk on sellise probleemi kõrvaldamine.

[3] Antud töös kasutatavad hierarhiad on leitavad repositoorimust https://gitlab.cs.ut.ee/pqder/loputoo-failid-tuubituvastaja.

Tabel 2. Osa andmete anonüümija jaoks genereeritud näidisandmetest.

patient_id	patient_gender	patient_birthdate	patient_ehak_code	dgn_code	case_start
402630355	N	08.07.2019	130	I84.5	23.04.2020
251934949	M	21.10.2017	130	M79.6	23.04.2020
3530572859	M	05.04.1969	130	Z02.4	22.04.2020
1708457701	N	26.02.1973	130	Z09.4	23.04.2020
1606034887	N	15.10.1955	130	Z71.1	23.04.2020
3500436139	M	08.02.1973	141	A63.0	23.04.2020

Konkreetsete andmete üldistamine toimib hierarhiliselt ülespoole. Kui eelnevalt on antud kuupäev 08.07.2019, siis seda on võimalik üldistada kuu või aasta täpsuseks: 07.2019 või 2019. Samuti on võimalik üldistada ka teisi veerge näiteks EHAK-u koodiga. Kood „130“ tähistab Alutaguse valda, mis asub Ida-Viru maakonnas. Antud maakonna kood on „45“. Kui näidisandmetel Alutaguse vald üldistada hierarhiliselt ülespoole maakonna tasemeni, siis antud andmete puhul ei saa andmete saaja enam öelda, et patsiendid elavad kindlasti Alutaguse vallas.

Konkreetse anonüümise metoodika valik ja kirjeldus ei kuulu käesoleva töö skoopi. Oluline on üksnes see, et anonüümimise teostamiseks on tarvis määrata igale sisendtabeli veerule õige tüüp ja hierarhia. Käesolev töö keskendub küsimusele, kuidas seda määramist teha automaatselt, et valmiv tarkvara oleks kasutajasõbralikum, sest terviseandmed võivad sisaldada kümneid veerge ja iga kord nende määramine on tülikas.

1.3 Tüübi tuvastamine

Projekti Health Sense andmete anonüümimise tarkvara juures on oluline roll andmetüübi tuvastamisel. Eelnevalt kirjeldatud RHK-10 ja ATC koodide puhul on näha, et nende esinemisviis on erinev, mis tähendab, et ka üldistamise hierarhiad peavad olema erinevad.

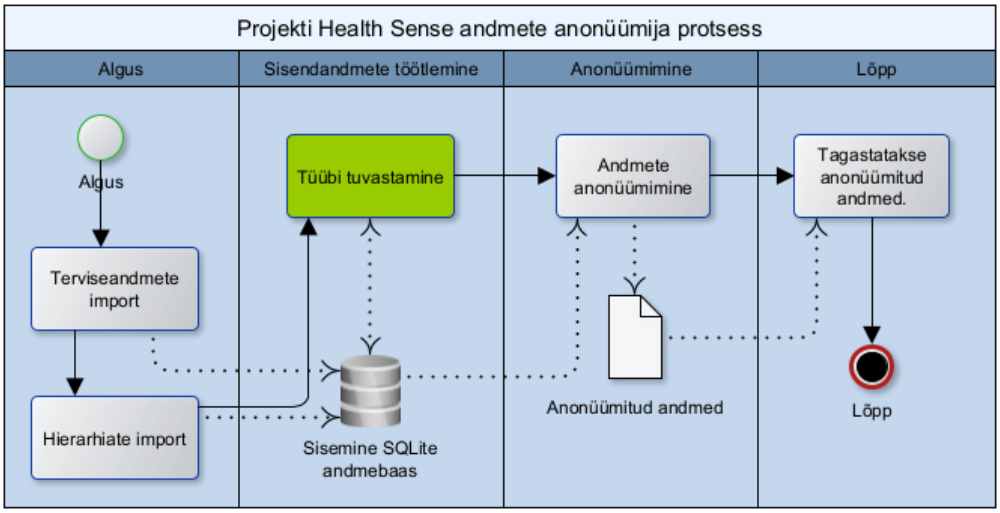
Tüüpi ei saa tuvastada vaid veeru ühe suvalise väärtuse põhjal, sest andmetes võib esineda vigu ning üks väärtus võib olla struktuuri poolest lubatud mitmes andmetüübis. Samuti ei ole mõistlik tuvastamist teostada veeru väärtuste põhjal, sest terviseandmed võivad sisaldada miljoneid ridu ja kümneid veerge, mille töötlemine võtab omajagu aega. Piisaks, kui valida minimaalne kogus suvalisi ridu analüüsimiseks, mille juures oleks veeru tüüp juba tuvastatav. Antud töös tehakse lisaks veel Eesti terviseandmete kohta analüüs, et selgitada välja, kui palju ja milliseid vigu esineb terviseandmetes. Analüüsi tulemused annaksid aimu, kui palju terviseandmetes esinevad vead võivad tüübi tuvastamist segada.

Tüübituvastus peab hakkama saama ka olukorraga, kus tuleb otsustada mitme sobiva andmetüübi vahel. Selline olukord tekib, kui ühe andmetüübi lubatud väärtused on alamhulgaks teise andmetüübi väärtustele. Näiteks kui veerg koosneb vaid väärtustest „M“ ja „N“, siis antud veeru tüübiks sobivad nii sugu kui ka ATC-kood. Tuvastaja peab antud olukorras eelistama veeru tüübiks sugu, sest selle andmetüübi lubatuid väärtusi kaetakse protsentuaalselt rohkem kui ATC-koodi lubatuid väärtusi.

Samuti võib tekkida olukord, kus osa veeru väärtustest on tühjad, näiteks veergude sisuks on patsiendi surma kuupäev, kuid ta patsient ise pole surnud, või väljaostetud ravimi toimeaine koodi, mis lisandub pärast ravimi ostmist. Sel juhul tuvastaja peab töötlema vaid tühisõnest pikemaid väärtusi.

Võib juhtuda, et on vaja üldistada kuupäeva 02.03.2020. Sel juhul ei ole teada, mismoodi täpselt seda teha. Antud kuupäev võib olla nii 2. märts kui ka 3. veebruar, mis tähendab, et antud kuupäeva kuu tasemeni saab üldistada viisidel „02.2020“ ja „03.2020“. Et andmete anonüümijale saaks anda õige kuupäeva formaadiga hierarhia, tuleb tuvastada lisaks ka veerus olevate kuupäeva formaat.

Tüübi tuvatamise komponent ei ole eraldi käivitatav, vaid on osa tarkvarast (Joonis 1). Sisendandmete töötlemise protsessi ajal antakse komponendile ette veergude nimed, mille tüüpe tuleb tuvastada, ja andmetabeli nimi, kus andmed asuvad. Komponent tagastab sõnastiku, kus võtmeteks on veergude nimed ja väärtusteks andmetüübid. Sõnastiku põhjal hiljem luuakse andmebaasi andmetabel metaandmetega, kust on mugav kätte saada iga veeru andmetüüpi.



Joonis 1. Projekti Health Sense andmete anonüümija protsess. Rohelisega on märgitud osa, millele käesolev töö pakub lahendust.

Teadaolevalt pole varasemalt tehtud töid, mis automaatselt tuvastaksid tabeli kujul olevate andmete kategoorilisi veerge, mis on anonüümimise tarkvara juures kõige olulisemad. Leitud tööd ja õpetused toovad vaid välja, kuidas tuvastada, kas veerg koosneb täis- või reaalarvuduest, kuupäevadest või sõnedest. Kategoorilised andmed on aga sõneliste väärtuste alamhulk.

2. Metoodika

Käesolevas peatükis tutvustatakse terviseandmete analüüsi ja tüübi tuvastamise metoodikaid.

