TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Milena Suvorova 206039IABB

Eestikeelsete meditsiinitekstide vastendamine SNOMED CT-ga

Bakalaureusetöö

Juhendaja: Bahdan Yanovich
BSc
Kaasjuhendaja: Gunnar Piho
PhD

Tallinn 2024

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Milena Suvorova

26.05.2024

Annotatsioon

Käesoleva lõputöö eesmärgiks on tuvastada ja analüüsida probleeme, mis tekivad eestikeelsete vabade meditsiiniliste tekstide vastendamisel SNOMED CT taksonoomiakoodidega. Metoodika hõlmas 71 valitud kõhu- ja vaagnapiirkonna ultraheliuuringu meditsiinilise teksti käsitsi analüüsimist, kus tekstid jagati sõnadeks ja fraasideks ning vastendati SNOMED CT koodidega. Veel koostati sagedussõnastiku vabatekstide alusel, kasutades Pythoni skripti, mille lõi ChatGPT abiga, et tuvastada kõige sagedamini esinevad sõnad ja fraasid. Lisaks katsetati vastendada tekste tehisintellekti abil.

Käsitsi vastendamise tulemused näitasid, et SNOMED CT eestikeelse versiooni osad koodid olid tuvastatavad, kuid ilmnesid ka mitmed probleemid nagu konteksti määramatus, eituse vormid ja lühendite kasutamine. Leiti, et tõlke kvaliteet mängib olulist rolli vastendamise täpsuses. Lisaks selgus, et mõned terminid ja mõisted puudusid SNOMED CT-s eesti keelses versioonis, mis tegi vastendamise protsessi keerulisemaks.

Järeldustest selgus, et SNOMED CT taksonoomia kasutamine eesti keele vabatekstide vastendamiseks on võimalik, kuid nõuab täiendavat tõlketegevust ja metoodika täpsustamist. Tulevikus võiks kaaluda tehisintellekti ja masinõppe meetodite integreerimist, et parandada automaatse vastendamise täpsust ja efektiivsust. Lõppkokkuvõttes annab see töö väärtusliku ülevaate SNOMED CT kasutamise väljakutsetest ja võimalustest meditsiinilise dokumentatsiooni parendamiseks Eestis.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 34 leheküljel, 5 peatükki, 17 joonist, 5 tabelit.

Abstract

Mapping of Estonian-language medical texts into SNOMED CT

The goal of this thesis is to identify and analyse the problems that arise when matching Estonian free medical texts with SNOMED CT taxonomy codes. The methodology involved manually analyzing 71 selected medical texts of abdominal and pelvic ultrasound examinations, where the texts were divided into words and phrases and mapped with SNOMED CT codes. Additionally, a frequency dictionary of free texts was compiled using a Python script created with the help of ChatGPT to identify the most frequently occurring words and phrases. Furthermore, an attempt was made to map the texts using artificial intelligence.

The results of the manual mapping showed that some codes from the Estonian version of SNOMED CT were identifiable, but several problems were encountered, such as contextual ambiguity, negation forms, and the use of abbreviations. It was found that the quality of translation plays a significant role in the accuracy of mapping. Additionally, it was discovered that some terms and concepts were entirely missing from the Estonian version of SNOMED CT, which made the mapping process more challenging.

The conclusions revealed that using SNOMED CT taxonomy for mapping Estonian free texts is possible but requires additional translation efforts and methodological refinement. In the future, it would be worth considering the integration of artificial intelligence and machine learning methods to improve the accuracy and efficiency of automatic mapping. Ultimately, this thesis provides a valuable overview of the challenges and possibilities of using SNOMED CT for enhancing medical documentation in Estonia.

The thesis is in Estonian language and contains 34 pages of text, 5 chapters, 17 figures, 5 tables.

Lühendite ja mõistete sõnastik

SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
UMLS	Unified Medical Language System
NLP	Natural Language Processing
MI	Myocardial infarction

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Käesolevas bakalaureusetöös uuritakse meditsiinilise teksti vastendamise asjakohasust SNOMED CT taksonoomiakoodide abil kõhu- ja vaagnapiirkonna ultraheliuuringute näitel.

