TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Andmeteaduse õppekava

Sander Tamm

Algoritmiline definitsioon patsientide trajektooride sarnasusele

Magistritöö (15 EAP)

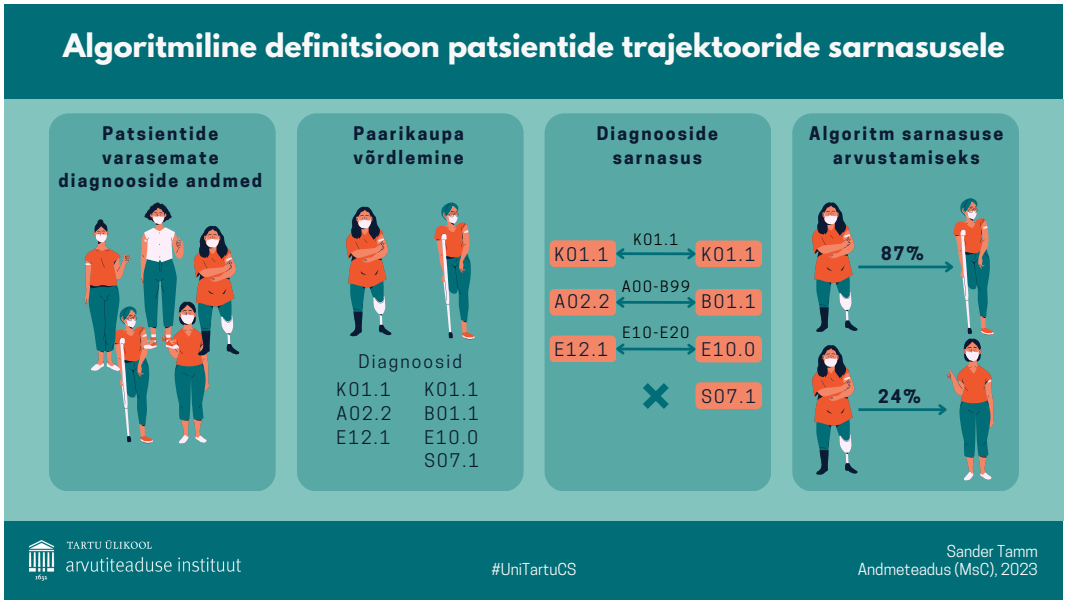
Juhendaja:
Jaak Vilo, PhD

Tartu 2023

Algoritmiline definitsioon patsientide trajektooride sarnasusele

Magistritöö
Sander Tamm

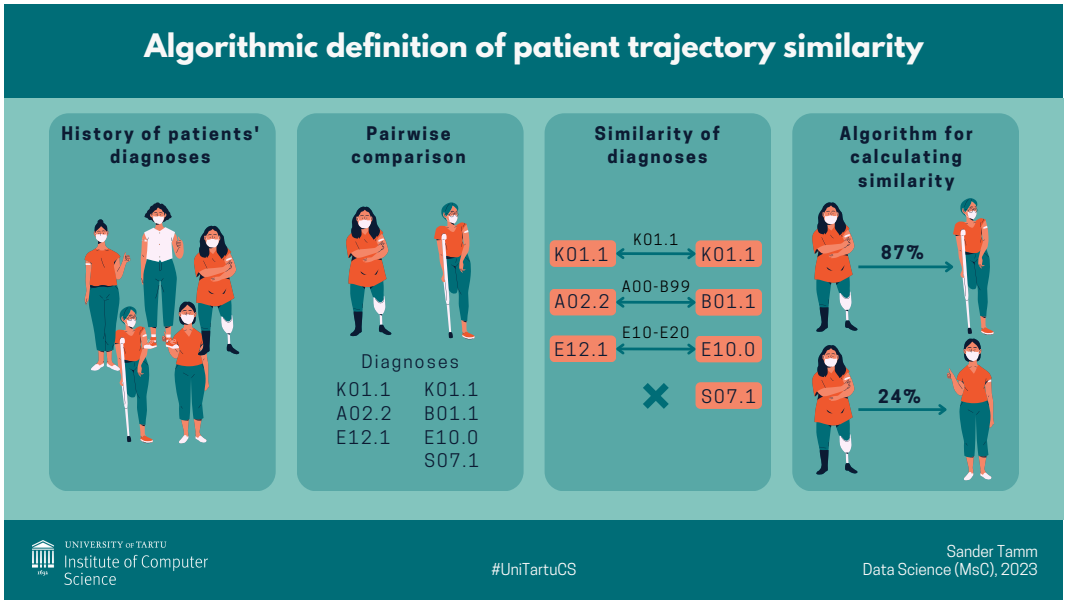
Lühikokkuvõte. Terviseandmete digitaliseerimine on kaasa toonud võimaluse patsiente omavahel võrrelda. Magistritöö eesmärgiks on luua algoritm, mis võimaldaks hinnata patsientide omavahelist sarnasust, kasutades nende varasemaid haiguste trajektoore. Selleks defineeritakse kaks Rahvusvahelise Haiguste Klassifikatsiooni (RHK) põhist diagnoosi tasemel algoritmi, millest üks tugineb RHK hierarhilisele ülesehitusele ning teine diagnooside tekstipõhisele sarnasusele. Lisaks arvestatakse iga diagnoosi jaoks nende haruldust, tõsidust ning kroonilisust. Nendest arendatakse edasi kaks patsiendi tasemel algoritmi, millel on erinevad eelised ja puudujäägid võrreldes teisega. Lisaks tuuakse välja nende omadusi ning võrreldakse kõiki algoritme omavahel, kasutades genereeritud andmeid.
CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika
Märksõnad. Terviseandmed, RHK-10, haiguste trajektoorid, patsientide sarnasuse arvutamine



Algorithmic definition of patient trajectory similarity

Master's thesis
Sander Tamm

Abstract. The digitalisation of medical data has provided the possibility to compare patients with each other. The objective of this Master's thesis is to define an algorithm, which could measure the similarity between a pair of patients based on their previous disease records. Two diagnose-based algorithms are defined, based on the International Classification of Diseases (ICD). One of them is based on the hierarchy of ICD and the other focuses on the contextual similarities of diagnoses. In addition, diagnose rarity, severeness and whether the diagnose is chronic or acute is considered. These algorithms are used to develop two different patient-based algorithms, which have different benefits and caveats. Features and comparisons of the two algorithms are provided by using synthetic data.
CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics
Key words. Medical data, ICD-10, diagnosis trajectories, patient similarity calculations



Sisukord

Sissejuhatus

Ülevaade patsientide varasematest terviseandmetest on oluline tööriist patsientide ravimeetodite hindamisel ja rakendamisel. Patsientide terviseandmete ajalugu ehk trajektoor on kogum diagnoosidest, ravist, uuringutest ja paljust muust meditsiinilisest infost. Peamiselt pööratakse tähelepanu ühe patsiendi enda varasemale terviseajaloole, kuid kasulik on patsienti võrrelda ka haigusloo mõttes sarnaste patsientidega. Kahjuks ei ole meditsiinitöötajatel tihti aega ega ressurssi iga patsiendi jaoks leida sarnaseid juhtumeid. Terviseandmete digitaliseerimine on kaasa toonud võimaluse seda algoritmide abil teatud määral automatiseerida.

Patsientide sarnasuse hindamine võimaldab meditsiiniteenuste osutamist mitmes aspektis parandada. Näiteks võimaldab see asendada teatud haiguste üldistatud ravimeetodid indiviidi tasemel raviga, vaadeldes varasemate sarnaste patsientide ravitrajektoore. Samuti võimaldab see paremini prognoosida haiguste teket, mõõta ravi efektiivsust, leida seoseid erinevate haiguste vahel ja palju muud. Eelnevad tegurid mängivad olulist rolli ka ravimiuuringute läbiviimisel.

Üks viis sarnasuste leidmiseks on kasutada andmeid, mis on korrastatud hierarhilise struktuurina. Selliseks sobib näiteks Maailma Terviseorganisatsiooni (WHO) loodud diagnooside kodeerimise meetod Rahvusvaheline Haiguste Klassifikatsioon (RHK, ing. k. ICD). Eestis on kasutusel selle kümnes väljaanne, RHK-10 [1], mille on ingliskeelest eesti keelde tõlkinud Eesti Sotsiaalministeeriumi Meditsiiniterminoloogia Komisjon.

Käesolevas töös loome algoritmi, mis suudab võrrelda erinevate patsientide RHK-10 diagnooside trajektoore ning leida iga patsiendi jaoks kõige sarnasemad teised patsiendid.

Töö on jaotatud neljaks peatükiks. Esimeses peatükis kirjeldatakse kasutatud andmestikke ning nende töötlust. Teises peatükis defineeritakse kaks diagnooside tasemel sarnasuse algoritmi, kirjeldatakse nende omadusi ning võrreldakse nende peamisi erinevusi. Lisaks on selles peatükis kirjeldatud muud tunnused, mida sarnasuse leidmiseks hinnata. Kolmandas peatükis defineeritakse patsiendi tasemel sarnasuste algoritmid, kasutades eelmises peatükis defineeritud diagnoosi tasemel sarnasusi. Neljandas peatükis võrreldakse varasemates peatükkides defineeritud algoritme omavahel, kasutades esimeses peatükis kirjeldatud genereeritud andmeid. Töö lõpus on lõik, mis võtab käesoleva magistritöö sisu ning tulemused kokku.

1 Kasutatud andmestikud

Selles peatükis käsitleme ning kirjeldame andmeid, mida on töös kasutatud. Reaalsed patsientide meditsiinilised andmed kuuluvad eriliiki isikuandmete alla [2] ning juurdepääs nendele on rangelt piiratud. Algoritmide loomiseks piisas reaalseid andmeid kirjeldavatest genereeritud andmetest ning näidete toomisel kasutati RHK-10 koode, mis kirjeldasid hästi algoritmide tulemusi.

1.1 RHK-10 diagnooside koodid

See alampeatükk tugineb raamatul [1]. Eesti kasutab diagnoosimisel RHK-10 süsteemi. Selle järgi on diagnoosi tähistuseks:

1. Tähestiku täht, mis määrab ära haiguse kõige laiema klassifikatsiooni, mida nimetame peatükiks. Üldiselt tähistab ühte rühma üks täht, kuid esinevad mõned erandid. Diagnooside tähe kattumine tähistab kõige nõrgemat sarnasust kahe diagnoosi vahel.

   Näide: Täht K tähistab seedeelundite haigust. Tähed S ja T tähistavad vigastusi, mürgistusi ja teatavaid muid välispõhjuste toime tagajärgi. Tähed $V - Y$ tähistavad haigestumise ja surma välispõhjuseid.

2. Otse tähe järgi kahenumbriline kombinatsioon, mis jaotatakse kitsamatesse haiguste kategooriatesse. Nende kahe numbri langemine samasse kategooriasse tähistab tugevamat sarnasust kahe diagnoosi vahel, täpne kattumine väga suurt sarnasust.

   Näide: $K42$ kuulub kategooriasse $[K40 - K46]$, ehk songad. $K42$ tähistab nabasonga.

3. Pärast eelmist kombinatsiooni järgneb veel üks või kaks numbrit, mis on eelmistest punktiga eraldatud. Neist esimene ehk alamjaotis tähistab üldiselt haiguse kõige spetsiifilisemat tähistust. Teine number ehk 5. koha alamjaotis (paljudel nii täpset tähistust ei olegi) tähistab üldiselt kas kaasuvaid haigusi või tekkepõhjust.

   Näide: $K42.0$ tähistab nabasonga, mis on sulgusega ja ilma gangreenita. $F01.01$ tähistab vaskulaarset dementsust ($F01$), mis on ägeda algusega ($F01.0$) ning sellega kaasuvad luululised sümptomid ($F01.01$).

Edaspidi käsitleme diagnoosi koode kuni alamjaotiseni, kuna 5. koha alamjaotis väga palju lisainformatsiooni ei anna. Samuti jätame analüüsist välja järgmiste tähtedega algavad diagnoosid:

1. S ja T - Vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed,
2. U - Ebakindla etioloogiaga uute haiguste ajutised määrangud (SARS, COVID-19),
3. V, W, X, Y - Haigestumise ja surma välispõhjused (õnnetused, kukkumised, rünnakud jms),
4. Z - Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega.

Kokku jääb analüüsimiseks 7298 RHK-10 koodi, millest 7074 on alamjaotise tasemel ning 224 jaotise tasemel, kuna neil kitsamat täpsustust RHK-10 süsteemis ei ole.

Näidete jaoks arvutatud algoritmides võtame RHK-10 koodidele vastavad tekstilised kirjeldused ning hierarhilise esituse Pythoni *simple_icd_10* [3] teegist, mis on omakorda üles ehitatud RHK-10 2019 aasta ingliskeelsele versioonile (ICD-10) [4]. Sellest tulenevalt tuginevad ka edaspidised mudelid RHK-10 koodide ingliskeelsetele tõlgendustele.

Töös on parema loetavuse eesmärgil välja toodud ka RHK-10 koodide eestikeelsed nimetused, mis on võetud Sotsiaalministeeriumi andmebaasist [5].

1.1.1 RHK-11 diagnooside koodid

Valminud on ka uus RHK versioon, nimelt RHK-11, mis jõustus aastal 2022. Eestis see veel kasutusele võetud ei ole ning seetõttu ei ole see ka käesoleva magistritöö loodud algoritmide fookuses.

RHK-11 täpsem struktuur on ära seletatud artiklis [6]. Uus klassifikatsioon on fundamentaalselt erinev, kuna ühele koodile võib vastata rohkem kui üks peatükk. See võimaldab näiteks seedeelundide kasvajad liigitada nii kasvajate alla (nagu see on RHK-10 süsteemis) kui ka seedeelundite haiguste alla. RHK-11 hierarhia lineaarseks muutmiseks on igale koodile määratud nii-öelda peamine klass, kuhu alla ta kuulub.

Kasutades vaid lineariseeritud RHK-11 süsteemi, on töös defineeritud algoritme võimalik otse rakendada. Samas RHK-11 täisvõimsuse rakendamiseks oleks hierarhiast sõltuvad algoritmid vaja teisiti defineerida.

1.2 Genereeritud patsientide andmed

Diagnooside harulduse leidmiseks on kasutusele võetud Artjom Valdase loodud sünteetiliste diagnooside genereerimise mudel (edaspidi SDG mudel) [7]. Mudeli koostamiseks on igale võimalikule RHK-10 koodile vastavusse viidud patsientide soo ja vanuse põhjal tekketõenäosused. Peatüki ja alampeatüki tasemel on tekketõenäosused tuletatud STACCi andmetest [8]. Jaotiste ning alamjaotiste jaoks selline andmestik puudus, seega originaalses mudelis on iga jaotise ja alamjaotise esinemise tõenäosus võrdeline tema naabritega. SDG mudelit algoritmide võrdlemiseks ei kasutata, kuna mudel kirjeldab hästi üldpopulatsiooni haiguste jaotumist, kuid mitte patsiendi tasemel haiguste tekkimist. Seda sellepärast, et patsiendi varasemad haigused ei mõjuta selles mudelis uute haiguste tekkimise tõenäosust.

1.2.1 Andmete täiendamine

Käesolevas töös on SDG mudeli tekketõenäosusi täiendatud Tervise Arengu Instituudi (TAI) Tervisestatistika ja terviseuuringute andmebaasi [9] haigestumuse andmete põhjal. TAI andmestike abil on võimalik tuletada paljude alampeatükkide jaotiste tekketõenäosused. Lihtsustamise huvides võtame haigestumised aastate lõikes kokku ning vanuselist erinevust ei arvesta. Alampeatükkide loetelu koos kasutatud andmestikuga on järgnev:

1. Alampeatüki $E10 - E14$ jaotuste tekketõenäosused saame tabelist EH01: Diabeedi esmashaigusjuhud soo ja vanuserühma järgi [10], valides aastate 2016-2021 andmed.