2.1 Terviseandmete kategooriliste väärtuste vigade analüüs

Terviseandmete analüüsimise eesmärk on kirjeldada sagedasi kategooriliste väärtuste vigu, et mõista nende mõju tüübituvastusele. Antud analüüsis loetakse väärtus vigaseks, kui see ei ole tühi, kuid ei vasta andmetüübi struktuurile või puudub lubatud väärtuste loendist. Analüüsimisel kasutatakse kolme Eesti keskset terviseandmekogu: Eesti Haigekassa raviarveid, Retseptikeskuse ning Tervise Infosüsteemi (TIS) andmeid. Analüüsitakse nende andmekogude kuupäevasid, diagnooside, tervishoiuteenuste ning ATC-koodide esinemissagedust. Andmed on osa projekti „Strateegilise TA tegevuse toetamine (RITA1)“ arendamisel kasutatud andmetest, mis on 10% juhuvalim nendest kolmest andmekogust ajavahemikust 2012-2019 [14]. Patsientide konfidentsiaalsuse tagamiseks saadud andmetes on vaid kaks veergu: väärtus ja väärtuse esinemiskordade arv (Tabel 3).

Tabel 3. Näide vigade analüüsi aluseks olnud andmetabeli kolmest esimesest reast. Analüüsitavad andmetes oli kaks veergu: väärtus ja esinemiskordade arv.

atc	atc_count
C07AB02	90012
A02BC01	38510
A10BA02	38014

Kategoorilisi ja kuupäevalisi väärtusi analüüsitakse sarnaselt. Mõlema andmetüüpide puhul on olemas lubatud väärtuste loend, kuid kuupäevaliste jaoks peab seda ise käsitsi genereerima teatud kuupäeva vahemiku jaoks. Esmalt filtreeritakse kõik õiged väärtused välja nii, et kontrollitakse andmetüübi lubatud väärtuste loendi pihta, kas sellised väärtused võivad eksisteerida. Seejärel alles jäänud vigaste väärtuste puhul uuritakse, miks väärtused on vigased, ja lisatakse vea põhjused. Uude veergu lisatavad vea põhjused:

- „väike täht“, kus väärtus oli vigane vaid seetõttu, et suurte tähtede asemel sisaldas väikeseid;
- „üleliigsed tühikud“, kus väärtus sisaldas üleliigseid tühikuid;

- „vale struktuur“, kus väärtus ei sobi andmetüübi struktuuriga kokku. Hiljem sellise veaga väärtusi uuritakse lähemalt ja üritatakse jaotada omakorda alamgruppideks;
- „pole loendis“, kus väärtust ei eksisteeri lubatud väärtuste loendis.

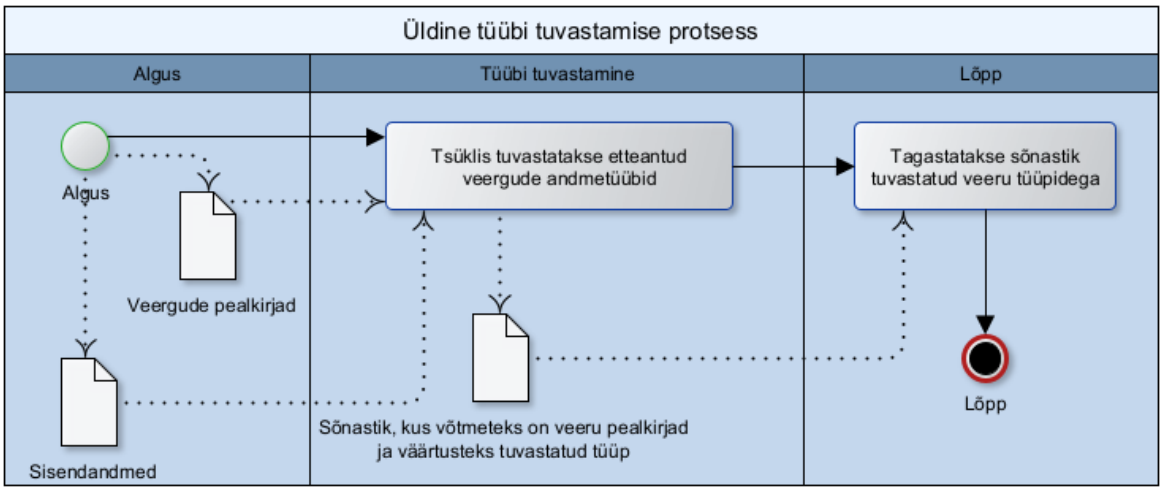
Vigu määratakse eelmainitud nimekirja järjekorras ning iga viga klassifitsserub täpselt ühte kategooriasse. Analüüsi tulemused tuuakse välja koos tabeli ja enim levinud vigadega.

2.2 Tüübi tuvastamine

Selles peatükis kirjeldatakse käesoleva töö raames loodava tüübituvastaja tööpõhimõte ning selle testimist.

2.2.1 Meetodi tööpõhimõte

Tüübi tuvastamise komponendi eesmärk on automaatselt tuvastada ette antud tabelikujul olevate terviseandmete veerud. Komponendile antakse ette veergude nimed, mille tüüpe tuleb tuvastada, ning andmetabeli nimi, kus veerud asuvad (Joonis 2). Seejärel kutsutakse tsüklis välja veeru tuvastamine kõikidele etteantud veergudele ning tagastatud tuvastatud tüübid kirjutatakse sõnastikku, kus võtmeteks on veeru pealkiri ning väärtusteks veeru tüüp. Tsükli lõppedes tagastatakse sõnastik.

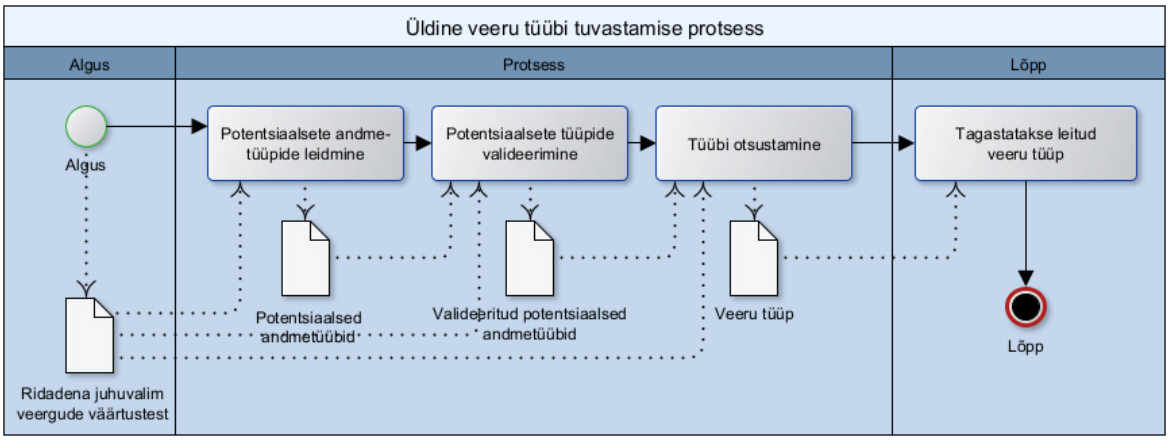


Joonis 2. Üldine tüübi tuvastamise protsess.

Veeru tüübi tuvastamiseks ei ole vaja kõiki väärtusi vaadelda, sest see on ressursikulukas. Piisab vaid võtta mingi arv, mida fikseeritakse realisatsioonis konstandina, veeru suvalisi väärtusi, mis on tühisõnest pikemad. Kui väärtusi on vähem kui fikseeritud konstant, siis sel juhul võetakse nii palju väärtusi, kui neid on.