1.1 Taust ja probleem

Igapäevaselt genereerivad haiglad ja kliinikud kogu maailmas tohutu hulga [1] andmeid, millest suur osa on struktureerimata tekstide kujul. Need tekstid sisaldavad tavaliselt uuringutulemuste kirjeldusi, diagnoose ja mitmesuguseid kliinilisi märkmeid. Meditsiinilise informatsiooni standardiseerimise ja struktureerimise abil saavad tervishoiuorganisatsioonid integreerida andmeid oma infosüsteemidesse, mis lihtsustab andmete haldamist, analüüsimist ja jagamist [2]. Suur osa neist andmetest jääb vabatekstide kujul, mis raskendab nende töötlemist ja analüüsi ning vähendab erinevate tervishoiu infosüsteemide koostöö efektiivsust [3].

Kuigi vabad tekstid on inimestele hästi mõistetavad, on nende töötlemine arvutitega keeruline. Ühelt poolt võib ühte ja sama nähtust kirjeldada erinevalt (nt südameatakk, MI ja müokardiinfarkt on sama tähendusega). Teiselt poolt võib sama sõna tähendus olenevalt kontekstist erineda (nt vererõhk kui kliiniline leid, vererõhk kui patsiendi kaebus, vererõhu mõõtmine kui protseduur jne). Selle probleemi lahendamiseks on loodud meditsiiniliste terminite nomenklatuurid, nagu UMLS (Unified Medical Language System) [4] ja SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) [5], [6]. Need nomenklatuurid pakuvad mitte ainult meditsiiniliste terminite loendit, vaid ka nende semantikat ja omavahelisi seoseid. SNOMED CT on üks ülemaailmselt kasutatavatest standarditest elektroonilistes tervisekaartides [5], [7].

On juba tehtud katseid vabade meditsiiniliste tekstide vastendamiseks SNOMED CT ja UMLS-iga. Kuna UMLS on väga suur (umbes 3 miljonit terminit), keskendun SNOMED CT-le, kus on umbes 360 000 terminit [3].

Kirjanduses on kirjeldatud erinevad lähenemisviise vastendamiseks: käsitsi, reeglipõhine *(rule based)*, loomuliku keele töötlemine (edaspidi NLP) ja masinõpe ning nende hübriidid [7]. Tulemused on olnud väga paljulubavad: kuni 97% terminidest oli SNOMED CT-s saadaval vastendamiseks [2], vastavuse täpsus oli 73%-94% [8], täpsus 67%-77% ja katvuse määr vastavuse leidmisel 73%-92% [1], täpsus 82%-97% ja katvuse määr 88%-90% [9]. Kuid selliseid kõrgeid tulemusi võib seletada väga kitsaste kasutusjuhtumitega (nt kopsuvähi staadiumide klassifitseerimine [8], gastrektoomiapatsiendid [2], patsientide valimine kliinilisteks uuringuteks [9]) ja ingliskeelsete tekstide vastendamisega.

Enamik uurimustöid on tehtud inglise keele jaoks. On mõned artiklid rootsi ja hiina keele kohta [7]. Olemas on ka eestikeelne versioon SNOMED CT-st [6]. Senini, minu teadmise kohaselt pole keegi proovinud automatiseerida eesti keele vabatekstide vastendamist SNOMED CT-ga. Lisaks ei pruugi olemasolevad ingliskeelsed vastendamise mudelid teistesse keeltesse kergesti üle kantavad olla [7]. Samuti väärib märkimist, et eestikeelne versioon ei ole täielikult tõlgitud. Enne automaatse vastendamise proovimist on vaja teha käsitsi vastendamist. Käsitsi lähenemine aitaks hinnata, kuidas mõjutab tõlkekvaliteet eestikeelse versiooni kasutatavust ning tuvastada potentsiaalseid väljakutseid automatiseeritud vastendamise käigus.

1.2 Eesmärk ja oodatud tulemused

Käesoleva lõputöö eesmärk on tuvastada ja analüüsida probleeme, mis tekivad eestikeelsete vabade meditsiiniliste tekstide vastendamisel SNOMED CT taksonoomiakoodidega.

Oodatud tulemusena peaks töö käigus valmima vastendamisel tekkivate väljakutsete nimekiri. Nende kaardistamine aitaks mõista, millised aspektid mõjutavad vastendamise kvaliteeti. See omakorda võimaldaks tulevikus luua parem automatiseeritud vastendamise mudel.