2. Nakkushaiguste ehk peatüki $A00 - B99$ alampeatükkide $A00 - A09$ ning $B15 - B19$ jaotuste tekketõenäosused saame tabelist NH02: Valitud nakkushaiguste registreeritud juhtude arv ja kordaja 100 000 elaniku kohta soo ja vanuserühma järgi [11], valides 2008.-2012. aasta andmed, kuna selles vahemikus oli kõige rohkem erinevate diagnooside väärtuseid.

3. Alampeatüki $A15 - A19$ jaotuste tekketõnäosused saame tabelist TB10: Tuberkuloosi esmasjuhud paikme, soo ja vanuserühma järgi [12], valides 2000.-2021.

aasta andmed.

4. Paljude pahaloomuliste kasvajate ehk $C00 - C97$ jaotuste tekketõenäosused saame tabelist PK10: Pahaloomuliste kasvajate esmasjuhud paikme, soo ja vanuserühma järgi [13], valides aastate 2000-2020 andmed.

5. Paljude psüühika- ja käitumishäirete ehk $F00 - F99$ jaotuste tekketõenäosused saame tabelist PKH1: Psühhiaatri poolt ambulatoorselt konsulteeritud isikud diagnoosi, soo ja vanuserühma järgi [14], valides 2000.-2021. aasta andmed.

6. Jaotuste $I21$ ja $I22$ tekketõenäosused saame kahest tabelist:

   (a) $I21 - I22$ osahulga alampeatüki $I20 - I25$ haigustest tabelist EH10: Esmashaigusjuhud soo ja vanuserühma järgi (1998-2016) [15], valides 2000.-2016. aasta andmed.

   (b) $I21$ ja $I22$ omavahelise suhte tabelist AMI02: Ägeda müokardiinfarktiga hospitaliseeritud patsiendid (juhud) soo, vanuserühma ja infarkti alatüübi järgi [16], valides aastate 2015-2021 andmed.

1.3 Näidete andmed

Reaalsete andmete peal algoritme testimisest hoidusime andmede konfidentsiaalsuse tõttu ning selleks sobivaid genereeritud andmeid ka ei leidunud. Seda arvestades valis autor diagnoosi tasemel algoritmide jaoks käsitsi välja RHK-10 diagnooside kolmikud, mis kirjeldasid hästi erinevate algoritmide iseärasusi. Patsientide tasemel algoritmide ülevaateks kasutame ka ChatGPT abi[1]. ChatGPT mudeli abil loodi kaks tabelit:

1. Kümme diabeedihaiget patsienti, kellel võib, aga ei pruugi diabeet (RHK-10 kood $E10 - E14$ koos alamjaotise täpsustusega) eraldi diagnoositud olla. Patsientidele on genereeritud vähemalt 2 diagnoosi, mis on selgelt diabeediga seotud.

   ChatGPT mudelile esitatud sisend: „*Make an example ICD-10 dataset of diabetic patients with related diagnoses.*“

2. Kümme patsienti, kellele on kõigile määratud üks peamine haigus, mis kolme patsiendi jaoks on diabeet, kahe jaoks müokardiinfarkt (RHK-10 kood $I21$ või

[1] ChatGPT on OpenAI poolt loodud keelemudel, mis on treenitud paljudel erinevatel tekstiallikatel. Lisateave: `https://openai.com`. Mudelide esitatud päringud tehtud 26.04.2023.

$I22$ koos täpsustava alamjaotisega), kolmel mingi psüühika- või käitumishäire (RHK-10 alamjaotis peatükist $F00 - F99$) ning kahel midagi muu diagnoos. Kõigil patsientidel on 2 kuni 5 sellega seonduvat RHK-10 diagnoosi veel lisaks.

ChatGPT mudeli sisend: „*Generate* 10 *patients with one main ICD-10 diagnosis (*3 *have some form of diabetes,* 2 *have myocardial infarction,* 2 *have F category diagnoses and* 3 *have something else), also add* $2 - 5$ *related diagnoses.*“

Autori valitud koodide kolmikud ja ChatGPT abiga loodud diagnoosi koodide tabelid koos nende peal rakendatud algoritmide tulemustega on peatükis 4.

2 Diagnooside sarnasus

Selleks, et patsiente omavahel võrrelda, on vaja defineerida sarnasused diagnooside tasemel. Diagnooside kategoriseerimiseks kasutame RHK-10 koode.

Et leida parim algoritm kahe diagnoosi võrdlemiseks, katsetame läbi mitu erinevat varianti ja meetodit. Järgnevates alampeatükkides on need välja toodud.

2.1 Koodide kategooriline sarnasus

Kahe diagnoosikoodi a ja b peamine sarnasuse tegur on ilmselt nende sisuline sarnasus. Selle leidmiseks kasutame kahte erinevat versiooni - diagnooside RHK-10 hierarhilise kauguse kaudu ning diagnooside tekstilise esituse konteksti kaudu. Neist esimene on kirjeldatud alampeatükis 2.1.1 ja teine alampeatükis 2.1.2.

2.1.1 RHK-10 taseme sarnasus

Selle alampeatüki tulemus tugineb artiklile [17].

Selleks, et ära defineerida kahe diagnoosi a ja b vaheline RHK-10 hierarhiline sarnasusfunktsioon $S_{RHK10}(a,b)$, läheb vaja kahte abifunktsiooni.

Definitsioon 1. Olgu meil RHK-10 klassifikaator x, mis on kas peatükk, alampeatükk, jaotis või alamjaotis. Defineerime klassifikaatori x taseme $L(x)$ RHK-10 süsteemis järgmiselt

$$L(x) := \begin{cases} 1, & \text{kui } x \text{ on peatükk,} \\ 2, & \text{kui } x \text{ on alampeatükk,} \\ 3, & \text{kui } x \text{ on jaotis,} \\ 4, & \text{kui } x \text{ on alamjaotis.} \end{cases}$$

Näide 1. Alampeatüki $G20$-$G26$ hierarhiline tase on

$$L(G20\text{-}G26) = 2,$$

sest temast kõrgemaid tasemeid on vaid üks, ehk peatükk VI : $G00$-$G99$.

Diagnoosi $I20.1$ hierarhiline tase on

$$L(I20.1) = 4,$$

sest temast kõrgemad tasemed on jaotis $I20$, alampeatükk $I20$-$I25$ ning peatükk XI : $I00$-$I99$

Definitsioon 2. Olgu meil RHK-10 koodid x ja y. Olgu RHK-10 kõikide klassifikaatorite hulk $\mathbf{R}$. Koodide x ja y lähimaks ühiseks klassifikaatoriks (*nearest common ancestor*) $NCA(x,y)$ nimetame klassifikaatorit $A \in \mathbf{R}$, mille alla kuuluvad nii x kui ka y, kusjuures ei leidu kitsamat klassifikaatorit $B \in \mathbf{R} : B \subset A$, kuhu mõlemad koodid kuuluvad. Juhul kui ei leidu sellist klassifikaatorit $A \in \mathbf{R}$, kuhu kuuluksid nii x kui y, siis $NCA(x,y) = \emptyset$.

Näide 2. Diagnooside $E10.0$ ja $E12.0$ lähim ühine klassifikaator on

$$NCA(E10.0, E12.0) = E10\text{-}E14,$$

kuna mõlemad on alampeatüki $E10$-$E14$ alamjaotised, aga ühine jaotis neil puudub, sest $E10.0$ kuulub jaotisesse $E10$ ning $E12.0$ jaotisesse $E12$.

Diagnooside $E10.0$ ja $A01.0$ lähim ühine klassifikaator on

$$NCA(E10.0, A01.0) = \emptyset,$$

kuna $E10.0$ kuulub peatükki IV: $E00-E90$, aga $A01.0$ kuulub peatükki I: $A00-B99$ ning laiemat klassifikaatorit kui peatükk RHK-10 süsteemis pole.

Definitsioon 3. Olgu meil 2 RHK-10 diagnoosi a ja b. Siis nende kahe diagnoosi kategooriline sarnasus $S_{RHK10}(a,b)$ on esitatav võrrandiga

$$S_{RHK10}(a,b) := \begin{cases} \frac{2 \cdot L(NCA(a,b))}{L(a)+L(b)} & \text{, kui } NCA(a,b) \neq \emptyset \\ 0 & \text{kui } NCA(a,b) = \emptyset \end{cases}.$$

Näide 3. koodide $E10.0$ ja $E12.0$ kategooriline sarnasus on

$$\begin{aligned} S_{RHK10}(E10.0, E12.0) &= \frac{2 \cdot L(NCA(E10.0, E12.0))}{L(E10.0) + L(E12.0)} \\ &= \frac{2 \cdot L(E10\text{-}E14)}{L(E10.0) + L(E12.0)} \\ &= \frac{2 \cdot 2}{4+4} \\ &= \frac{4}{8} \\ &= \frac{1}{2}. \end{aligned}$$

Lause 1. *Kahe diagnoosi kategoorilise sarnasuse $S_{RHK10}(a,b)$ jaoks kehtivad järgmised väited:*

1. $S_{RHK10}(a,b) = S_{RHK10}(b,a)$,
2. $\max\limits_{a,b \in \text{RHK-10}} S_{RHK10}(a,b) = 1$,
3. $\min\limits_{a,b \in \text{RHK-10}} S_{RHK10}(a,b) = 0$,
4. $S_{RHK10}(a,b) = 1 \Leftrightarrow a = b$.

2.1.2 Kontekstipõhine sarnasus

See alampeatükk kasutab artiklis [18] kirjeldatud loomuliku keele töötluse algoritmi *word2vec*, mille abil defineeritakse käesolevas töös uus diagnooside sarnasuse mudel.

Koodide kontekstipõhiseks võrdlemiseks on vaja esmalt leida diagnooside kirjeldused. RHK-10 süsteemis vastab igale koodile teksti kujul kirjeldus. Edaspidi tähistame koodi $a \in$ RHK-10 tekstilist kirjapilti kui a_t.

Näide 4. Koodi $A15.1$ eestikeelne tekstiline vaste RHK-10 süsteemis on „kopsutuberkuloos, kinnitatud ainult kultuuriga“.

Tekstide omavaheliseks võrdlemiseks oleks vaja need vektoriseerida. Teeme seda sõnade kaupa *word2vec* (edaspidi w2v) algoritmi abil. *Word2vec* sai valitud sellepärast, et antud algoritmiga saab leida nii süntaksi poolest kui semantiliselt sarnaseid sõnu - mida meditsiiniliste diagnooside hulgas on palju.

W2v algoritm vajab treenimiseks sõnade korpust, mille pealt tekitab vektorruumi, kus igale unikaalsele sõnale määratakse vektor. Tulemuste analüüsimisel kasutame töös eeltreenitud w2v mudelit EVEX [19] andmebaasist[2], mis on treenitud meditsiiniliste tekstide peal.

W2v algoritmid ei suuda leida sarnasust korpusevälistele sõnadele. Kasutatud RHK-10 diagnooside hulgas tekitas see probleemi 19 erineva diagnoosi jaoks (kokku on kasutusel 7298 diagnoosi). Neist 17 juhul sai diagnoosi asendada alternatiivse

[2]Kasutatud mudel on wikipedia-pubmed-and-PMC-w2v.bin, mis on kättesaadav veebist: `http://evexdb.org/pmresources/vec-space-models/`. Andmestik on alla laetud 12.07.2022.

kirjapildiga, mille sõnad olid korpuses olemas. Kahel juhul, koodidele $L68.3$ ja $J66.2$, oli vaja asendus teha vastavalt jaotise $L68$ ja $J66$ kirjeldusega, kuna neil alternatiivne kirjapilt puudus. Samuti asendasime viie diagnoosi nime, mille esmane versioon esines korpuses liiga harva, et piisavaid seoseid luua. Kõik need diagnoosid koos nende kirjapildi muudatustega on kirjas lisas A.

Pärast sõnade vektorkujule kirjutamist on neid omavahel võimalik võrrelda. Selle jaoks leiame vektoritevahelise koosinussarnasuse.

Definitsioon 4. [20, lk 77-78] Olgu **A** ja **B** n-dimensionaalsed vektorid. Nendevaheliseks koosinussarnasuseks nimetatakse funktsiooni

$$S_C(\mathbf{A},\mathbf{B}) := \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\,\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

kus A_i ja B_i on vastavalt vektorite **A** ja **B** komponendid kohal i.

Enamik diagnooside kirjeldusi on pikemad kui üks sõna, seega diagnooside peal otse definitsiooni 4 rakendada ei saa. Selleks, et koosinussarnasust rakendada ka lausete jaoks, peame leidma vektorid diagnooside kogukirjelduse jaoks.

Definitsioon 5. Olgu meil RHK-10 koodi a tekstiline kirjapilt a_t. Sellisel juhul leidub meil vektorite list $[\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^n]$, kus iga vektor $\mathbf{A}^j, j \in \{1, 2, \ldots, n\}$ vastab a_t mingile sõnale, ehk n on a_t sõnade arv. Defineerime vektorite keskmise

$$\bar{\mathbf{A}} := \{\bar{A}_1, \bar{A}_2 \ldots \bar{A}_k\},$$

kus $\bar{A}_i, i \in \{1, 2, \ldots, k\}$ on vektorite $[\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^n]$ liikmete $A_i^1, A_i^2, \ldots A_i^n$ aritmeetiline keskmine ja k on w2v mudeli vektorruumi vektorite pikkus.

Järeldus 1. *Paneme tähele, et kui koodi a kirjapilt a_t on ühesõnaline, siis talle vastav vektorite list on $[\mathbf{A}^1]$, ehk $\bar{A}_i = A_i^1$ iga $i \in \{1, \ldots, k\}$ korral, mistõttu ka $\bar{\mathbf{A}} = \mathbf{A}^1$.*

Nüüd saamegi ära defineerida kahe koodi kontekstipõhise sarnasuse.

Definitsioon 6. Olgu meil RHK-10 koodid a ja b ning nende tekstiliste kirjapiltide keskmised vektorid $\bar{\mathbf{A}}$ ja $\bar{\mathbf{B}}$. Defineerime koodide a ja b kontekstilise sarnasuse $S_{kontekst}(a,b)$:

$$S_{kontekst}(a,b) := \max\left(0, S_C(\bar{\mathbf{A}},\bar{\mathbf{B}})\right) = \frac{\max(0, \bar{\mathbf{A}}\cdot\bar{\mathbf{B}})}{\left\|\bar{\mathbf{A}}\right\|\left\|\bar{\mathbf{B}}\right\|}.$$

Lause 2. *Kahe diagnoosi kontekstipõhise sarnasuse $S_{kontekst}(a,b)$ jaoks kehtivad järgmised väited:*

1. $S_{kontekst}(a,b) = S_{kontekst}(b,a)$,
2. $\max\limits_{a,b \in \text{RHK-10}} S_{kontekst}(a,b) = 1$,
3. $\min\limits_{a,b \in \text{RHK-10}} S_{kontekst}(a,b) = 0$,
4. $a = b \Rightarrow S_{kontekst}(a,b) = 1$,
5. $a_t = b_t \Leftrightarrow S_{kontekst}(a,b) = 1$.

2.1.3 Koodide kategoorilise sarnasuse algoritmide võrdlus

Siin alampeatükis võrdleme omavahel RHK-10 taseme kaudu leitud sarnasust S_{RHK10} ja koodide sisulise konteksti kaudu leitud sarnasust $S_{kontekst}$.