Tuvastatavad andmetüübid on kategoorilised ja kuupäevalised. Antud töös tuvastatakse diagnooside koode, ATC-koode, EHAK-u koode, sugu, vanust ja kuupäevi formaadis „YYYY-MM-DD“, „MM-DD-YYYY“ ning „DD-MM-YYYY“ “, kuid meetod on kergesti laiendatav ka teistele hierarhilistele tüüpidele. Samuti on olemas ka tüüp „tundmatu“, mis märgib ära veerge, mida ei suudetud tuvastada. Ühine osa tuvastatavates ja töös kirjeldatud andmetüüpides on see, et igal andmetüübil on olemas kindel struktuur. Lubatud väärtused ei ole suvalise pikkusega ja suvaliste sümbolitega. See tähendab, et tüübi tuvastamisel saab kasutada regulaaravaldisi ehk *regex*'eid (Lisa 1). Iga konkreetse andmetüübi jaoks saab kirjeldada regulaaravaldise, mis kirjeldab kõigi lubatud väärtuste vormi. Regulaaravaldisi on mõistlik kasutada vaid esmaste potentsiaalsete tüüpide välja selgitamiseks, sest kuigi see viis on kiirem, kui kontrollida igat väärtust, kas see eksisteerib andmetüüpide lubatud väärtuste loendis, ei kontrollita regulaaravaldistega, kas väärtus on tõesti lubatud. Regulaaravaldistega vaid kontrollitakse, kas väärtused vastavad antud andmetüübi sarnase struktuuriga väärtustele või mitte. Seetõtti pärast esmast tüübi tuvastamist on vaja leitud tüübid ka valideerida.

Veeru tüübi tuvastamine koosneb kolmest osast: potentsiaalsete veerutüüpide leidmine, nende valideerimine ning valideeritud tüüpide seast kõige sobilikuma valimine (Joonis 3).

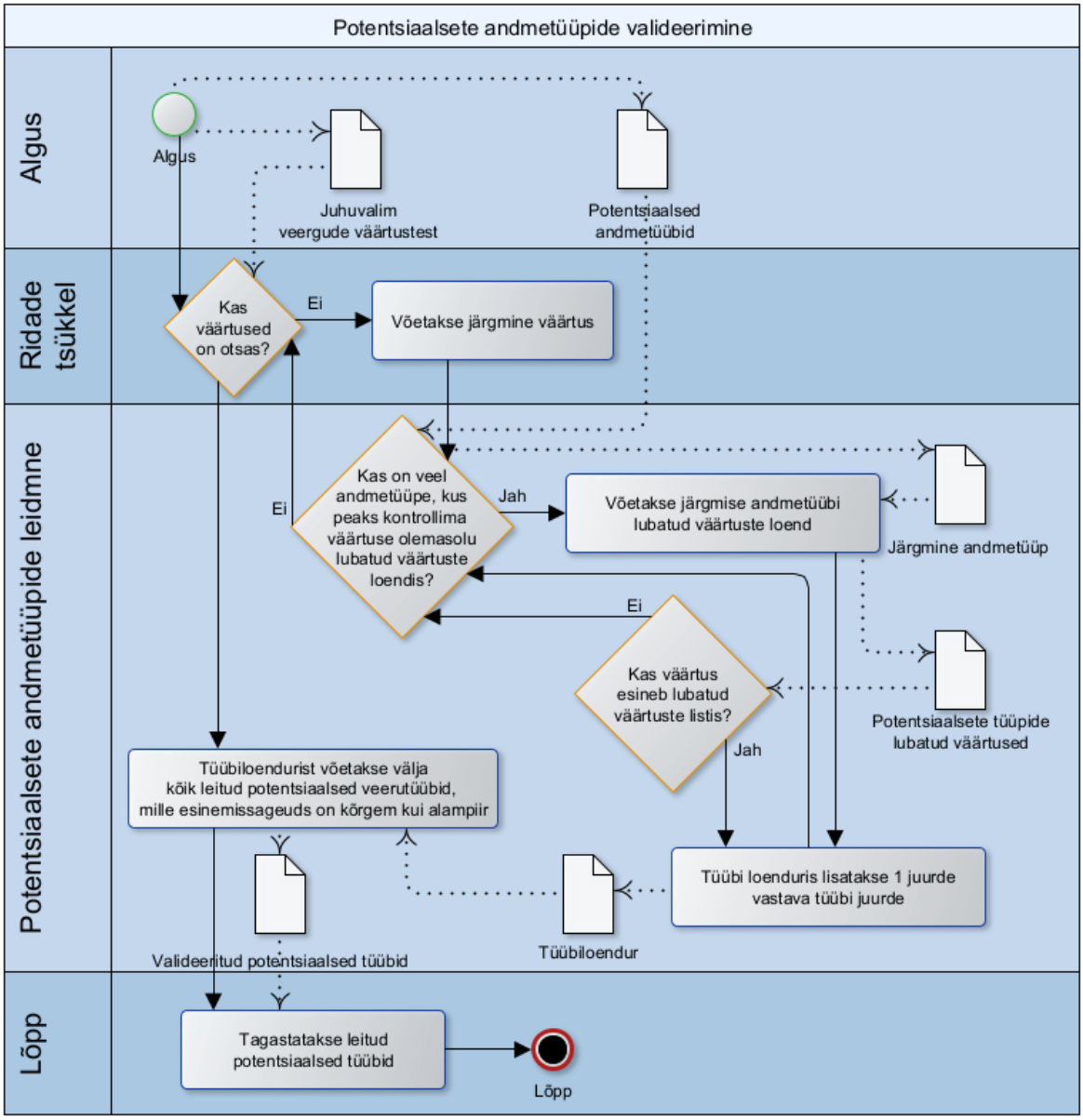


Joonis 3. Skeem ette antud veeru väärtuste põhjal tüübi tuvastamine.

Potentsiaalsete veerutüüpide leidmisel kasutatakse tüübiloendurit (Joonis 4), mis loendab kokku, mitmel korral veeru juhuvalimina valitud väärtused vastavad millistele regulaaravaldistele (Joonis 5). Võib juhtuda, et üks väärtust vastab mitmele regulaaravaldisele, sel juhul tüübiloenduris suurendatakse mõlema andmetüübi esinemiskordade arvu 1 võrra. Kui väärtus ei vasta ühelegi regulaaravaldisele, siis suurendatakse andmetüübi „tundmatu“ esinemiskordade arvu. Kuupäeva formaadi tuvastamisel suurendatakse lisaks tüübi esinemisele ka formaadi esinemist (Joonis 4). Pärast väärtuste läbimist võetakse tüübiloendurist välja potentsiaalsed andmetüübid. Andmetüüp muutub potentsiaalseks, kui kindlaks määratud protsent väärtustest vastab regulaaravaldisele. Näiteks kui võetakse andmetabelist 5000 suvalist väärtust ja on määratud, et 80% väärtustest peab kindlasti tüübi regulaaravaldisele vastama, siis andmetüüp muutub potentsiaalseks, kui vähemalt 4000 väärtust vastavad. Nii välditakse olukordi, kus andmetüüp peaks olema „tundmatu“, kuid veerus esinesid näiteks üksikud ATC-koodid, mille esinemisi märgiti ära tüübiloenduris. Iga veeru jaoks arvutatakse alampiir eraldi, sest alati ei ole võimalik tühjade väärtuste või väikese andmestiku puhul tõttu võtta fikseeritud konstant arvu ridu. Kui üheks potentsiaalseks andmetüübiks on ka kuupäev, siis see tagastatakse kuupäeva formaadiga, mida esines kõige rohkem. Antud Joonise 4 korral tagastatakse formaat „DD-MM-YYYY“.