1.3 Töö struktuur

Bakalaureusetöö töö koosneb 5 peatükist. Esimeses neist antakse ülevaade teema aktuaalsusest ja sõnastatakse lõputöö eesmärke. Teises osas tõstakse esile peamist objekti

ja kasutatud tööriistu. Järgmisena kirjeldatakse kirjanduse ülevaadet ja töö praktilist osa ning tuuakse ette peamised tulemused. Neljandas osas tehakse analüüsi koos järeldustega, selgitatakse töö edasiarenduse võimalusi. Viimasena tehakse kokkuvõte tehtud tööst.

2 Metoodika

Metoodika osas edastan infot kuidas praktiline osa tööst oli tehtud ja miks see oli tehtud.

2.1 Peamine objekt

Peamiseks töö objektiks on eestikeelsete meditsiiniliste tekstide vastendamise protsess SNOMED CT eestikeelse versiooniga ja sellega kaasnevad väljakutsed. Töö raames katsetakse käsitsi vastendamist ja vastendamist ChatGPT [10] tehisintellekti abil. Samuti koostatakse ette antud tekstide põhjal sõnade ja väljendite sagedussõnastik. Töö tulemusena koondati vastendamist mõjutavad aspektid ja väljakutsed. Tulevikus see lihtsustab automatiseerimist ja pärast põhjalikku töötlemist saab sagedussõnastikku kasutada individuaalse mudeli loomiseks konkreetse arsti jaoks.

SNOMED CT eestikeelne versioon uuendatakse 2 korda aastas – mai kuus ja novembris. Uurimistöös kasutatakse töö kirjutamise hetkel kehtiv versioon, mis oli uuendatud viimase väljalaskega 30. novembril 2023 [6]. SNOMED CT eestikeelne versioon ei ole täiesti tõlgitud.

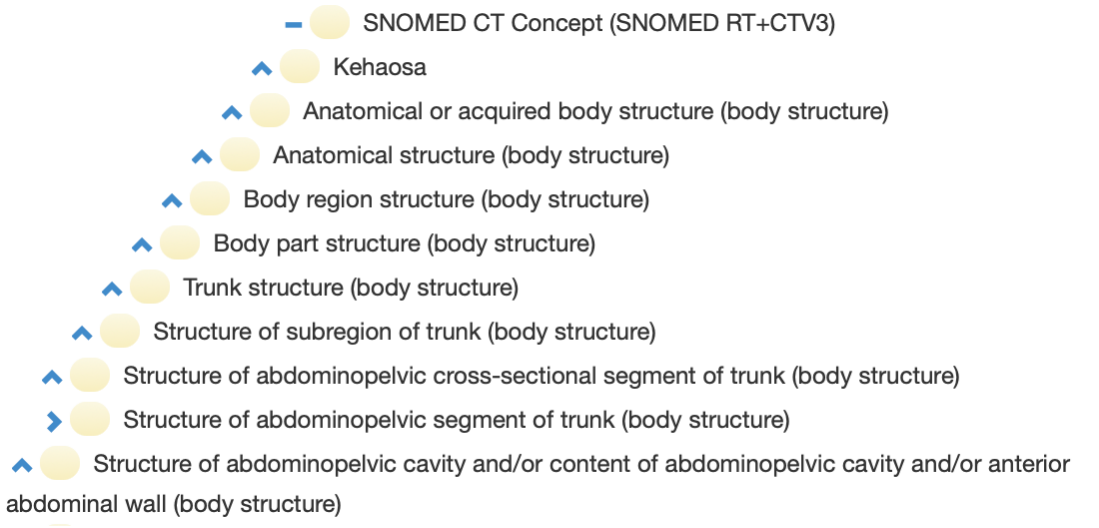
2.2 Tööriistad

2.2.1 SNOMED CT Browser

SNOMED CT brauser [6] on veebirakendus, mille abil on võimalik SNOMED CT taksonoomiat kasutada. Brauseris on leitav üle 400 000 meditsiinilise termini, mis on jaotatud 19 erineva kontsepti vahel. Vajalike väljendite otsimine on võimalik nii keeleliste kui ka semantiliste filtrite abil, mis muudab spetsiifiliste meditsiiniliste terminite leidmise lihtsamaks ja efektiivsemaks. See süsteem võimaldab meditsiinilist teavet otsida erinevate kategooriate, näiteks haiguste, meditsiiniliste protseduuride, kehaosade või sümptomite järgi. See võimaldab otsida mõisteid mitte ainult nende täpse nime, vaid ka seotud märksõnade või sünonüümide järgi. SNOMED CT lihtsustab