Peamine erinevus on selles, et kontekstipõhine sarnasus on dünaamilisem ja kirjeldab paremini eelkõige RHK-10 süsteemi jaotiste omavahelist seost. Nimelt mõnes jaotises on diagnoosid üksteise suhtes sarnasemad, mõnes on seos diagnooside vahel väga väike.

Näide 5. Jaotise $E10$ (insuliinisõltuv suhkurtõbi) alamjaotised on omavahel väga sarnased:

$E10.0$ - „Insuliinisõltuv suhkurtõbi koomaga“,

$E10.1$ - „Insuliinisõltuv suhkurtõbi ketoatsidoosiga“,

$E10.2$ - „Insuliinisõltuv suhkurtõbi neerutüsistustega“,

$E10.3$ - „Insuliinisõltuv suhkurtõbi silmatüsistustega“,

$E10.4$ - „Insuliinisõltuv suhkurtõbi neuroloogiliste tüsistustega“,

$E10.5$ - „Insuliinisõltuv suhkurtõbi perifeersete vereringetüsistustega“,

$E10.6$ - „Insuliinisõltuv suhkurtõbi muude täpsustatud tüsistustega“,

$E10.7$ - „Insuliinisõltuv suhkurtõbi hulgitüsistustega“,

$E10.8$ - „Insuliinisõltuv suhkurtõbi täpsustamata tüsistustega“,

$E10.9$ - „Insuliinisõltuv suhkurtõbi tüsistusteta“.

Jaotise $E34$ - (muud sisesekretsioonihäired) alamjaotised on omavahel suurema erinevusega:

$E34.0$ - „Kartsinoidsündroom“

$E34.1$ - „Soolehormoonide muu liignõristus“

$E34.2$ - „Mujal klassifitseerimata ektoopiline hormooninõristus“

$E34.3$ - „Mujal klassifitseerimata vaegkasv“

$E34.4$ - „Konstitutsionaalne liigkasv“

$E34.5$ - „Androgeenresistentsuse sündroom“

$E34.8$ - „Muud täpsustatud endokriinsed haigusseisundid“

$E34.9$ - „Täpsustamata endokriinne haigusseisund“

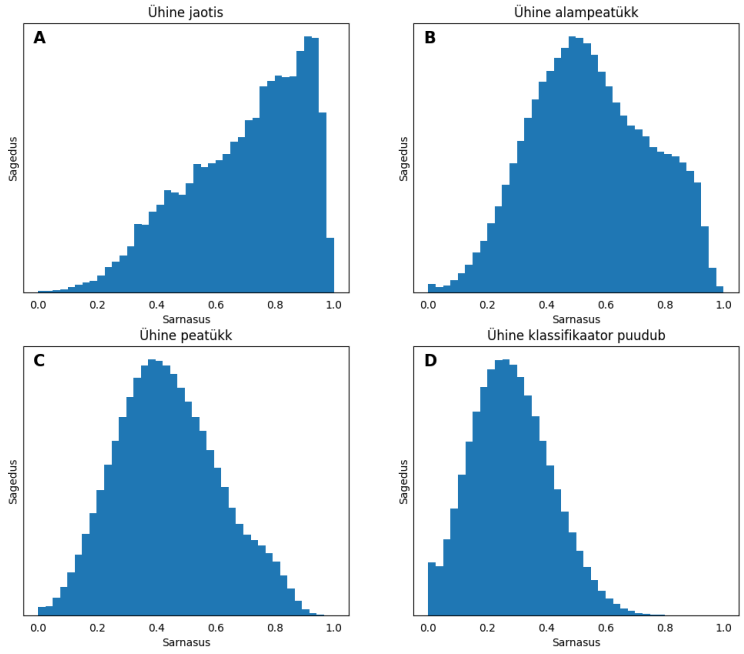
Jaotise $E10$ ja $E34$ erinevate alamjaotiste keskmine sarnasus RHK-10 taseme järgi on mõlemal juhul $0{,}75$, aga kontektsi põhine keskmine sarnasus on vastavalt $0{,}9500$ ja $0{,}3133$.

Tabelist 1 näeme, et suures plaanis on ühise jaotise ja ühise alampeatüki sarnasused võrreldavad, aga peatüki jaoks leiab kontekstipõhine mudel $S_{kontekst}$ tugevamad seosed diagnooside vahel kui RHK-10 taseme sarnasuse mudel S_{RHK10}. Suurim erinevus on diagnooside vahel, mis kuuluvad täiesti erinevatesse peatükkidesse. Nimelt sellisel juhul S_{RHK10} väärtuseks on alati 0, aga $S_{kontekst}$ väärtus keskmiselt $0{,}2818$.

	Alamjaotis	Jaotis	Alampeatükk	Peatükk	NCA puudub
S_{RHK-10}	1	0,75	0,5034	0,2522	0
$S_{kontekst}$	1	0,6978	0,5498	0,4417	0,2818
Erinevus	0	0,0522	0,0464	0,1895	0,2818

Tabel 1: Sarnasuste S_{RHK10} ja $S_{kontekst}$ võrdlemise jaoks loodud tabel, kus tunnusteks on kahe diagnoosi lähim ühine klassifikaator ehk NCA ning ridades vastavalt RHK-10 tasemel sarnasuse keskmine tulemus.

Jooniselt 1 on näha, et vaatamata S_{RHK10} ja $S_{kontekst}$ jaotise ja alampeatükkide keskmiste sarnasusele, on variatsioon sarnasuse $S_{kontekst}$ siseselt küllaltki suur. Ühise RHK-10 klassifikaatori puudumisel on diagnoosipaaride sarnasus ligikaudselt normaaljaotusest. Sellest lähtuvalt võib öelda, et sarnasuse $S_{kontekst}$ baasväärtus on ligikaudu $0{,}28$.



Joonis 1: Diagnooside sarnasuse histogrammid, kus joonisel A on iga jaotise jaoks leitud kõikide selle jaotise alamjaotiste omavahelised sarnasused, joonisel B iga alampeatüki selliste alamjaotiste sarnasus, mis pole ühisest jaotisest ning joonisel C analooglselt peatüki tasemel sarnasused. Joonisel D on diagnoosipaaride sarnasused, millel pole ühist RHK-10 klassifikaatorit.

2.2 Muud diagnoosi tunnused

Eesmärgiks oli luua algoritm, mis võtab arvesse ka muid diagnooside iseärasusi. Raamatus [1] on mainitud, et RHK-10 on välja arnenenud praktilise klassifitsioonina, võttes peamiselt arvesse haiguste tekitajaid, paiknemisi ning tekketingimusi. Vähem on klassifitseerimisel arvestatud näiteks haigustega seotud suremust ning haigestumust.

Diagnooside omavahel võrdlemisel võivad sellised tegurid teatud rolli mängida. Selle tõttu kirjeldame järgnevates peatükkides haiguse tunnuseid, mille võtame algoritmi loomisel arvesse.

2.2.1 Koodide haruldus

Selle peatüki arvulised väärtused põhinevad peatükis 1.2 kirjeldatud andmetel.

On võimalik, et leidub kaks paari diagnoose, mis on RHK-10 hierarhias omavahel sama kaugel, kuid üks paar on esinemise sageduse poolest enam-vähem võrdne, teine mitte.

Näide 6. Diagnoosi paaride ($E10$,$E11$) ja ($C53, C56$) kategoorilised sarnasused on vastavalt

$$\begin{aligned} S_{RHK10}(E10,E11) &= \frac{2 \cdot L(NCA(E10,E11))}{L(E10) + L(E11)} = \frac{2 \cdot L(E10 - E14)}{L(E10) + L(E11)} \\ &= \frac{2 \cdot 2}{3+3} = \frac{2}{3} \end{aligned}$$

ja

$$\begin{aligned} S_{RHK10}(C53, C56) &= \frac{2 \cdot L(NCA(C53, C56))}{L(C53) + L(C56)} = \frac{2 \cdot L(C51 - C58)}{L(C53) + L(C56)} \\ &= \frac{2 \cdot 2}{3+3} = \frac{2}{3}. \end{aligned}$$

Samas aastatel 2016-2020 sai esmakordselt insuliinisõltuva suhkurtõve diagnoosi ($E10$) 1027 inimest, insuliinisõltumatu suhkurtõve ($E11$) 24043 inimest, emakakaela pahaloomulise kasvaja ($C53$) 706 inimest ja munasarja pahaloomulise kasvaja ($C56$) 684 inimest.

Intuitiivselt võiks näites toodud paari ($C53, C56$) pidada omavahel sarnasemaks kui paari ($E10$,$E11$), kuna esinemissageduse mõttes on esimese paari diagnoosid peaaegu võrdsed, aga $E11$ esineb tunduvalt sagedamalt kui $E10$.

Selleks, et seda algoritmiliselt leida, peab igale diagnoosile määrama esinemise harulduse. Kasutame peatükis 1.2 kirjeldatud mudelit, et genereerida 20000 patsienti ning leiame iga haiguse levimuse antud populatsioonis.

Olgu a RHK-10 kood, tähistame haiguse esinemise sageduse kui $F(a)$.

Genereeritud andmetes on sagedused määratud kuni alampeatükkideni, seega saame kuni selle tasemeni esinemissagedused. Alampeatükkide $A00 - A09$, $A15 - A19$, $B15-B19$, $C00-C14$, $C15-C26$, $C30-C39$, $C43-C44$, $C51-C58$, $C60-C63$, $C64-C68$, $C69-C72$, $C73-C75$, $C81-C96$, $E10-E14$, $F00-F09$, $F20-F29$, $F30 - F29$, $F40 - F49$, $F50 - F59$, $F60 - F69$, $F80 - F89$, $F90 - F98$ ning $I20 - I25$ jaotiste esinemissageduse saame tuletada andmetest, mis on kirjeldatud peatükis 1.2.1.

Ülejäänud alampeatükkide jaotistele määrame esinemissageduseks vastava alampeatüki esinemissageduse.

Seejärel leiame haiguse osakaalu patsientide hulgas, jagades iga koodi a esinemissageduse $F(a)$ patsientide koguarvuga. Harulduse saame, kui lahutame selle väärtuse ühest.

Definitsioon 7. Olgu meil RHK-10 kood a. Tähistame koodi a harulduse funktsiooniga $H(a)$:

$$H(a) := 1 - \frac{F(a)}{N},$$

kus $F(a)$ on diagnoosi a esinemiste arv $N \in \mathbb{N}$ patsiendi hulgas.

Edasised haruldused on leitud nii, et patsientide arv $N = 20000$. Nii leitud harulduste korral on näites 6 toodud diagnooside haruldused:

$$H(E10) = 0{,}972,$$
$$H(E11) = 0{,}355,$$
$$H(C53) = 0{,}961,$$
$$H(C56) = 0{,}964.$$

2.2.2 Koodidega seotud suremus

Sarnaselt eelmisele alampeatükile ei ole RHK-10 klassifikatsiooni eesmärk ka diagnoose nende tõsiduse põhjal grupeerida. Autori hinnangul võiks tõenäosus, et diagnoos põhjustab patsiendi surma, mõjutada kahe diagnoosi vahelist sarnasust. Selle mõõtmine on siin alampeatükis ära kirjeldatud ja tulevatesse algoritmidesse kaasatud.

Selleks, et määrata diagnoosiga seotud suremust, kasutame Statistikaameti andmestikku RV56: Surnud Surmapõhjuse, soo ja vanuserühma järgi [21]. Suremuse analüüsiks võtame andmestikust ette andmed aastatest 2000 kuni 2022 ning lihtsustamise eesmärgil arvestame meeste ja naiste suremuse kokku ning vanuserühmade eristust samuti ei tee. Nii tehes saame ülevaate, kui palju on kindlas ajavahemikus igat diagnoosi kokku määratud surma põhjuseks.

Suremused on RV56 andmetes määratletud erinevatel tasemetel:

1. kõik põhjused kokku, $A00$ - $Y89$,
2. peatüki tasemel, näiteks $A00$ - $B99$ (nakkus- ja parasiithaigused) ja $L00$ - $L99$ (naha- ja nahaaluskoe haigused),
3. alampeatüki või muu sarnase klassifitseerimise tasemel, näiteks $B20$ - $B24$ (HIV-tõbi) aga ka $F11 - F16$, $F18$, $F19$ (uimastisõltuvus ja toksikomaania),
4. jaotise tasemel, näiteks $C53$ (emakakaela pahaloomuline kasvaja) ja $K70$ (maksa alkoholtõbi).

Paljude jaotiste kohta eraldi täpsustus puudub, mistõttu tuleb nendega seotud suremus andmetest tuletada.

Tähistame tabeli RV56 andmetes leiduvadate tasemete hulga $\mathbf{D}$ ning iga tasemega $X \in \mathbf{D}$ seotud surmade arvu kui $s(X)$.

Definitsioon 8. Olgu a RHK-10 jaotis. Olgu $A \in \mathbf{D}$ väikseim klassifikatsioon, kuhu a kuulub. Tähistame A alamhulkade kogumit $\{A_1, A_2, \dots A_n : A_i \in \mathbf{D}, A_i \subset A\}$ tähisega A'. See tähendab, et $\forall A_i \in A' : a \notin A_i$. Sellisel juhul jaotisega a seotud surmade arvu $s(a)$ leiame valemiga

$$s(a) = \frac{s(A) - \sum\limits_{A_i \in A'} s(A_i)}{|\{x : x \in A \ \wedge \ \forall A_i \in A', x \notin A_i\}|}.$$

Näide 7. Aastatel 2000-2022 suri juhuslikku mürgitusse $A = (X40$ - $X49)$ kokku 5340 inimest, ehk

$$s(A) = 5340.$$

Nendest täpsustus on ainult alkoholimürgitusele $A_1 = (X45)$, millesse suri 3231 inimest, ehk

$$s(A_1) = 3231 \text{ ja } A' = \{A_1\}.$$

Seega diagnoosidesse $X40$ - $X44$ ja $X46$ - $X49$ suri kokku 2109 inimest, ehk

$$s(A) - s(A_1) = 5340 - 3231 = 2109.$$

Saame iga diagnoosi

$$a_i \in \{X40, X41, X42, X43, X44, X46, X47, X48, X49\}$$

jaoks surmade arvu:

$$\begin{aligned} s(a_i) &= \frac{s(A) - s(A_1)}{|\{x : x \in A \ \wedge \ x \notin A_1\}|} \\ &= \frac{s(A) - s(A_1)}{|\{X40, X41, X42, X43, X44, X46, X47, X48, X49\}|} \\ &= \frac{2109}{9} \\ &= 234\frac{1}{3}. \end{aligned}$$

Seega määrame igale sellisele diagnoosile $234\frac{1}{3}$ surmajuhtumit.

Näites 7 ilmnenud mitte täisarvuline surmade arv ei määra edaspidi suurt rolli, kuna arvutamisel kasutame skaleeritud suremust.

Samuti ei ole andmestikus täpsustusi alamjaotiste kohta, mille puhul teeme samasuguse suremuste määramise nagu definitsioonis 8.

Ideaalis kasutaksime suremuse koefitsendi määramiseks ära diagnooside esinemissagedust ning suremust, näiteks kui 2000.-2022. aastal sai haiguse a 500 inimest ja samal ajaperioodil suri sellesse 400 inimest, võiks suremuskoefitsent olla $0{,}8$. Nagu eelmises peatükis mainitud, siis sünteetilitest andmetest ei saa piisava kvaliteediga esinemissagedust tuletada, seega tuleb suremus defineerida teisiti. Määrame diagnoosiga a seotud suremuseks $S(a)$ tõenäosuse, et just a on patsiendi surma põhjustajaks. Seejärel skaleerime tõenäosused nii, et 0 vastab diagnoosile, mis on kõige harvem surma põhjustaja ning 1 vastavalt kõige sagedamale surma põhjustajale.