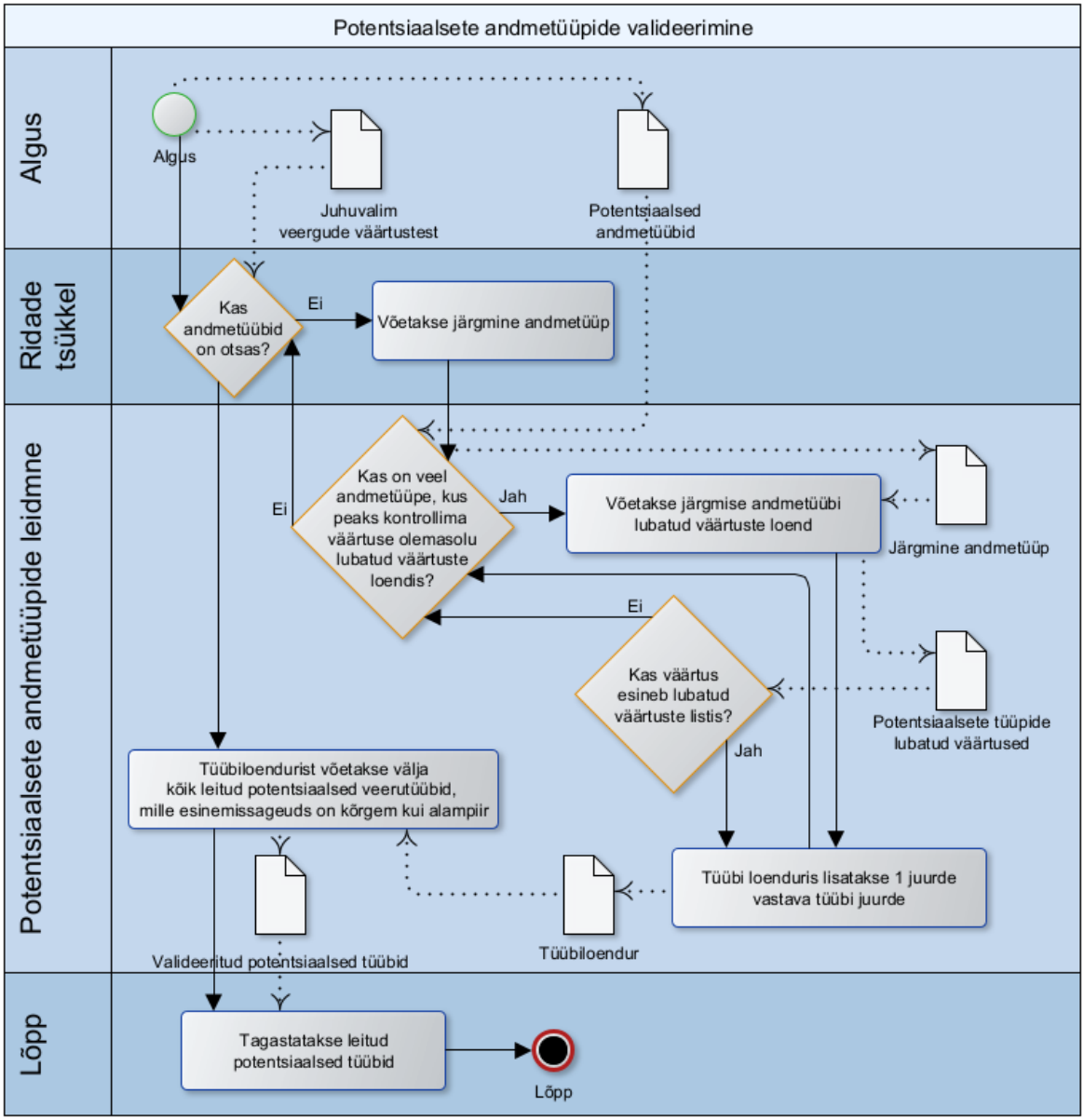
```
{'date': {'sum': 1000, 'format': {'DD-MM-YYYY': 950, 'MM-DD-YYYY': 437}}}

{'age': {'sum': 995}, 'ehak': {'sum': 701}}
```

Joonis 4. Näited sõnastiku kujul tüübiloenduritest. Esimesel juhul 1000 väärtusest 1000 väärtust olid vastavuses kuupäeva regulaaravaldistega, teisel juhul 1000 väärtustest olid 995 väärtust vastavuses vanuse ning 701 väärtust EHAK-u regulaaravaldisega.



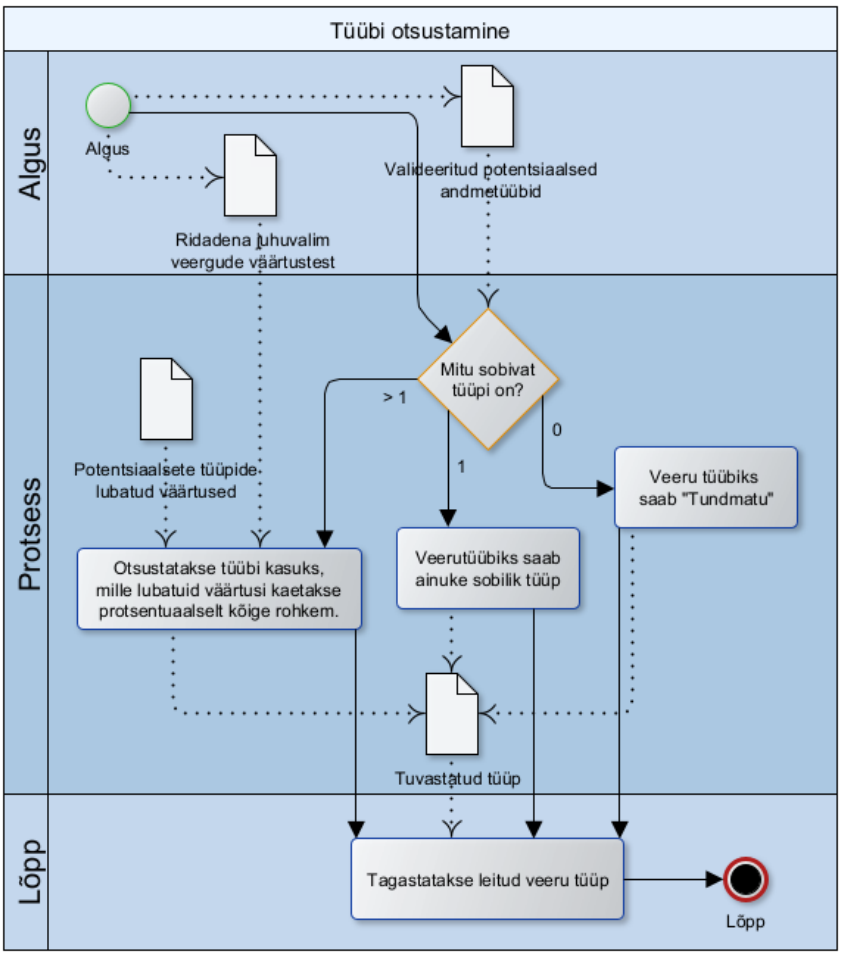
Joonis 5. Veeru tüübi tuvastamisel potentsiaalsete andmetüüpide leidmine.

Järgmisena valideeritakse potentsiaalseid andmetüüpe (Joonis 6). Valideerimise all peetakse silmas kontrollimist, kas veeru väärtused lisaks andmetüübi struktuuri sobivusele, on olemas ka lubatud väärtuste loendis. Protsess sarnaneb potentsiaalsete andmetüüpide leidmisele, kus samuti kasutatakse tüübiloendurit. Seekord loetakse kokku, mitmel korral juhuvalimina võetud väärtused esinevad andmetüüpide lubatud väärtuste loendis. Seejärel tüübiloendurist võetakse välja valideeritud potentsiaalsed andmetüübid, mille esinemiskordade arv ületab etteantud alampiiri.



Joonis 6. Veeru tüübi tuvastamisel potentsiaalsete andmetüüpide valideerimine.

Viimase sammuna valitakse valideeritud potentsiaalsete andmetüüpide vahel välja kõige sobivam tüüp (Joonis 7). Kui potentsiaalseid ei ole, siis tagastatav tüüp on „tundmatu“. Kui potentsiaalseid on üks, siis tagastatakse see. Kui neid on mitu, siis valitakse välja andmetüüp, mille lubatuid väärtusi kaetakse protsentuaalselt kõige rohkem võrreldes teiste potentsiaalsete andmetüüpidega. Näiteks kui valik jääb ATC-koodi ja soo vahel, siis valitakse sobivaks tüübiks antud juhul „sugu“.



Joonis 7. Veeru tüübi tuvastamisel tüübi otsustamine valideeritud potentsiaalsete andmetüüpide seast.

Projekti Health Sense andmete anonüümija tüübi tuvastamise komponent implementeeritakse Pythoni programmeerimiskeeles, sest seda keelt kasutatakse tarkvara sisendandmete töötlemiseks. Sisendandmed ja hierarhiad imporditakse SQLite andmebaasi, kust hakatakse võtma tüübi tuvastamise komponendi jaoks vajalikke väärtusi SQL päringutega.