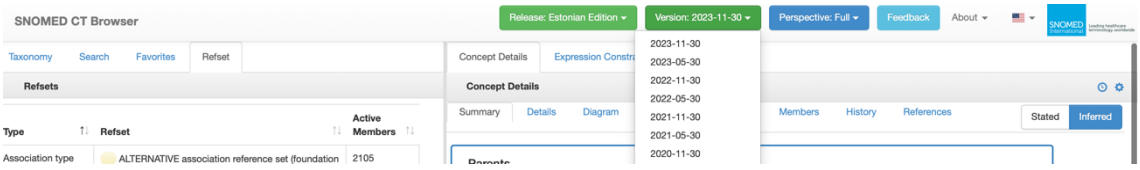
oluliselt õige termini leidmist, eriti kui täpne meditsiiniline nimetus on teadmata või unustatud [3].

Taksonoomia kasutab hierarhilist süsteemi, kus terminid on korraldatud ülalt alla, moodustades „vanemate“ (parent) ja „laste“ (children) suhted. Kasutajad saavad selle mõiste puu abil täpsustada oma otsingut, valides konkreetse kategooria, mis vastab kõige paremini nende päringule (Joonis 1).



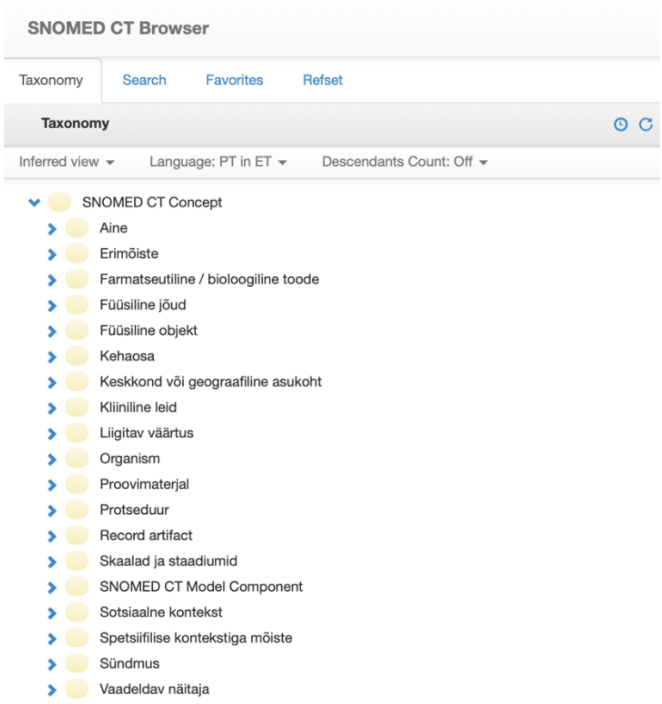
Joonis 1. SNOMED CT Browser hierarhiline süsteem

Samuti võimaldab brauser valida konkreetse keeleversiooni (Edition) ja väljalaskekuupäeva (Version). Keeleversioone uuendatakse kaks korda aastas (Joonis 2).