Definitsioon 9. Olgu a RHK-10 kood ning selle surmaga lõppenud juhtumite arv $s(a)$. Kõikide diagnooside kogu surmade arvuks on $s(\mathbf{D})$. Tähistame $P_S(a)$ kui tõenäosuse, et a on surma põhjustaja. Siis $P_S(a)$ avaldis on järgmine:

$$P_S(a) := \frac{s(a)}{s(\mathbf{D})}.$$

Definitsioon 10. Olgu a RHK-10 kood ning tõenäosus, et see on suvalise patsiendi surma põhjustajaks $P_S(a)$. Sellisel juhul koodiga a seotud skaleeritud suremus $S(a)$ on defineeritud järgmiselt:

$$S(a) := \frac{P_S(a) - \min\limits_{x \in \text{RHK-10}} P_S(x)}{\max\limits_{x \in \text{RHK-10}} P_S(x) - \min\limits_{x \in \text{RHK-10}} P_S(x)}.$$

2.2.3 Koodide kroonilisus

See alampeatükk tugineb Ameerika Ühendriikides välja töötatud tervishoiukulude ja -rakendamise projekti HCUP (*Healthcare Cost and Utilization Project*) käigus loodud kroonilise seisundi indikaatoril CCIR (*Chronic Condition Indicator Refined*) [22].

Haigusi liigitatakse reeglina kaheks - kroonilisteks ja ägedateks. HCUP CCIR klassifikatsiooni kasutusõpetuses[3] on defineeritud kroonilised haigused järgmiselt:

Definitsioon 11. Krooniliseks haiguseks nimetatakse diagnoosi, mis kestab vähemalt 12 kuud ning mille korral on täidetud vähemalt üks järgmistest tingimustest:

1. Haigus vajab pidevat meditsiinilist sekkumist.
2. Haigus piirab iseseisvat elu ning sotsiaalset suhtlemist.

Vaatame siin peatükis ka korduvate haiguste arvestamist:

1. Korduv kroonilise haiguse diagnoos.

   Sellise haiguse mitmekordne lugemine ei ole sarnasuse määramisel oluline, kuna kui krooniline haigus on juba kord diagnoositud, siis see üldjuhul enam ei tervistuta.

[3]HCUP CCIR lehelt leitav kasutusjuhend „User Guide: Chronic Condition Indicator Refined (CCIR) for ICD-10-CM, v2023.1“, vaadatud 08.05.2023.

Näiteks ei määra olulist rolli diabeedi korduv diagnoosimine kahe patsiendi sarnasuse hindamisel. See tähendab, et pole mõtet liigitada patsient P_a ja patsient P_b, kes on mõlemad saanud 10+ korda diabeedi diagnoosi, omavahel sarnasemaks kui P_a ja P_c, kui P_c on diabeedi diagnoosi saanud vaid 3 korral.

2. Korduv ägeda haiguse diagnoos.

   Ägedate haiguste mitmekordne diagnoosimine võib määrata rolli kahe patsiendi sarnasuse vahel. Näiteks võiks sarnaseks hinnata kahte patsienti, kellel on korduvalt esinenud $J13$, ehk *Streptococcus pneumoniae* tekkene kopsupõletik, kuna neil võib olla teatud soodumus selle korduvaks tekkeks.

Algoritmi lihtsustamise huvides jätame aga välja kõik korduvad haigused ning käsitleme diagnoosi kroonilisust üksnes kui sarnasuse tegurit. See tähendab, et kui võrrelda mingit kroonilist diagnoosi $x \in$ RHK-10 omavahel küllaltki sarnaste diagnoosidega y,$z \in$ RHK-10, kus vastavalt esimene on krooniline ja teine äge haigus, siis diagnoosiga x sarnasemaks tuleks pidada diagnoosi y. Määrame ära kroonilisuse indikaatori.

Definitsioon 12. Olgu a RHK-10 kood. Defineerime a kroonilisuse $K(a)$ järgmiselt

$$K(a) = \begin{cases} 1, & \text{kui } a \text{ on krooniline haigus,} \\ 0{,}5, & \text{kui } a \text{ kroonilisust ei saa määrata,} \\ 0, & \text{kui } a \text{ on äge haigus.} \end{cases}$$

Selleks, et määrata diagnoosi kroonilisus, võtame abiks HCUP CCIR klassifikaatori tabeli[4]. CCIR tabel määrab igale diagnoosile, kas ta on krooniline (väärtus 1) või mitte (väärtus 0). Erandjuhul ei saa kroonilisust määrata, sellisel juhul on tabelis väärtuseks 9. Arvestame sellistele diagnoosidele väärtuseks $0{,}5$, nii saame neile hiljem määrata sama kauguse kroonilistest ning ägedatest diagnoosidest.

CCIR tabelis on saadaval kroonilisuse informatsioon 2015.-2023. aasta ICD-10-CM diagnooside jaoks. See on täpsem versioon kui RHK-10 süsteem. See tähendab, et CCIR tabelis leiduvad kitsamad klassifikatsioonid kui alamjaotis. Sellest lähtuvalt leidub käesolevas töös analüüsitavaid diagnoose, mida CCIR tabelis sellisel kujul ei leidu. Nendel juhtudel võtame CCIR tabelist vastava koodi all olevad täpsustatud koodid ning leiame nende hulgast kõige sagedamalt esineva kroonilisuse märgendi.

Näide 8. Diagnoos $B27.0$ (gammaherpesviirus-mononukleoos) puudub selles CCIR tabelis, kuid leiduvad diagnoosid $B27.00, B27.01, B27.02, B27.09$, mis on kõik ka

[4]HCUP CCIR lehelt leitav tabel CCIR_v2023-1.csv, alla laetud 08.05.2023.

gammaherpesviirus-mononukleoosid, kuid lisatud on kaasuva haiguse täpsustus. Kõikidel nendel on CCIR tabelis kroonilisuse vaste 1 ehk krooniline haigus, mistõttu määrame ka diagnoosi $B27.0$ krooniliseks, ehk $K(B27.0) = 1$.

Leidub 52 RHK-10 jaotist, mille ühtki täpsustatud klassifikatsiooni ei ole mingil põhjusel CCI tabelis. Nendele määras autor käsitsi kroonilisuse märgendi ning see on leitav lisas B.

2.3 Koodide kombineeritud sarnasus

Selleks, et kahte koodi omavahel võrrelda, tuleb kombineerida varasemalt mainitud tunnused üheks.

Kõige lihtsam viis peatükis 2.2 kirjeldatud kahe diagnoosi tunnuste võrdlemiseks on iga tunnuse jaoks võtta nende kahe diagnoosi tunnuse väärtuste vahe. Nii saame nende erinevuse, ehk sarnasuse leidmiseks lahutame selle tulemuse ühest. Sellisel juhul tagame, et ühe ja sama diagnoosi võrdlemisel saame alati sarnasuseks väärtuse 1.

Definitsioon 13. Olgu meil RHK-10 alamjaotised a ja b. Siis saame

1. harulduse sarnasuseks $H_s(a,b) = 1 - |H(a) - H(b)|$,
2. suremuse sarnasuseks $S_s(a,b) = 1 - |S(a) - S(b)|$,
3. kroonilisuse sarnasuseks $K_s(a,b) = 1 - |K(a) - K(b)|$.

Lause 3. *Kõigi definitsioonis 13 nimetatud sarnasuste maksimaalne väärtus on* 1 *ja minimaalne väärtus on* 0.

Definitsioon 14. Olgu a ja b RHK-10 alamjaotised. Defineerime diagnooside sarnasused $S_1(a,b)$ ja $S_2(a,b)$ järgmiselt:

$$S_1(a,b) := w_1 \cdot S_{RHK10}(a,b) + w_2 \cdot H_s(a,b) + w_3 \cdot S_s(a,b) + w_4 \cdot K_s(a,b), \quad (1)$$
$$S_2(a,b) := w_1 \cdot S_{kontekst}(a,b) + w_2 \cdot H_s(a,b) + w_3 \cdot S_s(a,b) + w_4 \cdot K_s(a,b), \quad (2)$$

kus

$$w_1, w_2, w_3, w_4 \in [0,1] : w_1 + w_2 + w_3 + w_4 = 1$$

.

Lause 4. *Diagnooside sarnasuste $S_1(a,b)$ ja $S_2(a,b)$ jaoks kehtivad järgmised väited:*

1. $\max\limits_{a,b\in\text{RHK-10}} S_1(a,b) = 1$,
2. $\max\limits_{a,b\in\text{RHK-10}} S_2(a,b) = 1$,
3. $S_1(a,b) = 1 \Leftrightarrow a = b$,
4. $a = b \Rightarrow S_2(a,b) = 1$,
5. $S_1(a,b) = S_1(b,a)$,
6. $S_2(a,b) = S_2(b,a)$.

3 Patsientide sarnasus

Siin peatükis defineerime ära algoritmid, millega saab patsientide sarnasust mõõta ning toome välja nende peamised erinevused. Tähistagu edaspidi P_x suvaliselt valitud patsienti. Olgu patsiendile P_x määratud diagnooside hulk

$$\mathbf{D_x} := \left\{dt_{(x,1)}, dt_{(x,2)}, \ldots, dt_{(x,n)}\right\},$$

kus $n \in \mathbb{N}$ tähistab P_x diagnooside koguhulka ning

$$dt_{(x,i)} := (d_{(x,i)}, t_{(x,i)}),$$

kus $d_{(x,i)}$ tähistab koodi RHK-10 süsteemist ning $t_{(x,i)}$ koodi $d_{(x,i)}$ määramise hetkel patsiendi P_x vanust (aastates).

Olgu edaspidi

$$\mathbf{P} := \{P_1, P_2, \ldots, P_k\}$$

kõikide patsientide hulk, kus $k \in \mathbb{N}$ tähistab, kui palju patsiente on valimis.

3.1 Patsiendi vanus diagnoosi määramise hetkel

Eesmärgiks on luua algoritm, mis hindab patsientide sarnasust üleüldiselt, mitte ühe haiglavisiidi siseselt. Sellest tulenevalt võiks teguriks olla ka diagnooside saamise vanus. Näiteks kui võrrelda 20-aastaselt infarkti saanud patsienti vastavalt 24- ja 89-aastaselt infarkti saanud patsiendiga, siis võiks sarnasemaks pidada 20- ja 24-aastaselt infarkti diagnoosi saanud patsiente.

Ilmselt mängib osade haiguste (eelkõige nakkushaiguste) sarnasuse määramise puhul rolli ka absoluutne põdemise aeg. Näiteks võiks sarnaseks lugeda 2018. aastal grippi põdenud patsiendid. Algoritmide lihtsustamise huvides on see tegur välja jäetud ning arvestame ainult patsientide suhtelist vanust põdemise hetkel.

Olgu meil patsiendid P_u ja P_v ning neile määratud suvaliselt valitud RHK-10 koodid vastavalt $d_{(u,a)}$ ja $d_{(v,b)}$. Sellisel juhul on nende diagnoosimise hetkel patsientide vanused vastavalt $t_{(u,a)}$ ja $t_{(v,b)}$. Defineerime diagnooside $d_{(u,a)}$ ja $d_{(v,b)}$ põdemise hetke vanuste sarnasuse kui $T_s(t_{(u,a)}, t_{(v,b)})$.

Vanuste võrdlemise funktsioon $T_s(t_{(u,a)}, t_{(v,b)})$ peab täitma kahte intuitiivset tingimust:

1. Vanuste vahe erinevus sõltub absoluutsest vanusest. Näiteks 10- ja 15-aastaselt diabeedi saanud patsiendid on vähem sarnased, kui 75- ja 80- aastaselt diabeedi diagnoosi saanud patsiendid.

2. Vanuste vahe erinevus ei tohi olla ühe patsiendi suhtes lineaarne. Näiteks 10-aastane patsient ei ole 2 korda sarnasem 20-aastasega kui 40-aastasega. 10- ja 20-aastase patsiendi sarnasus peaks olema suurem kui kahekordne.

Defineerime esmalt funktsiooni $T_s^1(t_{(u,a)}, t_{(v,b)})$, mis täidaks esimest tingimust:

$$T_s^1(t_{(u,a)}, t_{(v,b)}) := \frac{\min(t_{(u,a)}, t_{(v,b)})}{\max(t_{(u,a)}, t_{(v,b)})}. \tag{3}$$

Lause 5. *Funktsioon $T_s^1(t_{(u,a)}, t_{(v,b)})$ täidab esimest tingimust, aga mitte teist.*

Tõestus. Tõestame esmalt, et funktsioon $T_s^1(t_{(u,a)}, t_{(v,b)})$ ei täida teist tingimust.

Olgu meil patsiendid P_x, P_y ja P_z, kes kõik on põdenud sama diagnoosi vastavalt $t_{(x,a)} = A$, $t_{(y,a)} = 2A$ ja $t_{(z,a)} = 4A$ aasta vanuselt, kus $A \in \mathbb{N} \setminus \{0\}$. Näitame, et

$$T_s^1(t_{(x,a)}, t_{(y,a)}) = 2 \cdot T_s^1(t_{(x,a)}, t_{(z,a)}).$$

Kasutades defineeritud võrrandit (3), saame et

$$T_s^1(t_{(x,a)}, t_{(y,a)}) = \frac{A}{2A} = \frac{1}{2}$$

ja

$$T_s^1(t_{(x,a)}, t_{(z,a)}) = \frac{A}{4A} = \frac{1}{4},$$

ehk

$$T_s^1(t_{(x,a)}, t_{(y,a)}) = \frac{1}{2} = 2 \cdot \frac{1}{4} = 2 \cdot T_s^1(t_{(x,a)}, t_{(z,a)}).$$

Näitame, et esimene tingimus kehtib. Olgu meil patsiendid P_x, P_y, P_z ja P_w. Kehtigu, et kõik on põdenud sama diagnoosi vastavalt $t_{(x,a)} = A$, $t_{(y,a)} = A + I$, $t_{(z,a)} = B$ ja $t_{(w,a)} = B + I$ vanuselt, kus $A, B, I \in \mathbb{N} \setminus \{0\}$. Tingimus on täidetud, kui tingimusel $A > B$ kehtib

$$T_s^1(t_{(x,a)}, t_{(y,a)}) < T_s^1(t_{(z,a)}, t_{(w,a)}).$$

Teame, et

$$T_s^1(t_{(x,a)}, t_{(y,a)}) = \frac{A}{A + I}$$

ja

$$T_s^1(t_{(y,a)}, t_{(w,a)}) = \frac{B}{B+I},$$

ehk vaja on näidata, et

$$\frac{A}{A+I} < \frac{B}{B+I}.$$

Seega piisab, kui näidata järgmise võrratuse kehtivust:

$$\frac{B}{B+I} - \frac{A}{A+I} > 0.$$

Näitame seda:

$$\frac{B}{B+I} - \frac{A}{A+I} = \frac{B(A+I) - A(B+I)}{(B+1)(A+1)}.$$

On ilmne, et $(B+1)(A+1) > 0$, seega on vaja veel näidata, et

$$B(A+I) - A(B+I) > 0.$$

$$B(A+I) - A(B+I) = BA + BI - AB - AI = BI - AI = I(B-A) > 0,$$

kuna $I > 0$ ja $B > A$, ehk $B - A > 0$. □

Defineerime funktsiooni $T_s^2(t_{(u,a)}, t_{(v,b)})$, mis täidaks teist tingimust:

$$T_s^2(t_{(u,a)}, t_{(v,b)}) := \begin{cases} \frac{1+\cos(0{,}1|t_{(u,a)} - t_{(v,b)}|)}{2}, & \text{kui } |t_{(u,a)} - t_{(v,b)}| \leq 10\pi \\ 0 & \text{mujal} \end{cases}. \quad (4)$$

Lause 6. *Funktsioon $T_s^2(t_{(u,a)}, t_{(v,b)})$ täidab teist tingimust, aga mitte esimest.*

Tõestus. On ilmne, et funktsioon $T_s^2(t_{(u,a)}, t_{(v,b)})$ ei täida esimest tingimust, kuna samade vanusevahede korral on funktsiooni väärtused võrdsed.