2.2.2 Meetodi testimine

Meetodit testitakse läbi projekti Health Sense terviseandemte anonüümimise tarkvara. Testimiseks kasutatakse arvutit, millel on peal operatsioonisüsteem Windows 10, 8-tuumaline AMD protsessor „Ryzen 7 2700“ ning 16 GB-i muutmälu ehk RAM-i.

Meetodi testimisel pööratakse tähelepanu tüübi tuvastamise õigsusele ja ajakulule. Testimisel kasutatakse 6 isekoostatud testfaili ning anonüümimise tarkvara testimiseks TEHIK-u poolt

antud näidisfaili. Kaks isekoostatud testfailidest immiteerivad väikeseid terviseandmistikke. Failid koosnevad 10 ja 1000 reast ning järgmistest andmetest:

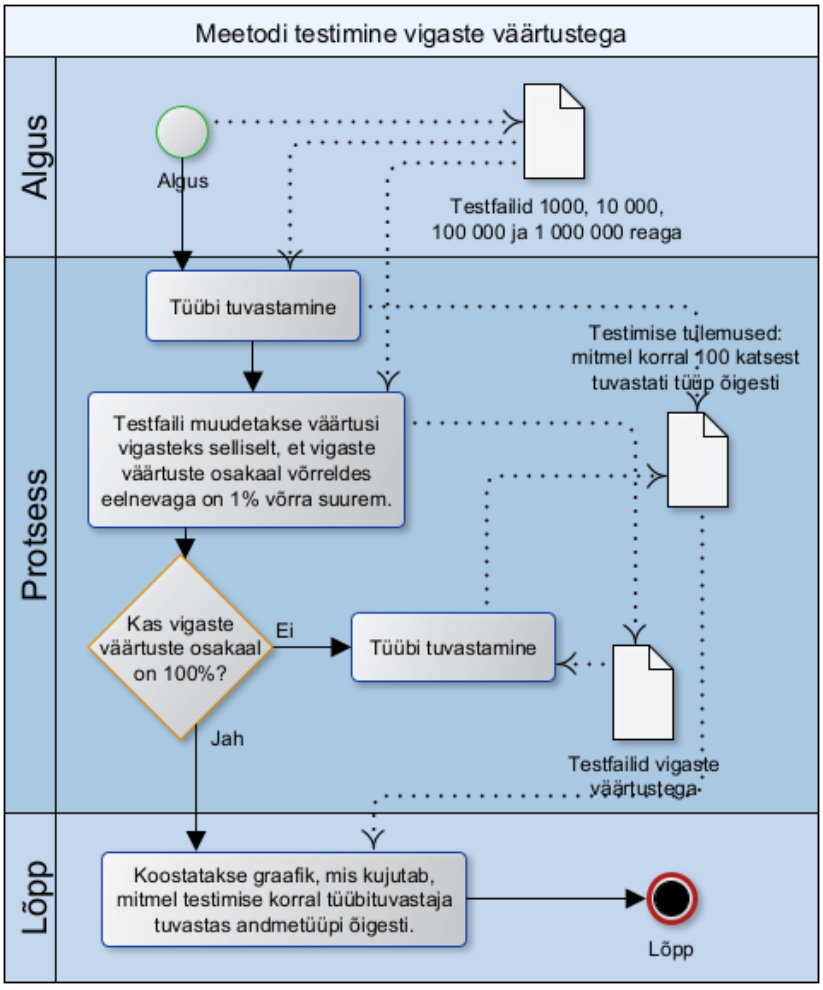
- patsiendi id, kus väärtused on 4-kohalised arvud. Väärtused võivad segamini minna EHAK-u neljakohaliste koodidega, seega sellega kontrollitakse potentsiaalsete andmetüüpide valideerimist;
- sugu väärtustega „M“ ja „N“;
- patsiendi vanus, mis võib segamini minna EHAK-u kahekohaliste koodidega, seega sellega kontrollitakse potentsiaalsete andmetüüpide valideerimist;
- sünnikuupäev, kus enamus kuupäevad on formaadis DD-MM-YYYY, kuid esineb ka üksikuid kuupäevi, mis on selgesõnaliselt formaadis MM-DD-YYYY nagu näiteks „01.31.2020“. Nii kontrollitakse, kas tuvastatakse ka õige kuupäeva formaat;
- elukoht, mis on tähistatud EHAK-u koodiga 3 erineval viisil: nullideta lühike kood, nullidega lühike kodd ja pikk kood. Nullideta lühikese koodi puhul tehakse kaks veergu: ühes väärtused vahemikus 100-999 (vallad, linnaosad, linnad haldusüksusena), teises vahemiksu 30-89 (maakonnad);
- diagnoos;
- diagnoosile sarnased koodid, kuid väärtused ei esine diagnooside lubatud väärtuste loendis. Sellega kontrollitakse potentsiaalsete andmetüüpide valideerimist;
- väljakirjutatud ravimi toimeaine, kus esinevad ka tühjad väärtused. Sellega kontrollitakse, kas meetod saab hakkama, kui veerud koosnevad ka tühjadest väärtustest;

Veerud ja nende väärtused on valitud selliselt, et oleksid kaetud andmetüüpide kõik võimalikud esinemisviisid ning ka erijuhud, mis on välja toodud peatükis 1.3.

Failide genereerimisel EHAK-u, diagnoosi, ATC, soo ja kuupäevade veergude väärtused võetakse suvaliselt vastavate andmetüüpide lubatud väärtuste loendist. Diagnoosile sarnasete koodide väärtuste jaoks luuakse eraldi loend 20 väärtusega, mille väärtused vastavad diagnoosikoodi struktuurile, kuid Eestis neid ei esine. Patsiendi vanus ja sünnikuupäeva ning diagnoosi ja välja kirjutatud ravimi toimeaine väärtustel puudub omavaheline seos, sest ka andmete anonüümimisel ei arvestata veergude omavaheliste sisuliste seostega.

Neli ülejäänud isekoostatud testfaili koosnevad 1000, 10 000, 100 000, 1 000 000 reast ning ühest veerust diagnoosi väärtustest. Need on alusfailid, sest testimise käigus muudetakse tsüklis nendes failides olevaid väärtusi osaliselt vigasteks, et vaadata, kuidas sisendandmete

suurus ning vigaste väärtuste osakaal mõjutavad tüübi tuvastamise õigsust (Joonis 8). Iga tsükliga muudetakse diagnoosi väärtusi vigasteks nii, et nende esinemissagedus failis oleks 1% võrra suurem võrreldes eelnevaga. Kui vigaste väärtuste arv ei ole täisarv, siis ümardatakse see alla. Tüüpi proovitakse tuvastada samade väärtustega 100 korral. Kuna tüübi tuvastamiseks võetakse sisendandmetest juhuvalimina fikseeritud arv väärtusi ja juhuslikkus on mängus, siis vigaste väärtuste korral võib tuvastaja mingitel kordadel veeru andmetüüpi tuvastada ja mingitel kordadel mitte.



Joonis 8. Skeem tüübi tuvastamise testimise protsessist vigaste väärtustega.

TEHIK-u näidisfail on genereeritud TEHIK poolt vaktsineerimiste andmete näitel ja koosneb 43 veerust ning veidi üle 500 000 reast. Andmed selles failis on küll tehislikud, kuid genereerimisel üritati jälgida päriselu mustreid. Seega selle failiga pannakse meetod proovile

päriselu sarnasele andmestikuga, mille sarnaseid faile soovitakse tervisandmete anonüümimise tarkvarast läbi lasta. Fail koosneb kõikidest tuvastavatest andmetüüpidega veergudest välja arvatud diagnoosikoodidest. Järgnevalt tuuakse välja veeru kirjeldused, mida meetod peab suutma tuvastada:

- vaktsiini kehtivuse algus;
- vaktsiini kehtivuse lõpp, mis osaliselt koosneb ka tühjadest väärtustest;
- dokumendi loomise aeg;
- dokumendi vastuvõtmise aeg;
- vaktsiini ATC-kood;
- immuniseerimise kuupäev;
- järgmise immuniseerimise kuupäev;
- patsiendi sünniaeg;
- patsiendi sugu;
- patsiendi 2-kohaline EHAK-u kood;
- patsiendi 3-kohaline EHAK-u kood;
- patsiendi 4-kohaline EHAK-u kood.