Teise tingimuse täitmiseks piisab näidata, et $T_s^2(10, 20) \neq 2 \cdot T_s^2(10, 40)$. See kehtib, sest võrrandist (4) saame, et

$$T_s^2(10, 20) = \frac{1 + \cos(0{,}1 \cdot 10)}{2} \approx 0{,}770$$

ja

$$T_s^2(10, 40) = \frac{1 + \cos(0{,}1 \cdot 30)}{2} \approx 0{,}005.$$

□

Definitsioon 15. Defineerime põdemise hetke vanuste sarnasuse $T_s(t_{(u,a)}, t_{(v,b)})$ kui $T_s^1(t_{(u,a)}, t_{(v,b)})$ ja $T_s^2(t_{(u,a)}, t_{(v,b)})$ kaalutud keskmisena, kuna sellisel juhul on täidetud mõlemad kriteeriumid:

$$T_s(t_{(u,a)}, t_{(v,b)}) := m_1 \cdot T_s^1(t_{(u,a)}, t_{(v,b)}) + m_2 \cdot T_s^2(t_{(u,a)}, t_{(v,b)}),$$

kus m_1, m_2 on valitud konstandid selliselt, et $m_1, m_2 \in [0{,}1]$ ja $m_1 + m_2 = 1$.

Intuitiivselt on esimese tingimuse täitmine olulisem, seega edaspidistes näidetes valime $m_1 = 0{,}75$ ja $m_2 = 0{,}25$.

Definitsioon 16. Defineerime ajast sõltuvad koodide $d_{(x,i)}$ ja $d_{(y,j)}$ kombineeritud sarnasused $S_1^t(dt_{(x,i)}, dt_{(y,j)})$ ning $S_2^t(dt_{(x,i)}, dt_{(y,j)})$ järgmiselt:

$$S_1^t(dt_{(x,i)}, dt_{(y,j)}) := W_1 \cdot S_1(d_{(x,i)}, d_{(y,j)}) + W_2 \cdot T_s(t_{(x,i)}, t_{(y,j)}),$$
$$S_2^t(dt_{(x,i)}, dt_{(y,j)}) := W_1 \cdot S_2(d_{(x,i)}, d_{(y,j)}) + W_2 \cdot T_s(t_{(x,i)}, t_{(y,j)}),$$

kus W_1 määrab koodide tasemel sarnasuse kaalu ja W_2 ajalise sarnasuse kaalu nii, et $W_1, W_2 \in [0{,}1]$ ja $W_1 + W_2 = 1$.

3.2 Patsientide sarnasuse algoritmid

Defineerime kaks algoritmi, mis võrdlevad omavahel patsientide paare. Esimene algoritm tugineb artiklile [23] ning teine artiklile [24].

3.2.1 Suhtelise sarnasuse algoritm

Defineerime esmalt algoritmi, mille eesmärgiks on leida patsiendi P_u jaoks patsient hulgast $\mathbf{P} \setminus P_u$ nii, et igale patsiendi P_u diagnoosile vastab võimalikult sarnane diagnoos koos diagnoosimise ajaga valitud patsiendi diagnooside hulgast. Selleks leiame iga patsiendi

$$P_v \in \mathbf{P} \setminus P_u$$

korral igale patsiendi P_u diagnoosile $dt_{(u,i)}$ kõige sarnasema diagnoosi

$$dt_{(v,j)} \in P_v,$$

liidame kõik tulemused kokku ning jagame P_u diagnooside arvuga läbi. Mida lähemale tuleb sellise algoritmi väärtus arvule 1, seda sarnasem on patsient P_v patsiendile P_u.

Definitsioon 17. Olgu meil patsiendid $P_u \in \mathbf{P}$ diagnoosidega

$$\mathbf{D_u} = \left\{dt_{(u,1)}, dt_{(u,2)}, \ldots, dt_{(u,n)}\right\}$$

ja $P_v \in \mathbf{P}$ diagnoosidega

$$\mathbf{D_v} = \left\{dt_{(v,1)}, dt_{(v,2)}, \ldots, dt_{(v,m)}\right\}.$$

Defineerime patsiendi P_u sarnasuse patsiendiga P_v tähisega $S_{(1,1)}(P_u, P_v)$ ning võrrandiga

$$S_{(1,1)}(P_u, P_v) := \frac{1}{|\mathbf{D_u}|} \sum_{i=1}^{n} \max_{j \in \{1,2,\ldots,m\}} S_1^t(dt_{(u,i)}, dt_{(v,j)}),$$

kus $S_1^t(dt_{(u,i)}, dt_{(v,j)})$ on defineeritud definitsioonis 14 võrrandiga (1). Analoogselt defineerime $S_{(1,2)}(P_u, P_v)$:

$$S_{(1,2)}(P_u, P_v) := \frac{1}{|\mathbf{D_u}|} \sum_{i=1}^{n} \max_{j \in \{1,2,\ldots,m\}} S_2^t(dt_{(u,i)}, dt_{(v,j)}),$$

kus $S_2^t(dt_{(u,i)}, dt_{(v,j)})$ on defineeritud definitsioonis 14 võrrandiga (2).

Nii defineeritud sarnasuse jaoks peab patsientide P_u ja P_v võrdlemiseks arvutama $n \cdot m$ diagnoosi kombinatsiooni. See tähendab, et ühe patsiendi P_1 jaoks kõige sarnasema patsiendi leidmiseks tuleb arvutada

$$\sum_{i=1}^{|\mathbf{P}|} |\mathbf{D_1}| \cdot |\mathbf{D_i}|$$

kombinatsiooni.

Lause 7. *Definitsioonis 17 loodud algoritmid ei ole kommutatiivsed, ehk leiduvad patsiendid P_u ja P_v nii, et*

$$S_{(1,1)}(P_u, P_v) \neq S_{(1,1)}(P_v, P_u)$$

ja

$$S_{(1,2)}(P_u, P_v) \neq S_{(1,2)}(P_v, P_u)$$

.

Lause 8. *Algoritmide $S_{(1,x)}(P_u, P_v), x \in \{1,2\}$ kohta kehtivad järgmised väited:*

1. $\max_{P_u, P_v \in \mathbf{P}} S_{(1,x)}(P_u, P_v) = 1$,
2. $P_u = P_v \Rightarrow S_{(1,x)}(P_u, P_v) = 1$.

Seega oleme defineerinud algoritmid $S_{(1,*)}(P_u, P_v)$, mis leiavad igale patsiendi P_u koodile $dt_{(u,i)}$ kõige sarnasema sisu ja diagnoosimise hetkega koodi $d_{(v,j)}$ patsiendi P_v koodide hulgast. Seda tehakse, kasutades ära varem defineeritud diagnooside tasemel sarnasuse algoritme. Seejärel liidetakse kõik diagnoosi tasemel sarnasused kokku ning jagatakse tulemus patsiendile D_u määratud diagnooside arvuga. Nii tehes saame väärtuse lõigust $[0,1]$, mis vastabki patsiendi P_v suhtelisele sarnasusele patsiendiga P_u.

Kuna selliselt defineeritud sarnasus on suhteline patsiendi P_u suhtes, siis ei tähenda $S_{(1,*)}(P_u, P_v) = 1$, et $P_u = P_v$.

Näide 9. Patsiendi P_u diagnoosiga $\mathbf{D_u} = \{"E10.0"\}$ ja patsiendi P_v diagnoosidega $\mathbf{D_v} = \{"E10.0", "E12.0"\}$ korral $P_u \neq P_v$, aga nende suhteline sarnasus $S_{(1,1)}(P_u, P_v)$ on

$$\begin{aligned} S_{(1,1)}(P_u, P_v) &= \frac{1}{1} \sum_{i=1}^{1} \max_{j \in \{1,2\}} S_1^t(dt_{(u,i)}, dt_{(v,j)}) \\ &= \max\left(S_1(E10.0, E10.0), S_1(E10.0, E12.0)\right) \\ &= \max\left(1; 0{,}5\right) \\ &= 1. \end{aligned}$$

Selliselt defineeritud algoritmi järgi saavad patsiendiga P_u võrdlemisel eelise sellised patsiendid, kellel on elu jooksul olnud rohkem diagnoose. Seda sellepärast, et summa jagatakse läbi vaid P_u diagnooside arvuga ja järelikult on rohkemate diagnoosidega patsiendiga võrreldes lihtsalt rohkem võimalikke kombinatsioone, mis võiksid olla sarnased.

3.2.2 Sarnaseima paari leidmise algoritm

Defineerime algoritmi, mille eesmärgiks on leida kõige sarnasemate diagnoosidega patsientide paar. Ehk patsiendi P_u jaoks on eesmärk leida selline patsient P_v, kellega võrreldes leidub neil kahe peale kokku kõige rohkem sarnaseid diagnoose.

Selline algoritm lahendaks ära lauses 7 puuduva kommutatiivsuse ning rohkemate diagnoosidega patsiendi eelistamise probleemi.

Selleks kasutame algoritmi, mille pakkusid välja D. Girardi *et al* artiklis [24].

Definitsioon 18. Olgu meil patsiendid $P_u \in \mathbf{P}$ diagnoosidega

$$\mathbf{D_u} = \left\{dt_{(u,1)}, dt_{(u,2)}, \ldots, dt_{(u,n)}\right\}$$

ja $P_v \in \mathbf{P}$ diagnoosidega

$$\mathbf{D_v} = \left\{dt_{(v,1)}, dt_{(v,2)}, \ldots, dt_{(v,m)}\right\}.$$

Tähistame

$$\begin{aligned} X &:= \mathbf{D_u} \setminus \mathbf{D_v}, \\ Y &:= \mathbf{D_v} \setminus \mathbf{D_u}, \\ Z &:= \mathbf{D_v} \cup \mathbf{D_u}. \end{aligned}$$

Defineerime patsiendi P_u sarnasuse patsiendiga P_v tähisega $S_{(2,1)}(P_u, P_v)$ ning võrrandiga

$$S_{(2,1)}(P_u, P_v) = 1 - \frac{1}{|Z|}\left(\sum_{dt_{(u,i)} \in X} \frac{1}{|\mathbf{D_v}|} \sum_{j=1}^{m} D_1(dt_{(u,i)}, dt_{(v,j)}) + \sum_{dt_{(v,j)} \in Y} \frac{1}{|\mathbf{D_u}|} \sum_{i=1}^{n} D_1(dt_{(v,j)}, dt_{(u,i)})\right),$$

kus $D_1(a,b) = 1 - S_1^t(a,b)$. Sarnaselt saame defineerida $S_{(2,2)}(P_u, P_v)$:

$$S_{(2,2)}(P_u, P_v) = 1 - \frac{1}{|Z|}\left(\sum_{dt_{(u,i)} \in X} \frac{1}{|\mathbf{D_v}|} \sum_{j=1}^{m} D_2(dt_{(u,i)}, dt_{(v,j)}) + \sum_{dt_{(v,j)} \in Y} \frac{1}{|\mathbf{D_u}|} \sum_{i=1}^{n} D_2(dt_{(v,j)}, dt_{(u,i)})\right),$$

kus $D_2(a,b) = 1 - S_2^t(a,b)$.

Lause 9. *Definitsioonis 18 loodud algoritmid on kommutatiivsed, ehk iga patsiendi P_u ja P_v korral*

$$S_{(2,1)}(P_u, P_v) = S_{(2,1)}(P_v, P_u)$$

ja

$$S_{(2,2)}(P_u, P_v) = S_{(2,2)}(P_v, P_u)$$

.

Lause 10. *Algoritmide* $S_{(2,x)}(P_u, P_v), x \in \{1,2\}$ *kohta kehtivad järgmised väited:*

1. $\max\limits_{P_u, P_v \in \mathbf{P}} S_{(2,x)}(P_u, P_v) = 1$,
2. $P_u = P_v \Rightarrow S_{(2,x)}(P_u, P_v) = 1$.

Siin defineerisime ära algoritmi $S_{(2,*)}(P_u, P_v)$, mis leiab esmalt kõik patsiendi P_u koodid $dt_{(u,i)}$, mida pole patsiendi P_v koodide hulgas. Seejärel leiab algoritm kõigi selliste diagnooside jaoks keskmise kauguse kõigist patsiendi P_v diagnoosidest. Algoritmi esimeseks liidetavaks ongi nende keskmiste kauguste summa. Seejärel leitakse teine liidetav analooglselt patsiendi P_v jaoks. Kui liidetavad on leitud, jagatakse summa kõigi unikaalsete koodide hulgaga (arvestades diagnoosimise hetke), mis on kas patsiendile P_u või P_v määratud. Nii tehes saame kätte patsientide omavahelise erinevuse, ehk sarnasuse leidmiseks lahutame saadud tulemuse ühest. Kokku saame väärtuse lõigust $[0,1]$, mis vastabki patsientide P_u ja P_v omavahelisele sarnasusele. Kuna tehte tulemus ei sõltu liidetavate järjekorrast, ei sõltu tulemus sellest, kas alustame patsiendist P_u või P_v.

Kuna selline algoritm leiab iga P_u unikaalse diagnoosi ja diagnoosimise hetke jaoks keskmise kauguse kõigist P_v diagnoosidest (ja vastupidi), siis selline algoritm töötab paremini lühikese vaatlusaja jaoks. Seda sellepärast, et elu jooksul võivad inimesed saada väga erinevaid haiguseid, mistõttu ühe konkreetse diagnoosi keskmine kaugus kõigist võib tulla väga väike.

4 Näited

Siin peatükis vaatleme genereeritud andmete põhjal saadud tulemusi ning toome välja mõned näited.

4.1 Diagnoosi tasemel sarnasus

Vaatame diagnooside kolmikut:

1. $I60.2$ - subarahnoidaalne hemorraagia eesmisest ühendusarterist,
2. $I05.0$ - mitraalstenoos,
3. $I45.0$ - parem fastsikulaarblokaad.

Diagnoosipaaride $(I60.2, I05.0)$ ja $(I60.2, I45.0)$ kategoorilised sarnasused S_{RHK10} ja $S_{kontekst}$ on vastavalt

$$\begin{aligned} &S_{RHK10}(I60.2, I05.0) = S_{RHK10}(I60.2, I45.0) = 0{,}25, \\ &S_{kontekst}(I60.2, I05.0) = 0{,}501, \\ &S_{kontekst}(I60.2, I45.0) = 0{,}465. \end{aligned}$$

Seega mõlemal juhul on väärtused üsna sarnased: RHK-10 hierarhia järgi erinevus puudub, konteksti põhjal võiks $I60.2$ jaoks sarnasemaks lugeda diagnoosi $I05.0$.

Nende diagnooside suremust vaadates leiame, et $I60.2$ ja $I45.0$ on ohtlikumad kui $I05.0$:

$$\begin{aligned} S(I60.2) &= 0{,}471, \\ S(I05.0) &= 0{,}010, \\ S(I45.0) &= 0{,}202. \end{aligned}$$

Sellest saame, et suremuste sarnasused on

$$\begin{aligned} S_s(I60.2, I05.0) &= 0{,}540, \\ S_s(I60.2, I45.0) &= 0{,}731. \end{aligned}$$

Valides sarnasuste funktsioonide S_1 ja S_2 võrrandites 1 ja 2 kaaludeks

$$w_1 = 0{,}8, w_2 = 0, w_3 = 0{,}2, w_4 = 0,$$

saame sarnasusteks vastavalt

$$\begin{aligned} S_1(I60.2, I05.0) &= 0{,}8 \cdot S_{RHK10}(I60.2, I05.0) + 0{,}2 \cdot S_s(I60.2, I05.0) \\ &= 0{,}8 \cdot 0{,}25 + 0{,}2 \cdot 0{,}540 \\ &= 0{,}308, \end{aligned}$$

$$\begin{aligned} S_1(I60.2, I45.0) &= 0{,}8 \cdot S_{RHK10}(I60.2, I45.0) + 0{,}2 \cdot S_s(I60.2, I45.0) \\ &= 0{,}8 \cdot 0{,}25 + 0{,}2 \cdot 0{,}731 \\ &= 0{,}346 \end{aligned}$$

ja

$$\begin{aligned} S_2(I60.2, I05.0) &= 0{,}8 \cdot S_{kontekst}(I60.2, I05.0) + 0{,}2 \cdot S_s(I60.2, I05.0) \\ &= 0{,}8 \cdot 0{,}501 + 0{,}2 \cdot 0{,}540 \\ &= 0{,}508, \end{aligned}$$

$$\begin{aligned} S_2(I60.2, I45.0) &= 0{,}8 \cdot S_{kontekst}(I60.2, I45.0) + 0{,}2 \cdot S_s(I60.2, I45.0) \\ &= 0{,}8 \cdot 0{,}465 + 0{,}2 \cdot 0{,}731 \\ &= 0{,}518. \end{aligned}$$

Mõlemal juhul peaks selle mudeli järgi sarnasemaks hindama paari $(I60.2, I45.0)$.

Vaatame veel lisaks kolmikut, kus sarnasust mõjutavad kroonilisus ja haruldus:

1. $C18.0$ - „Umbsoole pahaloomuline kasvaja“,
2. $D12.1$ - „Ussripiku healoomuline kasvaja“,
3. $C48.2$- „Täpsustamata kõhukelme pahaloomuline kasvaja“.

Kontekstuaalselt on kõik kolm sarnased, kuna kõigil kolmel juhul on tegemist seedetrakti kasvajatega. Asukoha mõttes on $C18.0$ ja $D12.1$ omavahel sarnasemad, kuna ussripik on umbsoole jätk. Paari $C18.0$ ja $C48.2$ võiks jällegi sarnasemaks

pidada selle poolest, et nad on mõlemad seedetrakti pahaloomulised kasvajad, $D12.1$ aga healoomuline.

Ka selle kolmiku puhul ei aita kategoorilised sarnasused S_{RHK10} ja $S_{kontekst}$ otsust langetada:

$$\begin{aligned} S_{RHK10}(C18.0, D12.1) &= S_{RHK10}(C18.0, C48.2) = \frac{1}{4}, \\ S_{kontekst}(C18.0, D12.1) &= 0{,}899, \\ S_{kontekst}(C18.0, C48.2) &= 0{,}895. \end{aligned}$$

Samuti on kõigil kolmel võrdlemisi sarnane suremuse skoor:

$$\begin{aligned} S(C18.0) &= 0{,}071, \\ S(D12.1) &= 0{,}020, \\ S(C48.2) &= 0{,}204. \end{aligned}$$

Suremuste sarnasus on seega:

$$\begin{aligned} S_s(C18.0, D12.1) &= 0{,}950, \\ S_s(C18.0, C48.2) &= 0{,}866. \end{aligned}$$

Seega valides sarnaselt esimesele näitele funktsioonide S_1 ja S_2 kaaludeks

$$w_1 = 0{,}8, w_2 = 0, w_3 = 0{,}2, w_4 = 0$$

saame sarnasusteks vastavalt

$$\begin{aligned} S_1(C18.0, D12.1) &= 0{,}8 \cdot S_{RHK10}(C18.0, D12.1) + 0{,}2 \cdot S_s(C18.0, D12.1) \\ &= 0{,}8 \cdot 0{,}25 + 0{,}2 \cdot 0{,}950 \\ &= 0{,}390, \end{aligned}$$

$$\begin{aligned} S_1(C18.0, C48.2) &= 0{,}8 \cdot S_{RHK10}(C18.0, C48.2) + 0{,}2 \cdot S_s(C18.0, C48.2) \\ &= 0{,}8 \cdot 0{,}25 + 0{,}2 \cdot 0{,}866 \\ &= 0{,}373 \end{aligned}$$

ja

$$\begin{aligned} S_2(C18.0, D12.1) &= 0{,}8 \cdot S_{kontekst}(C18.0, D12.1) + 0{,}2 \cdot S_s(C18.0, D12.1) \\ &= 0{,}8 \cdot 0{,}899 + 0{,}2 \cdot 0{,}950 \\ &= 0{,}909, \end{aligned}$$

$$\begin{aligned} S_2(C18.0, C48.2) &= 0{,}8 \cdot S_{kontekst}(C18.0, C48.2) + 0{,}2 \cdot S_s(C18.0, C48.2) \\ &= 0{,}8 \cdot 0{,}895 + 0{,}2 \cdot 0{,}866 \\ &= 0{,}889. \end{aligned}$$

Mõlemal juhul on tulemused jätkuvalt väga sarnased, kuid paar ($C18.0, D12.1$) saab veidi kõrgemad tulemused.

Erinevus tuleb sisse, kui arvestada nende koodide haruldust ning kroonilisust, need on vastavalt:

$$\begin{aligned} H(C18.0) &= 0{,}903, \\ H(D12.1) &= 0{,}117, \\ H(C48.2) &= 0{,}974 \end{aligned}$$

ja

$$\begin{aligned} K(C18.0) &= 1, \\ K(D12.1) &= 0, \\ K(C48.2) &= 1. \end{aligned}$$

Ehk vastavalt sarnasused H_s ja K_s:

$$\begin{aligned} H_s(C18.0, D12.1) &= 0{,}214, \\ H_s(C18.0, C48.2) &= 0{,}929. \end{aligned}$$

ja

$$\begin{aligned} K_s(C18.0, D12.1) &= 0, \\ K_s(C18.0, C48.2) &= 1. \end{aligned}$$

Mõlemal juhul on sarnasused paari ($C18.0, C48.2$) jaoks tunduvalt suuremad.

Seega valides funktsioonide S_1 ja S_2 kaaludeks

$$w_1 = 0{,}8, w_2 = 0{,}1, w_3 = 0, w_4 = 0{,}1,$$

saame sarnasusteks vastavalt

$$\begin{aligned} S_1(C18.0, D12.1) &= 0{,}8 \cdot S_{RHK10}(C18.0, D12.1) + 0{,}1 \cdot H_s(C18.0, D12.1) \\ &\quad + 0{,}1 \cdot K_s(C18.0, D12.1) \\ &= 0{,}8 \cdot 0{,}25 + 0{,}1 \cdot 0{,}214 + 0{,}1 \cdot 0 \\ &= 0{,}221, \end{aligned}$$

$$\begin{aligned} S_1(C18.0, C48.2) &= 0{,}8 \cdot S_{RHK10}(C18.0, C48.2) + 0{,}1 \cdot H_s(C18.0, C48.2) \\ &\quad + 0{,}1 \cdot K_s(C18.0, C48.2) \\ &= 0{,}8 \cdot 0{,}25 + 0{,}1 \cdot 0{,}929 + 0{,}1 \cdot 1 \\ &= 0{,}393 \end{aligned}$$

ja

$$\begin{aligned} S_2(C18.0, D12.1) &= 0{,}8 \cdot S_{kontekst}(C18.0, D12.1) + 0{,}1 \cdot H_s(C18.0, D12.1) \\ &\quad + 0{,}1 \cdot K_s(C18.0, D12.1) \\ &= 0{,}8 \cdot 0{,}899 + 0{,}1 \cdot 0{,}214 + 0{,}1 \cdot 0 \\ &= 0{,}741, \end{aligned}$$

$$\begin{aligned} S_2(C18.0, C48.2) &= 0{,}8 \cdot S_{kontekst}(C18.0, C48.2) + 0{,}1 \cdot H_s(C18.0, C48.2) \\ &\quad + 0{,}1 \cdot K_s(C18.0, C48.2) \\ &= 0{,}8 \cdot 0{,}895 + 0{,}1 \cdot 0{,}929 + 0{,}1 \cdot 1 \\ &= 0{,}909. \end{aligned}$$

Mõlemal juhul on paari $(C18.0, C48.2)$ sarnasus märgatavalt suurem kui paari $(C18.0, D12.1)$ sarnasus.

4.2 Patsientide sarnasus

Siin alampeatükis näitame patisentide sarnasuste algoritmide

$$S_{(1,1)}, S_{(1,2)}, S_{(2,1)}, S_{(2,2)}$$

omavahelist erinevust. Selleks kasutame ChatGPT abiga loodud näidisandmeid.

4.2.1 Diabeedihaiged patsiendid

Vaatame kõigepealt kümmet diabeedihaiget patsienti, mis on välja toodud tabelis 2. Igal patsiendil on 2-3 RHK-10 diagnoosi, mis on kas diabeet ise ($E10 - E14$) või sellega kaasnev haigus. Osadel patsientidel otsene diabeedi diagnoos puudub, on vaid kaasnevad haigused. Intuitiivselt võiks kõiki näites toodud patsiente lugeda omavahel sarnaseks. Diagnoosimise ajahetke informatsioon puudub, seega eeldame

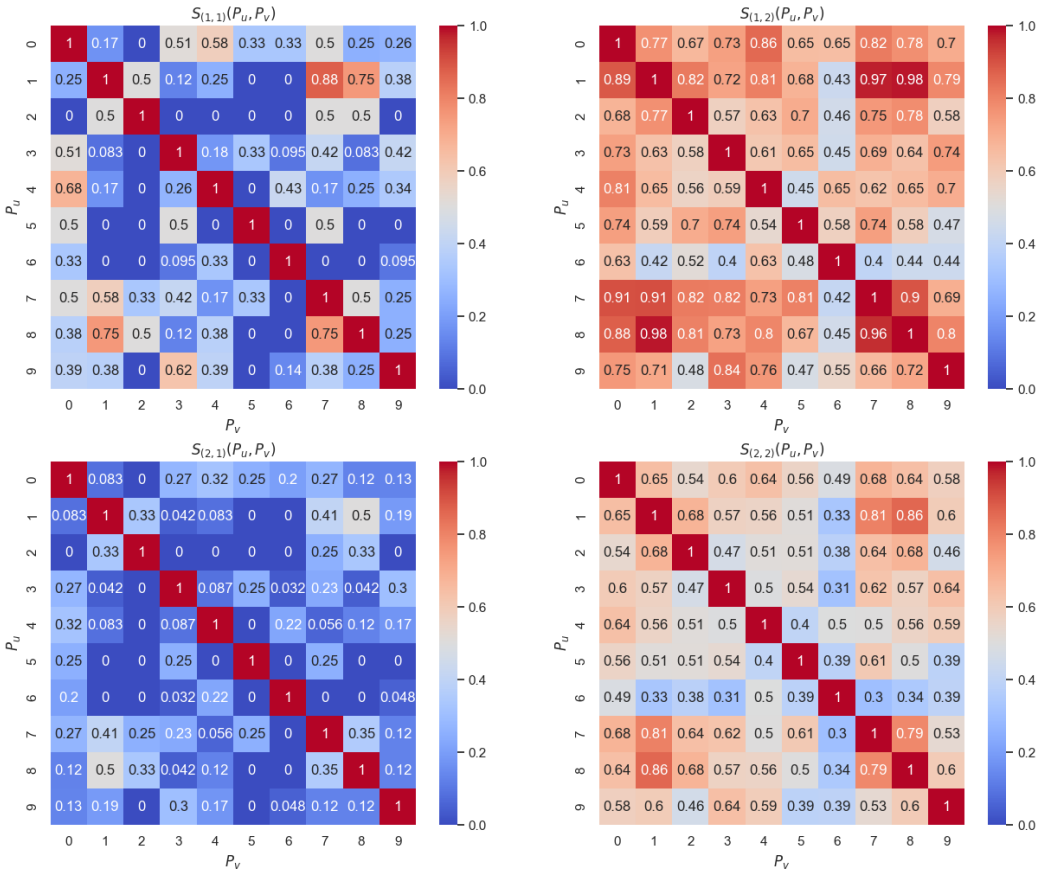
Patsient	Diagnoos	RHK-10 kood
0	Insuliinisõltumatu suhkurtõbi tüsistusteta, kõrgvererõhktõbi, diabeetiline polüneuropaatia	E11.9, I10, G63.2
1	Insuliinisõltuv suhkurtõbi tüsistusteta, diabeetiline retinopaatia	E10.9, H36.0
2	Diabeetiline retinopaatia, kroonilise neeruhaiguse lõppjärk	H36.0, N18.0
3	Diabeetiline polüneuropaatia, perifeersete veresoonte täpsustamata haigus, täpsustamata hüperlipideemia	G63.2, I73.9, E78.5
4	Insuliinisõltumatu suhkurtõbi neerutüsistustega, kõrgvererõhktõbi, aterosklerootiline südamehaigus	E11.2, I10, I25.1
5	Muu krooniline pankreatiit, diabeetiline polüneuropaatia	K86.1, G63.2
6	Mujal klassifitseerimata alajäsemehaavand, muu krooniline osteomüeliit, kõrgvererõhktõbi	L97, M86.6, I10
7	Insuliinisõltuv suhkurtõbi ketoatsidoosiga, diabeetiline polüneuropaatia, diabeetiline retinopaatia	E10.1, G63.2, H36.0
8	Insuliinisõltumatu suhkurtõbi muu täpsustatud tüsistusega, diabeetiline retinopaatia	E11.6, H36.0
9	Insuliinisõltuv suhkurtõbi muu täpsustatud tüsistusega, perifeersete veresoonte täpsustamata haigus	E10.6, I73.9

Tabel 2: ChatGPT genereeritud diabeedihaigete patsientide andmed

siin näites, et ajaline sõltuvus on 0. See tähendab, et antud näites $S_1^t(dt_{(x,i)}, dt_{(y,j)}) = S_1(d_{(x,i)}, d_{(y,j)})$.

Sarnasust hästi kirjeldav algoritm hindaks kõik need patsiendid vähemalt mingil määral sarnaseks. Jooniselt 2 näeme, et sarnasuste $S_{1,2}$ ja $S_{2,2}$, ehk diagnooside konteksti sarnasust $S_{kontekst}$ kasutavad algoritmid teevad seda päris hästi. Keskmine sarnasus nendel juhutdel on vastavalt $0{,}711$ ja $0{,}59$. Mõnevõrra oodatult on nendes kõige madalamad sarnasused patsiendi number 6 suhtes, kuna talle on määratud diagnoosid, mis on küll diabeedi kaasuvad haigused, aga võivad tekkida ka muudel põhjustel.

$S_{1,1}$ ja $S_{2,1}$ leiavad vaid üksikutel juhtudel tugevad sarnasused, kuna paljud diagnoosid on täiesti erinevates peatükkides, mis tähendab, et nendel juhtudel $S_{RHK} = 0$. Nende algoritmide puhul on keskmiseks patsiendide sarnasuseks vastavalt $0{,}331$ ja $0{,}224$.