Fail koosneb ka paljudest teistestki veergudest nagu näiteks „Arsti kood“, „Terviseteenuse osutaja nimi“ ja „Dokumendi versioon“, millest mõned võivad segamini minna vanuse või EHAK-u koodide väärtustega, sest lubatud väärtuste loendis on neil samuti täisarvulised väärtused.

3. Tulemused ja arutelu

Selles peatükis kirjeldatakse terviseandmete kategooriliste väärtuste vigade analüüsi tulemusi ning tüübituvastuse impelmentatsiooni koos testimise tulemustega.

3.1 Terviseandmetes esinevate kategooriliste väärtuste vigade analüüs

Terviseandmete analüüsimisel selgus, et vigu esines kõigis kolmes andmestikus, kuigi nende osakaal oli väike (Tabel 4)[4]. Kõigis andmestikus oli üle 99% korrektse väärtuse. Enamiku andmestike probleemiks on see, et kood on struktuuri poolest õige, kuid seda ei esine lubatud väärtuste loendis. Sellel võib olla mitu põhjust: andmed sisaldavad aegunud koode ja analüüsimisel kasutatud lubatud väärtuste loend ei sisaldanud neid või koodi trükkimisel sisestati valesid sümboleid.

Tabel 4. Vigade analüüsi tulemused, kus ATC 1 on välja kirjutatud ravimite toimeained ning ATC 2 on välja ostetud ravimite toimeained.

Tunnus / Andmekogu	**Koode kokku**	**Sobivaid koode %**	**Üleliigsed tühikud %**	**Väikesed tähed %**	**Vale struktuur %**	**Pole loendis %**	**Kokku vigaseid %**
Diagnoosid / Haigekassa raviarved	10 093 437	99,9999	0	0	0	$9{,}91*10^{-6}$	$9{,}91*10^{-6}$
Diagnoosid / TIS	5 852 310	99,9592	0.0001	0.0255	0.0086	0,0067	0,0408
Diagnoosid / Retseptikeskus	9 610 304	100	0	0	0	0	0
ATC / TIS	2 106 289	99,8141	0.0435	0	0	0,1424	0,1859
ATC 1 / Retseptikeskus	9 610 304	99,9100	0	0	0	0,0900	0,0900
ATC 2 / Retseptikeskus	9 610 304	99,9155	0	0	0	0,0845	0,0845

[4] Terviseandmete analüüsimiseks kasutatud Jupyter Notebook'i fail on leitav repositooriumist https://gitlab.cs.ut.ee/pqder/loputoo-failid-tuubituvastaja.

Väärtuste struktuuri mittevastavust esines vaid TIS diagnooside osas. Järgnevalt on välja toodud mõned näited:

- D-täht koodi ees („DN41.9“, „DN11“);
- punkt puudu („X1012“);
- üleliigne punkt („Z54..0“);
- üleliigsed sümbolid koodi lõpus („K51.8+“, „G01 *“, „H65 ?“);
- puuduvad sümbolid („K“);
- täiesti valed väärtused („AMBEPIKRIIS“, „???“, „A-+.9“, „39035006“).

On näha, et mõned vigased väärtused on potentsiaalselt parandatavad, eemaldades või lisades mõne sümboli. Sellist parandust võiks tulevikus teostada ka automaatselt. Samas on sel juhul äärmiselt tähtis, et enne parandamist oleks veeru tüüp õigesti määratud ja parandust ei hakataks teostama vales veerus.

Kokkuvõttes saab öelda, et kuigi terviseandmetes esineb vähe vigu, peab tüübituvastusel nende esinemisega sellegipoolest arvestama.

3.2 Tüübituvastuse realisatsioon

Tüübituvastuse metoodika implementeeriti Pythoni eraldi seisvasse faili „*column_detection.py*“. Meetodit välja kutsudes antakse kaasa veergude nimed, mida vaja tuvastada, ning andmebaasi nime, kus andmed asuvad. Lisaks loodi ka funktsioone teistesse failidesse, kus pöörduti SQL päringuga andmebaasi poole ning leiti kuupäeva väärtuse eraldaja või lubatude väärtuste loend. Funktsioone loodi teistesse failidesse, sest need võivad olla kasulikud ka teistele anonüümija tarkvara komponentidele. Tüübituvastuse komponent tagastab sõnastiku, kus igale etteantud veeru nimele vastab tuvastatud tüüp (Joonis 9). Loodud tarkvara kood koos kasutamisõpetusega on kättesaadav Git-i koodirepositooriumist https://gitlab.cs.ut.ee/pqder/loputoo-failid-tuubituvastaja.

```
{
    'patsiendi_id': {'type': 'UNK'},
    'sugu': {'type': 'gender'},
    'vanus': {'type': 'age'},
    'sunnikuupaev': {'type': 'date', 'format': 'DD-MM-YYYY'},
    'diagnoos': {'type': 'dgn'},
    'vale_diagnoos': {'type': 'UNK'},
    'valja_kirjutatud_ravim': {'type': 'atc'},
    'elukoht': {'type': 'ehak'}
}
```

Joonis 9. Näide tüübituvastuse tagastavast sõnastikust, kus võtmeteks on veerunimed ja väärtusteks tüüp ning formaat.

Terviseandmete analüüsimisel selgus, et 99% väärtustest on korrektsed. Metoodika realiseerimisel kasutati seda teadmist konstantide väärtustamisel: andmeveergudest võetaks võimalusel 1000 suvalist väärtust ning vähemalt 90% nendest peavad olema andmetüübi regulaaravaldisega vastavuses, et andmetüüp muutuks potentsiaalseks. Kui ridu veerus on vähem kui 1000, siis võetakse kõik, kuid 90% nendest peavad olema regulaaravaldisega vastavuses. Samuti potentsiaalse andmetüübi valideerimise puhul peavad 90% väärtustest esinema lubatud väärtuste loendis.

Testimiseks kasutati 5 isekoostatud andmefaili ning anonüümimise tarkvara testimiseks TEHIK-u poolt antud näidisfaili[5]. Kaks isekoostatud testfaili testisid tüübituvastaja õigsust juhtudel, kus kõik väärtused olid õiged, või erijuhtumite puhul, kus mõned väärtused olid tühjad, mõned väärtused olid struktuuri poolest õiged, kuid ei esinenud lubatud väärtuste loendis, ning mõned väärtused olid omased mitme andmetüübile. Mainitud kahe failiga testimine võttis aega 10 rea puhul 0.12 sekundit ning 1000 rea puhul 2.11 sekundit. Testimise käigus selgus, et isekoostatud failide põhjal tuvastati peaaegu kõik veerud õigesti (Tabel 5). Valesti tuvastati EHAK-u veerg, kus väärtusteks olid nullideta kahekohalised täisarvaud, mis pidid tähistama patsiendi asukoha maakonda. Selle veeru tüübiks tuvastati „vanus“. Hierarhiates on vanuse väärtusi vahemikus 0-130 ja maakondade EHAK-u koodi vahemikus 30-89. Vanuse väärtusi on rohkem, seega tüübituvastuse juures veeru valideeritud andmetüüpide seast kõige sobilikuma tüübi valimisel oleks pidanud langema valik EHAK-u peale. Teglikkuses on EHAK-u hierarhias kõik lubatud koodid: nii pikad kui ka lühikesed, nii

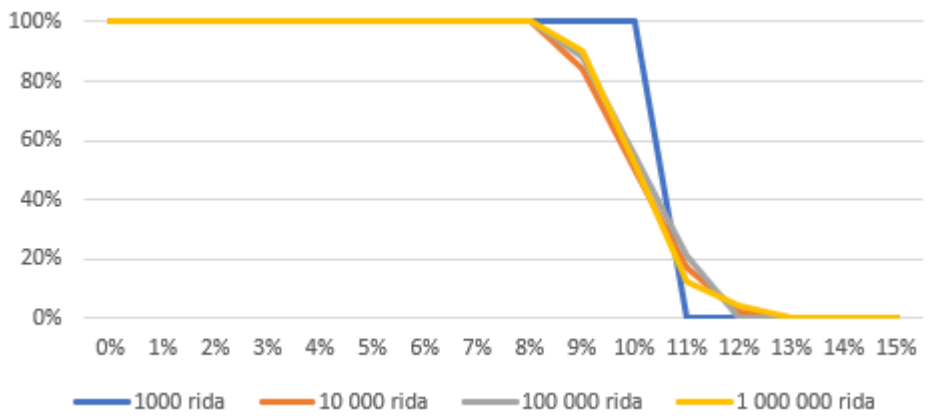
[5] Tüübituvastuse testfailid on leitavad repositooriumist https://gitlab.cs.ut.ee/pqder/loputoo-failid-tuubituvastaja. TEHIK-u poolt antud näidisfaili ei ole avalikult jagatavad konfidentsuaalsuse tõttu.

kahe-, kolme kui ka neljakohalised. Seetõttu tegelikult protsentuaalselt kaeti rohkem vanuse lubatuid väärtusi, kui EHAK-u omasid.