Joonis 2: Diabeedihaigete patsientide sarnasuse maatriksid, kus algoritmiks on kasutatud vastavalt funktsioone $S_{(1,1)}$, $S_{(1,2)}$, $S_{(2,1)}$ ja $S_{(2,2)}$.

4.2.2 Erinevad põhihaigused

Vaatame nüüd kümmet patsienti, kellest kolmel on põhiliseks haiguseks diabeet ($E10 - E14$), kahel müokardiinfarkt (RHK-10 kood $I21$ või $I22$), kolmel mingi psüühika- või käitumishäire (RHK-10 alamjaotis peatükist $F00 - F99$) ning kahel midagi muud - vastavalt $K52.9$, ehk täpsustamata mittenakkuslik gastroenteriit või koliit ning $R07.2$, ehk rinnakutagune valu. Kõigil patsientidel on põhidiagnoosiga 2-5 seonduvat RHK-10 diagnoosi veel lisaks.

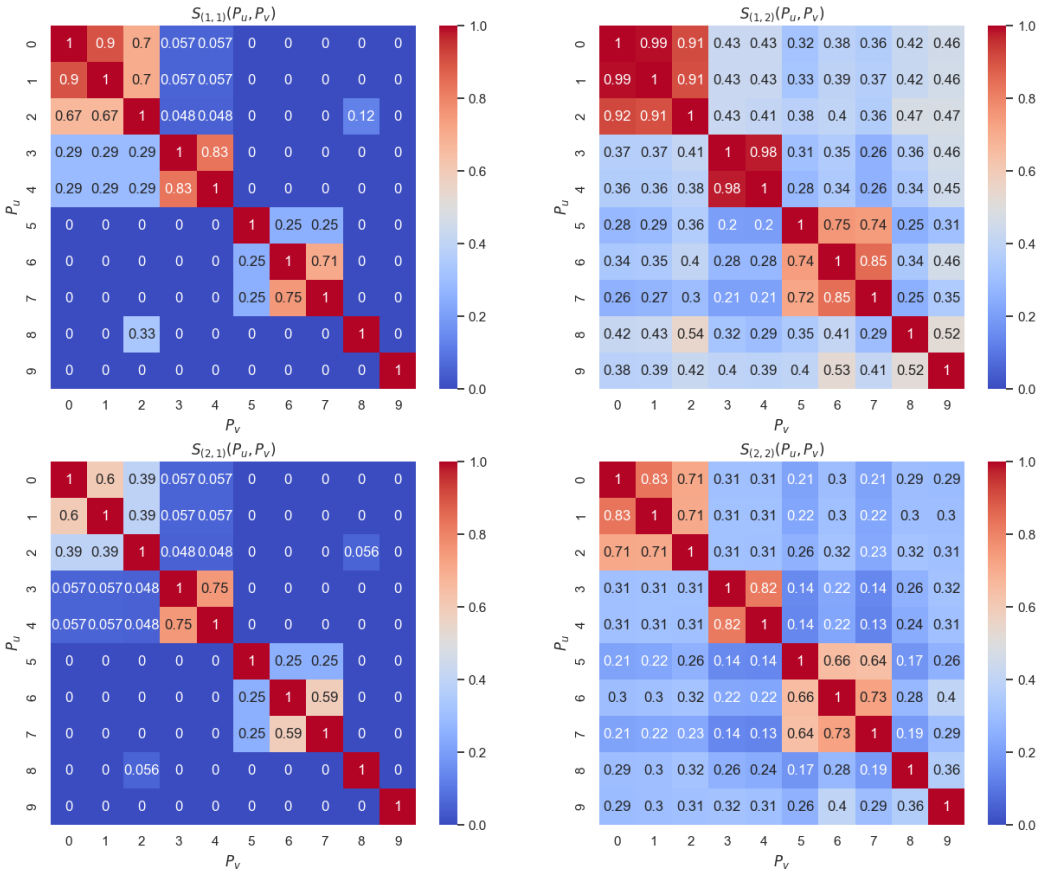
Täpsemad diagnoosid on kirjeldatud tabelis 3. Sellise andmestiku jaoks võiks sarnasuse algoritm sarnaseimaks pidada omavahel patsiente $0, 1, 2$ ning patsiente 3 ja 4. Samuti võiksid omavahel pigem sarnased olla 5,6 ja 7, kuid vähem kui eelmised grupid. Seda sellepärast, et side kõigi erinevate psüühikahäirete vahel on väiksem kui diabeedi- ja müokardiinfarkti erinevate vormide vahel. Teistest ning ka omavahel vähem sarnaseks võiks pidada patsiente 8 ja 9, kuna nende põhidiagnoosid on küllaltki erinevad varasemalt mainitud patsientide omadest.

Patsient	Põhidiagnoos	Seotud diagnoosid
0	E10.9	E11.1, E78.0, I10, N18.1
1	E11.9	E10.0, E78.0, I10, N18.1
2	E14.1	E10.1, E11.4, E78.0, I10, K86.0
3	I21.0	I20.0, I25.1
4	I22.1	I20.0, I25.1
5	F20.1	F22.0, F23.2, F25.1, F60.1, F60.2
6	F32.0	F33.0, F41.0, F41.1, F42.0, F43.2
7	F43.2	F40.1, F41.0, F42.0, F44.1, F45.1
8	K52.9	K50.0, K51.0, K86.1, K92.0, K92.1
9	R07.2	J44.9, J45.9, R06.0, R07.4, R10.4

Tabel 3: ChatGPT genereeritud patsientide põhidiagnoosid koos seotud haigustega.

Jooniselt 3 näeme, et sarnasuste $S_{(1,2)}$ ja $S_{(2,2)}$, ehk diagnooside konteksti sarnasust $S_{kontekst}$ kasutavad algoritmid kirjeldavad sarnasusi nii, nagu ootasime. Ülejäänud seosed ei ole ka väga nullilähedased, mis tuleneb ilmselt sellest, et $S_{kontekst}$ erinevate peatükkide vahel on keskmiselt 0,28.

Algoritm $S_{(1,1)}$ leiab tugeva sarnasuse patsientide 0,1,2 ning 3 ja 4 vahel. Patsientide $6, 7$ vahel leiab samuti suure sarnasuse, kuid patsiendiga 5 kumbagi väga sarnaseks ei hinda. Algoritm $S_{(2,1)}$ hindab sarnaseks vaid patsientide paari 3 ja 4. Mujal on nende kahe algoritmi sarnasuse väärtused kas 0 või väga selle lähedal, kuna paljud diagnoosid on täiesti erinevates peatükkides, mis tähendab, et nendel juhtudel $S_{RHK} = 0$.



Joonis 3: Erinevate põhidiagnoosidega patsientide sarnasuse maatriksid, kus algoritmiks on kasutatud vastavalt funktsioone $S_{(1,1)}$, $S_{(1,2)}$, $S_{(2,1)}$ ja $S_{(2,2)}$.

4.3 Ajast sõltuv sarnasus

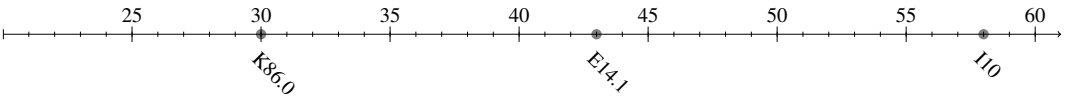
Valime 10 patsienti, kellel on kõigil kolm diagnoosi:

1. I10 - kõrgvererõhktõbi,
2. K86.9 - alkoholi põhjustatud krooniline pankreatiit,
3. E14.1 - täpsustamata suhkurtõbi ketoatsidoosiga.

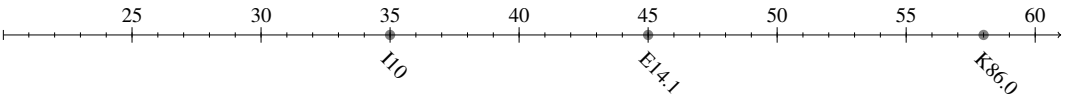
Patsientide $0-8$ jaoks valime igale diagnoosile suvaliselt diagnoosimise ajaks vanuse lõigust $[20{,}60]$. Patsiendi 9 diagnoosi ajad valime patsiendi 8 aegadest ühe aasta varasemad.

Nii tehes saame patsiendid:

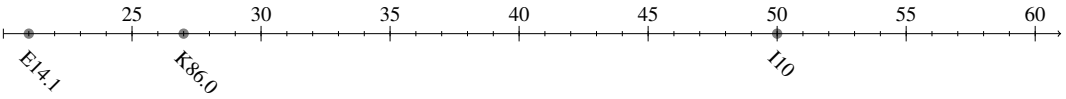
Patsient 0:



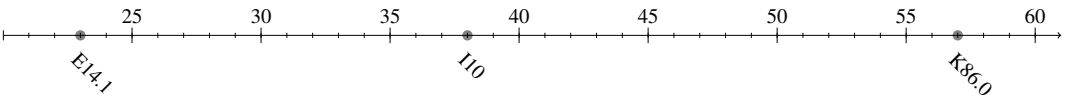
Patsient 1:



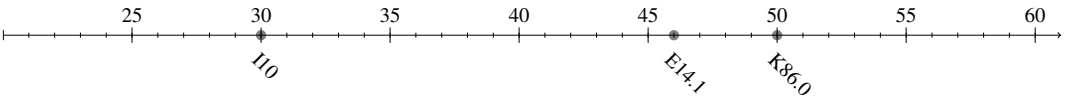
Patsient 2:



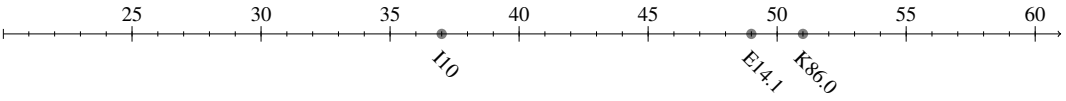
Patsient 3:



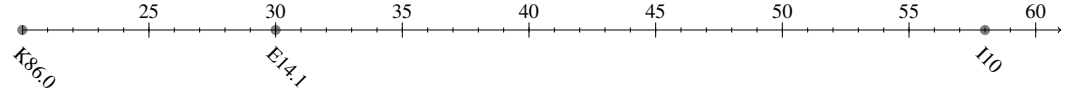
Patsient 4:



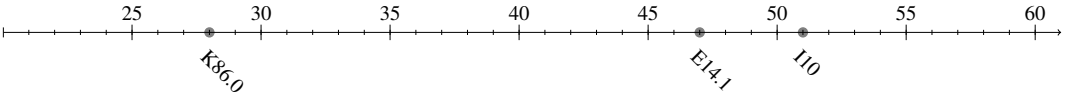
Patsient 5:



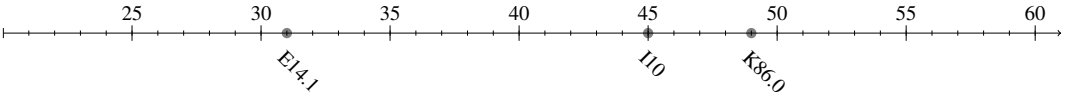
Patsient 6:



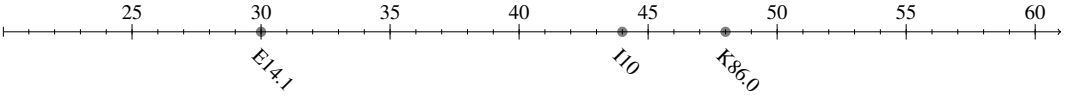
Patsient 7:



Patsient 8:



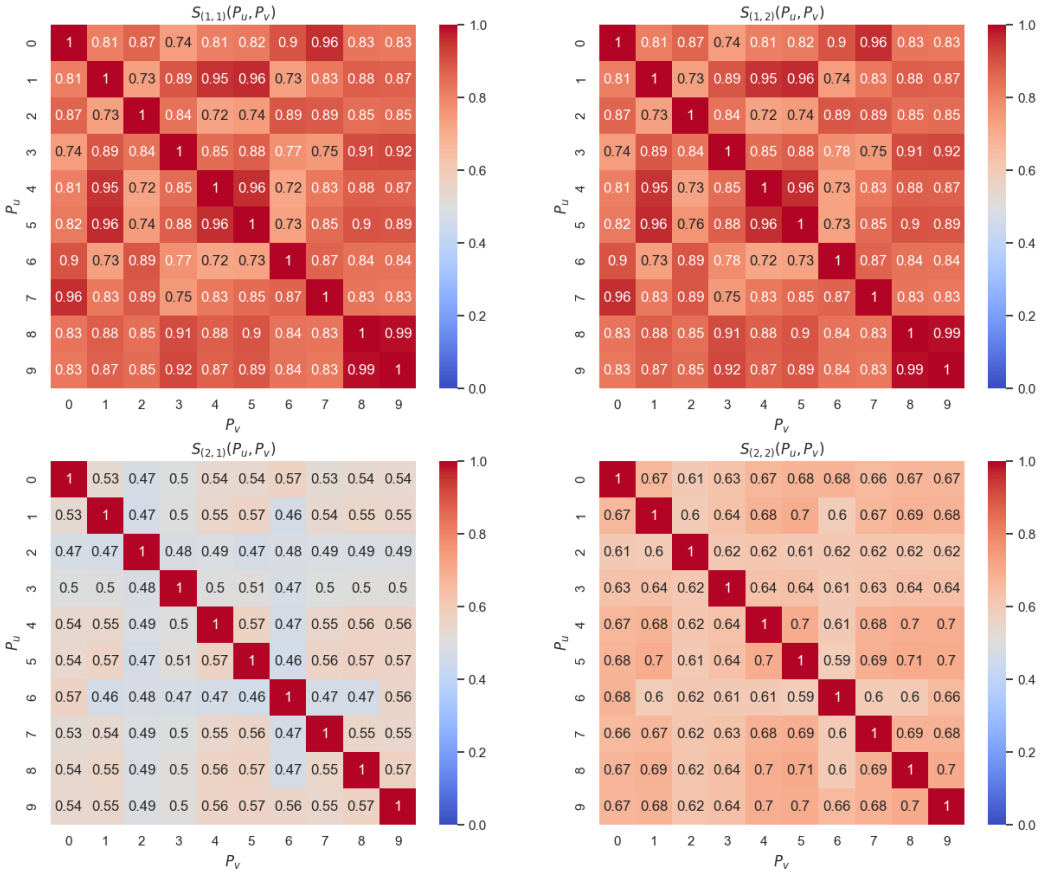
Patsient 9:



Ilmselt tuleks kõige sarnasemaks hinnata omavahel patsiendid 8 ja 9. Samuti võiksid 1, 4 ja 5 omavahel sarnased olla, kuna diagnoosid on samas järjekorras ja kõik haigused on enam-vähem sama vanalt diagnoositud. 0 jaoks kõige sarnasem võiks olla 7 ning 6. Patsiendi 2 jaoks täpselt sama patsienti ei leidu, aga patsiendil 7 on kaks diagnoosi peaaegu samal ajal diagnoositud. 3 on enam vähem sarnane patsientidega 8 ja 9.

Võtame funktsioonides $S_1^t(dt_{(x,i)}, dt_{(y,j)})$ ja $S_2^t(dt_{(x,i)}, dt_{(y,j)})$ kaaludeks $W_1 = 0{,}5$ ja $W_2 = 0{,}5$. Sellisel juhul funktsioonide $S_{(1,1)}$ ja $S_{(1,2)}$ minimaalseks väärtuseks on $0{,}5$.