Tabel 5. Tulemused, mida saadi tüübituvastuse komponendi testimisel isekoostatud failidega.

Andmetüüp failis	**Veerge failis**	**Õigesti tuvastatud 10 reaga failis**	**Õigesti tuvastatud 1000 reaga failis**	**Kommentaar**
kuupäev	1	1	1	Lisaks tüübile tuvastati ka õige formaat.
ATC	1	1	1	
sugu	1	1	1	
EHAK	4	3	3	Veerg nullideta lühikese koodiga, kus väärtused olid vahemikus 30-89, tuvastati kui vanus.
diagnoos	1	1	1	
vanus	1	1	1	
tundmatu	2	2	2	
KOKKU	11	10	10	

Neli ülejäänud isekoostatud testfaili koosnesid 1000, 10 000, 100 000 ja 1 000 000 reast ning ühest diagnoosi veerust. Nende failidega testiti, kui suur osa andmetest võivad olla vigased ilma, et see oluliselt mõjutaks ka tüübi tuvastamist erinevate suuruste failide puhul (Joonis 10). Joonisest 10 selgub, et kuni miljoni reaga veeru puhul võivad terviseandmed olla kuni 8% vigased, et tüübituvastaja 100%-lt tuvastaks õiget andmetüüpi. Kui vigaste andmete osakaal on suurem kui 8%, siis langeb ka tõenäosus, et tüüp tuvastatakse õigesti. Kui vigaste väärtuste osakaal on 13%, siis on õige tüübi tuvastamise tõenäosus juba 0%. 1000 rea puhul tüübituvastaja ei suuda tuvastada tüüpi alles 11% vigaste väärtuste puhul, sest juhuvalimina võetakse võimalusel andmestikus 1000 rida ning 90% juhtudest peab see vastama ka andmetüübile.



Joonis 10. Tüübi tuvastamise meetodi testimine vigaste väärtustega. X-teljel on vigade osakaal terve andmestiku peale, y-teljel näidatakse, mitmel korral 100 korrast tuvastati andmetüüp „diagnoos“ õigesti.

Terviseandmete kategooriliste väärtuste vigade analüüsist selgus, et analüüsitavate andmetest olid vigased ligikaudu 0.01%. See tähendab, et analüüsitud andmetele sarnasete terviseandmete veergude tüübi tuvastamist ei sega nendes esinevad vigased väärtused.

Testimine TEHIK-u poolt antud näidisfailiga võttis aega 11.89 sekundit. Tüübi tuvastamine õnnestus 38 veeru puhul (88%) (Tabel 6). Mitmed veerud tuvastati valesti. Soo veerg jäi tuvastamata, sest antud töös selle andmeveeru lubatuteks väärtusteks loeti „M“ ja „N“, mitte konkreetses testfailis esinenud *„Male“* ja *„Female“*. Samuti kuupäevade veerg koos kellaaegadega jäi tuvastamata, sest realisatsiooni ajal ei arvestatud sellega. Tundmatuid veerge oli vähem, kui pidi olema, sest veerud „Dokumendi versioon“ ja „Manustatud doosi kordsus“ tuvastati kui „vanus“, sest need veerud koosnesid 1-kohalistest arvudest ja muid konkureerivaid arvulisi andmetüüpe ei tarkvarale lisatud.

Tabel 6. Tulemused, mida saadi tüübituvastuse komponendi testimisel TEHIK-u poolt antud näidisfailiga.

Andmetüüp failis	**Veerge failis**	**Õigesti tuvastatud**	**Kommentaar**
kuupäev	7	5	2 veergu koos kellaaegadega jäid tuvastamata.
ATC	1	1	

sugu	1	0	Veerg jäi tuvastamata, sest väärtused failis olid *„Male“* ja *„Female“*, mitte „M“ ja „N“.
EHAK	3	3	
tundmatu	31	29	Veergude „Dokumendi versioon“ ja „Manustatud doosi kordsus“ andmetüübiks tuvastati „vanus“.
KOKKU	43	38	

Kokkuvõttes testi tulemused näitavad, et loodud tüübituvastuse komponent saab üldiselt hakkama õige tüübi tuvastamisega ning lisaks ka kuupäevaliste väärtuste puhul tuvastab õige kuupäeva formaadi ja eraldaja. Lisaks saab komponent hakkama veel veergudega, kus esinevad tühjad väärtusted ja väärtused, mis sobivad mingi andmetüübi struktuuriga kokku, kuid tegelikult need ei ole kasutusel. Siiski tuvastamises esineb puudusi.

Puudusi oleks võimalik likvideerida täiendades andmetüüpide hierarhiasid ning tüübi tuvastamise meetodid. Et EHAK-u kood ei läheks segamini vanusega nagu esimeses testjuhus, siis tuleks EHAK-u hierarhia tükeldada väiksemateks osadeks. Erinevate pikksutega EHAK-u koodid tuleks ära jagada erinevatesse hierarhiatesse. Lähtuvalt TEHIK-u poolt antud andmestikus tuleks lisada soo hierarhiale lubatuteks väärtusteks *„Female“* ja *„Male“*. Samuti tuleks välja mõelda, kuidas saaks tuvastada kuupäeva andmetüüpi, kui veeru väärtusteks on lisatud ka kellaaeg.

Metoodika realiseerimisel kasutati kahte konstanti: andmeveergudest võetaks võimalusel 1000 suvalist väärtust ning vähemalt 90% nendest peavad olema andmetüübi regulaaravaldisega vastavuses, et andmetüüp muutuks potentsiaalseks. Kuigi need konstandid sobisid antud testandmestikega, ei ole arvud leitud matemaatiliselt. Tüübituvstaja edasi arendamisel võiks leida kõige optimaalsemad arvud konstantide jaoks.

Kokkuvõte

Antud töö eesmärk oli luua projekti Health Sense raames arendatava terviseandmete anonüümise tarkvarale automaatse andmete veergude tüübituvastaja. Kuna terviseandmed võivad koosneda mitmekümnest veerust, siis selleks, et teha tarkvara mugavamaks ja kasutajasõbralikumaks, sooviti teha tüübi tuvastamine automaatseks.

Tüübi tuvastamise protsess jagati kolmeks osas. Regulaaravaldiste abil leitakse, milliste andmetüüpide struktuuriga veeru väärtused sobivad. Seejärel kontrollitakse veeru väärtuste vastavust nende andmetüüpide lubatud väärtuste loendiga. Viimases sammus valitakse sobivatest andmetüüpidest kõige sobilikum. Lisaks analüüsiti töös Eesti kategooriliste väärtustega terviseandmeid, et aru saada, kas terviseandmetes esinevad vead võivad olla probleemiks tüübi tuvastamise juures.

Realiseeritud tüübituvastamise meetod tuvastas testimise käigus suurema osa veergudest õigesti nii isekoostatud failidel kui ka TEHIK-u poolt näidisandmeks antud failil. Samuti testimise käigus selgus, et meetod suudab tuvastada veerge, mis sisaldavad kuni 8% vigaseid väärtusi, kui meetodis kasutada töös paika pandud konstante. Kuna eelnevalt mainitud terviseandmete kategoorilsite väärtuste vigade analüüsis selgus, et analüüsitavates andmetes on vaid 0,1% vigaseid väärtusi, siis antud andmestike puhul vigased väärtused ei ole tüübi tuvastamises segavaks faktoriks.