Jooniselt 4 näeme, et kõige paremini kirjeldavad sarnasusi funktsioonid $S_{(1,1)}$ ja $S_{(1,2)}$, kuna need saavad oodatud tulemustega võrreldavad tulemused. $S_{(2,1)}$ ja $S_{(2,2)}$ aega arvestavat sarnasust väga hästi ei leia, kuna arvestavad patsiendi P_u iga diagnoosi (ja diagnoosihetke) jaoks kõiki patsiendi P_v diagnoose koos diagnoosihetkedega.



Joonis 4: Kolme sama diagnoosi erineva ajajaotusega patsientide sarnasuse maatriksid, kus algoritmiks on kasutatud vastavalt funktsioone $S_{(1,1)}$, $S_{(1,2)}$, $S_{(2,1)}$ ja $S_{(2,2)}$.

4.4 Algoritmide hinnang

Näidete põhjal näeme, et reeglina töötavad paremini algoritmid, mis kasutavad diagnooside tasemel sarnasuse hindamiseks diagnooside konteksti, ehk sarnasust $S_{kontekst}$ kasutavad algoritmid $S_{(1,2)}$ ja $S_{(2,2)}$. Enamasti sai oodatud sarnaste patsientide jaoks kõige suuremad väärtused algoritm $S_{(1,2)}$, kuid kui aega mitte arvestada, siis sai häid tulemusi ka algoritm $S_{(2,2)}$. Sellistel juhtudel oleks soovitatav siiski kasutada algoritmi $S_{(2,2)}$, kuna siis saame patsientide suhtes sümmeetrilised tulemused ning väldime $S_{(1,*)}$ kalduvust eelistada erinevate diagnoosihulkadega patsiente.

Ajast sõltuvate sarnasuste leidmisega ei saa algoritm $S_{(2,*)}$ väga hästi hakkama. Seda sellepärast, et see leiab iga haiguse jaoks keskmise kauguse teise patsiendi haigustest ka juhutudel, kui diagnoosid on võrdsed, aga diagnoosimise hetked on erinevad. Parema tulemuse saamiseks peaks ajast sõltuvate sarnasuste leidmisel eelistama esimest algoritmi.

Kokkuvõte

Käesoleva töö eesmärgiks oli luua algoritm, mis võimaldaks omavahel võrrelda patsiente nende varasemate diagnooside põhjal.

Töö esimeses pooles kirjeldasime kahte peamist diagnoosi tasemel sarnasuse algoritmi. Nendest esimene, hierarhiapõhine algoritm, toetus varasematele tulemustele. Teine, kontekstipõhine algoritm, on käesolevas töös välja pakutud uus variant, mis toetub *word2vec* mudelile. Kahe algoritmi omavaheliseks võrdlemiseks esitasime mõned näited ning võrdlesime algoritmide tulemuste erinevusi. Lisaks kirjeldasime ning kaasasime sarnasuse arvutamisse ka muid diagnooside omadusi, mida saaks kasutada erinevate eesmärkide täitmiseks. Töö teises pooles keskendusime patsientide omavahelise sarnasuse leidmisele. Selleks kirjeldasime esmalt ära, kuidas hinnata diagnooside määramisaegade vahet. Seejärel kirjeldasime kaks erinevat algoritmi patsientide sarnasuse leidmiseks, millest mõlemad toetusid varasematele tulemustele. Algoritmide võrdlemiseks tõime lõpus välja mitu näidet, mille puhul algoritmi valimisega tulemus muutus.

Töö tulemusena valmis mitu algoritmi, mida erinevatel tingimustel tuleks eelistada teistele. Uue tulemusena loodud diagnooside kontekstipõhise algoritmi kasutamisel saime üldjuhul lähemale oodatud tulemustele, kui varasemalt kasutusel olnud hierarhilise mudeliga. Samuti aitavad diagnooside lisatunnused võrrelda omavahel muidu sarnaseid diagnoosipaare. Patsiendi tasemel loodud algoritmide valik sõltub eelkõige sarnasuse hindamise eesmärgist, kuna üks töötab paremini ajast sõltuvate diagnoosside võrdlemiseks ning individuaalse patsiendi jaoks sarnaseima leidmiseks, teine aga sarnaseima paari leidmiseks üldpopulatsioonist.

Viidatud kirjandus

[1] Eesti Sotsiaalministeerium. Rahvusvaheline haiguste ja nendega seotud terviseprobleemide statistiline klassifikatsioon : RHK-10 : kümnes väljaanne : 2. köide. Instruktsioonide käsiraamat. Tallinn: Sotsiaalministeerium, 1996.

[2] Andmekaitse Inspektsioon. Isikuandmed ja töötlemine. 2019. URL: https://www.aki.ee/et/eraelu-kaitse/isikuandmed-ja-tootlemine (vaadatud 07.05.2023).

[3] S. Travasci. simple_icd_10. GitHub. 2020. URL: https://github.com/StefanoTrv/simple_icd_10 (vaadatud 07.05.2023).

[4] WHO. International Statistical Classification of Diseases and Related Health Problems 10th Revision. 2019. URL: https://icd.who.int/browse10/2019/en (vaadatud 07.05.2023).

[5] Eesti Sotsiaalministeerium. Rahvusvahelise haiguste klassifikatsiooni RHK-10 andmebaas. URL: https://rhk.sm.ee/ (vaadatud 07.05.2023).

[6] C.G. Chute, C. Çelik. „Overview of ICD-11 architecture and structure“. *BMC Med Inform Decis Mak* 21 (Suppl 6) (2021), lk. 378.

[7] A. Valdas. Juhuslike diagnooside trajektooride generaator. TÜ arvutiteaduse instituudi bakalaureusetöö, 2021.

[8] STACC. Ülevaade eriarstiabi saanute arvust ja kuludest vanuse ning põhidiagnooside lõikes. 2020. URL: https://stacc.ee/ehif-stacked-area/?lng=Et (vaadatud 07.05.2023).

[9] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas. URL: https://statistika.tai.ee/pxweb/et/Andmebaas (vaadatud 07.05.2023).

[10] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas, tabel EH01: Diabeedi esmashaigusjuhud soo ja vanuserühma järgi. URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__01Esmashaigestumus/EH01.px/ (vaadatud 07.05.2023).

[11] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas, tabel NH02: Valitud nakkushaiguste registreeritud juhtude arv ja kordaja 100 000 elaniku kohta soo ja vanuserühma järgi. URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__02Nakkushaigused/NH02.px/ (vaadatud 07.05.2023).

[12] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas, tabel TB10: Tuberkuloosi esmasjuhud paikme, soo ja vanuserühma järgi. URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__03Tuberkuloos/TB10.px/ (vaadatud 07.05.2023).

[13] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas, tabel PK10: Pahaloomuliste kasvajate esmasjuhud paikme, soo ja vanuserühma järgi. URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__04PahaloomulisedKasvajad/PK10.px/ (vaadatud 07.05.2023).

[14] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas, tabel PKH1: Psühhiaatri poolt ambulatoorselt konsulteeritud isikud diagnoosi, soo ja vanuserühma järgi. URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__05Psyyhikahaired/PKH1.px/ (vaadatud 07.05.2023).

[15] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas, tabel EH10: Esmashaigusjuhud soo ja vanuserühma järgi (1998-2016). URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__01Esmashaigestumus/EH10.px/ (vaadatud 07.05.2023).

[16] Tervise Arengu Instituut. Tervisestatistika ja terviseuuringute andmebaas, tabel AMI02: Ägeda müokardiinfarktiga hospitaliseeritud patsiendid (juhud) soo, vanuserühma ja infarkti alatüübi järgi. URL: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__08AMI/AMI02.px/ (vaadatud 07.05.2023).

[17] A. Gottlieb, G. Stein, E. Ruppin, R. Altman, R. Sharan. „A method for inferring medical diagnoses from patient similarities“. *BMC medicine* 11 (2013), lk. 194.

[18] T. Mikolov, K. Chen, G. Corrado, J. Dean. „Efficient Estimation of Word Representations in Vector Space“ (2013).

[19] S. Van Landeghem *et al.* „Large-Scale Event Extraction from Literature with Multi-Level Gene Normalization“. *PLOS ONE* 8(4) (2013).

[20] J. Han, M. Kamber, J. Pei. Data Mining: Concepts and Techniques (Third Edition). The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012.

[21] Eesti Statistikaamet. RV56: Surnud surmapõhjuse, soo ja vanuserühma järgi. URL: https://andmed.stat.ee/et/stat/rahvastik__rahvastikusundmused_surmad/RV56 (vaadatud 07.05.2023).

[22] HCUP. Chronic Condition Indicator Refined for ICD-10-CM Diagnoses, v2023.1. Agency for Healthcare Research and Quality, Rockville, MD. URL: https://hcup-us.ahrq.gov/toolssoftware/chronic_icd10/chronic_icd10.jsp (vaadatud 08.05.2023).

[23] P. Haase, R. Siebes, F. van Harmelen. „Peer Selection in Peer-to-Peer Networks with Semantic Topologies“. *Semantics of a Networked World. Semantics for Grid Databases* (2004), lk. 108–125.

[24] D. Girardi *et al.* „Using concept hierarchies to improve calculation of patient similarity“. *Journal of Biomedical Informatics* 63 (2016), lk. 66–73.

Lisad

A Muudetud ingliskeelsed RHK-10 diagnooside kirjeldused

RHK-10 diagnoos	Originaalne tekst	Muudetud tekst
A56.1	Chlamydial infection of pelviperitoneum and other genitourinary organs	Chlamydial infection of pelvis and other genitourinary organs
B33.1	Ross River disease	RRV
B60.2	Naegleriasis	Infection by naegleria
B88.1	Tungiasis [sandflea infestation]	Tunga penetrans infestation
C96.5	Multifocal and unisystemic Langerhans-cell histiocytosis	Langerhans cell histiocytosis, multifocal unisystem
D74.8	Other methaemoglobinaemias	Other methemoglobinemia
D82.1	Di George syndrome	Digeorge
E07.1	Dyshormogenetic goitre	Dyshormonogenic goiter
E24.1	Nelson syndrome	Postadrenalectomy syndrome
E89.6	Postprocedural adrenocortical (-medullary) hypofunction	Postprocedural adrenocortical (medullary) hypofunction
J43.0	MacLeod syndrome	Hyperlucent lung syndrome
J66.2	Cannabinosis	Airway disease due to specific organic dust
J67.4	Maltworker lung	Maltworkers lung
J84.0	Alveolar and parietoalveolar conditions	Alveolar and parieto alveolar conditions
K02.4	Odontoclasia	Tooth root resorption
L68.3	Polytrichia	Hypertrichosis
M02.3	Reiter disease	Reactive arthritis
M71.9	Bursopathy, unspecified	Bursitis
N41.3	Prostatocystitis	Prostatic cystitis
O02.0	Blighted ovum and nonhydatidiform mole	Blighted ovum and non hydatidiform mole
Q06.0	Amyelia	Spinal cord agenesis
Q18.7	Microcheilia	Abnormal smallness of the lips
Q75.5	Oculomandibular dysostosis	Oculo mandibular dysostosis
R82.2	Biliuria	Bilirubinuria

Tabel 4: Ingliskeelsete RHK-10 diagnooside kirjelduste muutmine.

B Kroonilisuse käsitsi määramine

Jaotis	Nimi	Kroonilisus
A16	Bakterioloogiliselt või histoloogiliselt kinnitamata hingamiselundite tuberkuloos	0
B21	Pahaloomuliste kasvajatena avalduv HIV-tõbi	1
B22	Muude täpsustatud haigustena avalduv HIV-tõbi	1
B23	Muude haigusseisunditena avalduv HIV-tõbi	1
B24	Täpsustamata HIV-tõbi	1
C97	Sõltumatute (primaarsete) hulgipaikmete pahaloomulised kasvajad	1
E12	Väärtoitumussuhkurtõbi	1
E14	Täpsustamata suhkurtõbi	1
E90	Toitumus- ja ainevahetushäired mujal klassifitseeritud haiguste korral	0,5
F00	Dementsus Alzheimeri tõvest	1
F38	Muud meeleoluhäired	1
F61	Segatüüpi ja muud isiksushäired	1
F62	Ajukahjustuse ja -haigusega mitteseostatavad püsivad isiksusemuutused	1
F83	Segatüüpi spetsiifilised arenguhäired	1
F92	Segatüüpi käitumis- ja tundeeluhäired	1
G22	Parkinsonism mujal klassifitseeritud haiguste korral	1
G41	Epileptiline staatus e seisund	0
H03	Lau haigusseisundid mujal klassifitseeritud haiguste korral	0
H06	Pisaraelundite ja silmakoopa haigusseisundid mujal klassifitseeritud haiguste korral	0
H13	Konjunktiivi haigusseisundid mujal klassifitseeritud haiguste korral	0
H19	Skleera ja kornea haigusseisundid mujal klassifitseeritud haiguste korral	0
H45	Klaaskeha ja silmamuna haigusseisundid mujal klassifitseeritud haiguste korral	0
H48	Nägemisnärvi ja -kulglate haigusseisundid mujal klassifitseeritud haiguste korral	1
H58	Silma ja silmamanuste muud haigusseisundid mujal klassifitseeritud haiguste korral	0

Jaotis	Nimi	Kroonilisus
I64	Täpsustamata kas hemorraagia või infarktitekkene insult e rabandus	1
I98	Vereringeelundite muud haigusseisundid mujal klassifitseeritud haiguste korral	0
J46	Raskekujuline äge astma	0
K07	Hammaste ja näo arenguhäired [kaasa arvatud hambumushäired]	1
K10	Muud lõualuuhaigused	0
K93	Mujal klassifitseeritud haiguste korral esinevad seedeelundite haigusseisundid	0
M03	Postinfektsioossed e nakkusejärgsed ja reaktiivsed artropaatiad mujal klassifitseeritud haiguste korral	0
M09	Juveniilne artriit mujal klassifitseeritud haiguste korral	0,5
M68	Sünoviaalkestade ja kõõluste haigusseisundid mujal klassifitseeritud haiguste korral	0
M73	Pehmete kudede haigusseisundid mujal klassifitseeritud haiguste korral	0
M82	Osteoporoos mujal klassifitseeritud haiguste korral	1
O05	Muu abort	0
O06	Täpsustamata abort	0
O81	Vaakumekstraktsioon-ja tangüksiksünnitus	0
O83	Muul viisil abistatud üksiksünnitus	0
O84	Mitmiksünnitus	0
O95	Sünnitusabiga seotud täpsustamata põhjusega surm	0
O96	Sünnitusabiga seotud mistahes põhjusega surm ajavahemikul 42 päeva kuni üks aasta pärast sünnitust	0
O97	Surm sünnitusabiga seotud otseste põhjuste jääknähtude tõttu	0
P20	Üsasisene e intrauteriinne hüpoksia e hapnikuvaegus	0
P21	Sünniasfüksia e -lämbus	0
P75	Mekooniumiileus e esmasrooja-soolesulgus	0
R02	Mujal klassifitseerimata gangreen	0
R72	Mujal klassifitseerimata valgeliblede hälbeline leid	1
R95	Väikelapse äkksurma sündroom e haigustunnustik	0
R96	Muu äkksurm teadmata põhjusel	0
R98	Tunnistajateta surm	0

Tabel 5: Käsitsi määrtatud kroonilisused.

C Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Sander Tamm,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose, **Algoritmiline definitsioon patsientide trajektooride sarnasusele**, mille juhendaja on Jaak Vilo, 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 3.0, 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.

Sander Tamm ***09.05.2020***