Meetodi puudused ilmnesid, kui andmetüüpides esinesid uued väärtused, mida pole eelnevalt olnud lubatud väärtuste loendis: soo puhul väärtused *„Male“* ja *„Female“* ning kuupäevaliste väärtuste puhul lisaks kuupäevale ka kellaaeg. Tüübituvastaja meetodid annaks edasi arendada lisades arvuliste väärtustega andmetüüpe, et oleks võimalik tuvastada näiteks kehatemperatuuridega, vereõhu ja inimese pikkuse väärtustega veerge. Samuti tüübi tuvastamise meetodit on võimalik kasutada ka mõnes muus valdkonnas peale tervisevaldkonna.

Viidatud kirjandus

[1] Shen, X., Ma, S., Vemuri, P., Castro, M. R., Caraballo, P. J., Simon, G. J., A novel method for causal structure discovery from EHR data and its application to type-2 diabetes mellitus, *Scientific Reports* 11 (2021): 21025. https://doi.org/10.1038/s41598-021-99990-7 (09.05.2022)

[2] Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C., Iyengar, S. S., Computational Health Informatics in the Big Data Age: A Survey, *ACM Computing Surveys (CSUR)* 49, no 1 (2016): 1-36. https://www.cs.helsinki.fi/u/jilu/paper/HealthBigdata.pdf (09.05.2022)

[3] Lane, J., Balancing access to health data and privacy: a review of the issues and approaches for the future., *Health services research*, 45, no. 5p2 (2010): 1456-1467. https://doi-org.ezproxy.utlib.ut.ee/10.1111/j.1475-6773.2010.01141.x (14.11.2021)

[4] OECD. Glossary of Statistical Terms. Microdata. https://stats.oecd.org/glossary/detail.asp?ID=1656 (14.11.2021).

[5] J. Ficek, Wang, W., Chen, H., Dagne, G., Daley, E., Differential privacy in health research: A scoping review, *Journal of the American Medical Informatics Association*, Volume 28, Issue 10 (10.2021): 2269-2276.

https://doi-org.ezproxy.utlib.ut.ee/10.1093/jamia/ocab135 (14.11.2022)

[6] Inimõiguste ressurss, Terviseandmed. https://www.inimoigustegiid.ee/ee/teemad/andmed-ja-privaatsus/terviseandmed (06.05.2022)

[7] Inimõiguste ressurss, Mis on terviseadnmed? https://www.inimoigustegiid.ee/ee/teemad/andmed-ja-privaatsus/terviseandmed/mis-on-terviseandmed (06.05.2022)

[8] Ehrenstein, V., Kharrazi, H., Lehmann, H., Taylor, C. O., Obtaining Data From Electronic Health Records. In Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition, Addendum 2, *Agency for Healthcare Research and Quality*

https://www.ncbi.nlm.nih.gov/books/NBK551878/. (04.04.2022)

[9] Shickel B, Tighe PJ, Bihorac A, Rashidi P, Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis, *IEEE journal of biomedical and health informatics* 22, no. 5 (2017): 1589-1604. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6043423/#R24 (04.04.2022)

[10] World Health Organization, tõlkinud Küng, A., Rahvusvaheline haiguste ja nendega seotud terviseprobleemide statistiline klassifikatsioon, Instruktsioonide käsiraamat, *Eesti Sotsiaalministeerium* 10. väljaanne, 2.köide (1996).

https://www.sm.ee/sites/default/files/content-editors/eesmargid_ja_tegevused/Tervis/E-tervis_ja_e-tervisetoend/kaesiraamat.pdf (05.05.2022)

[11] Hinrikus, T.; Kogermann, K.; Laius, O.; Leito, S.; Raal, A.; Soosaar, A.; Teppor, T.; Volmer, D. (2019) Farmaatsiaterminoloogia 2. väljaanne, (2019).

[12] Riigi Teataja. Isikuandmete kaitse seadus § 16. https://www.riigiteataja.ee/akt/130122010011#para16. (09.05.2022)

[13] Rahvastik. Alutaguse vald. https://www.alutagusevald.ee/rahvastik (09.05.2022s)

[14] RITA tegevus 1: Strateegilise TA tegevuse toetamine, *Eesti Teadusagentuur* https://www.etag.ee/rahastamine/programmid/rita/rita-strateegilise-ta-tegevuse-toetamine/ (07.05.2022)

LISA 1

Järgnevalt tuuakse välja regulaaravaldised, mida kasutatakse tüübi tuvastamise juures:

- EHAK-u kood (Joonis 11);

```
FIRST_LEVEL = r'[3-8][0-9]'         # Maakonnad, kood 30-89
SECOND_LEVEL = r'[1-9][0-9]{2}'    # Linnad/vallad, kood alates 100 kuni 999
THIRD_LEVEL = r'[1-9][0-9]{3}'      # Külad/alevikud, kood 1000-9999

TYPE = r'[0MVL][0-9]' # 0x - maakond/vald, Mx - maaline,
                      #Vx  väikelinnaline, L - linnaline asustuspiirkond;

LONG_CODE = rf'00{FIRST_LEVEL}0{SECOND_LEVEL}{THIRD_LEVEL}{TYPE}'
SHORT_CODE = rf'((00)?{FIRST_LEVEL})|(0?{SECOND_LEVEL})|({THIRD_LEVEL})'

EHAK = rf'^({LONG_CODE}|{SHORT_CODE})$'
```

Joonis 11. Regulaaravaldis, mis vastab EHAK-u koodi struktuurile.

- diagnoos (Joonis 12);

```
# Näited sobivatest väärtustest: I00-J00, I00, I00.9, I00.07
    LET_NUM_2 = r'[A-Z][0-9]{2}'
    NUM = r'[0-9]{1,2}'

    DGN =  rf'^{LET_NUM_2}((\.{NUM})|(-{LET_NUM_2}))?$'
```

Joonis 12. Regulaaravaldis, mis vastab diagnoosi koodi struktuurile.

- ATC-kood (Joonis 13);

```
# Näited sobivatest väärtustest: A, A01, A01A, A01AB, A01AB01
    NUM_2 = r'[0-9]{2}'
    LET = r'[A-Z]'
    LET_1_2 = r'[A-Z]{1,2}'

    ATC = rf'^{LET}({NUM_2}({LET_1_2}({NUM_2})?)?)?$'
```

Joonis 13. Regulaaravaldis, mis vastab ATC-koodi struktuurile.

- kuupäev (Joonis 14);

```
# Kuupäeva formaadid: DD-MM-YYYY ja YYYY-MM-DD
# SEP - eraldaja, mis on kas punkt, sidekriips, kaldkriips või mitte midagi. (., -, /)
    DD = r'(([1-2][0-9])|(3[0-1])|(0?[1-9]))'  # Ühe või kahekohaline kuupäeva number
    MM = r'((0?[1-9])|(1[0-2]))'  # Ühe või kahekohaline kuu number
    YYYY = rf'((18)|(19)|(20)){YY}'

    DD_MM_YYYY = rf'^{DD}{SEP}{MM}{SEP}{YYYY}$'
    YYYY_MM_DD = rf'^{YYYY}{SEP}{MM}{SEP}{DD}$'
    MM_DD_YYYY = rf'^{MM}{SEP}{DD}{SEP}{YYYY}$'

    DATE = r'^(DD_MM_YYYY)|(YYYY_MM_DD)|(MM_DD_YYYY)$'
```

Joonis 14. Regulaaravaldis, mis vastab kuupäeva väärtuste struktuurile.

- sugu (Joonis 15).

```
    GENDER = r'^(M|N)$'
```

Joonis 15. Regulaaravaldis, mis vastab soo väärtuste struktuurile.

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristina Mumm,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Terviseandmete tabeli veergude automaatne tüübituvastus“, mille juhendaja on Sulev Reisberg, PhD reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristina Mumm

10.05.2022