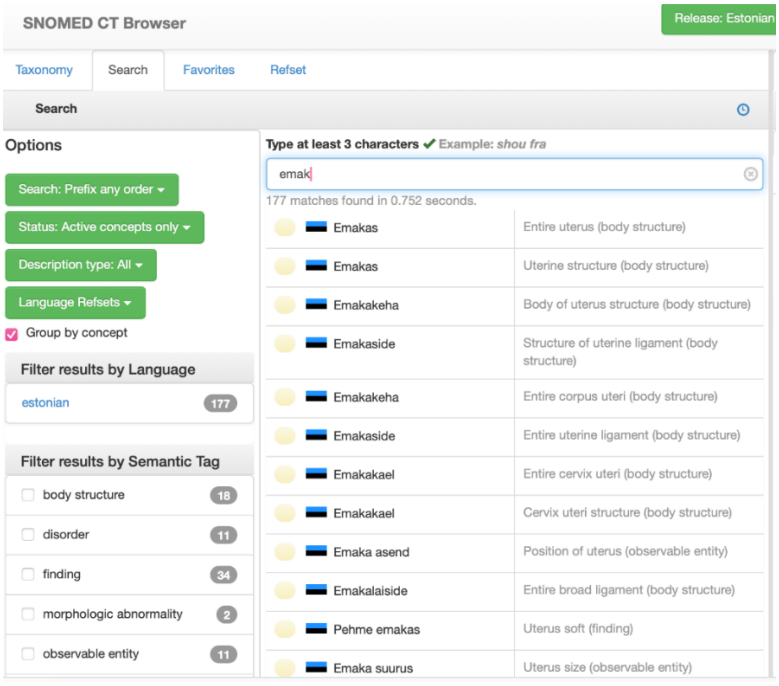
Joonis 2. SNOMED CT eestikeelse versiooni väljalaskekuupäevad

SNOMED CT brauseris on mitu vahelehte. Esimene neist on taksonoomia, mis sisaldab kirjeldatud kontsepte. Seda kasutatakse SNOMED CT struktuuri navigeerimiseks ja vaatamiseks, mis korraldab meditsiiniterminid hierarhilises järjekorras. See võimaldab kasutajatel hõlpsalt otsida ja vaadata spetsiifilisi meditsiinilisi kontseptsioone ja nende seoseid (Joonis 3).



Joonis 3. SNOMED CT Browser Taxonomy vaade

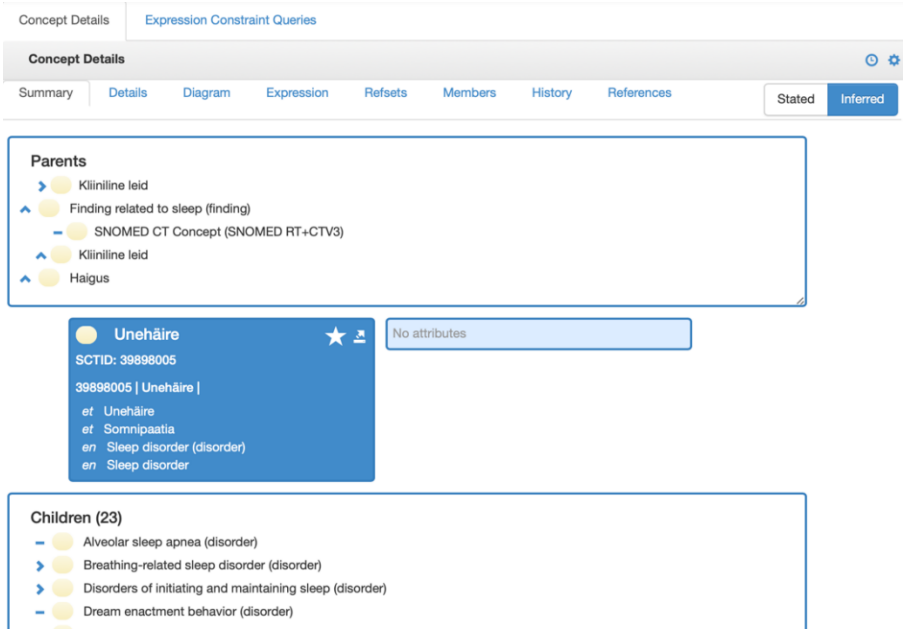
Teine vaheleht on „Search“, mis pakub tööriistu konkreetsete meditsiiniliste kontseptsioonide otsimiseks. Kasutajad saavad otsingut kohandada aktiivsuse staatuse, kirjelduse tüübi, keeleversioonide ja teiste parameetrite alusel. Samuti on saadaval valikud otsingutulemuste täpsustamiseks, mis lihtsustab vajaliku informatsiooni leidmist ulatuslikus meditsiinilises taksonoomias (Joonis 4).



Joonis 4. SNOMED CT Browser Search vaade

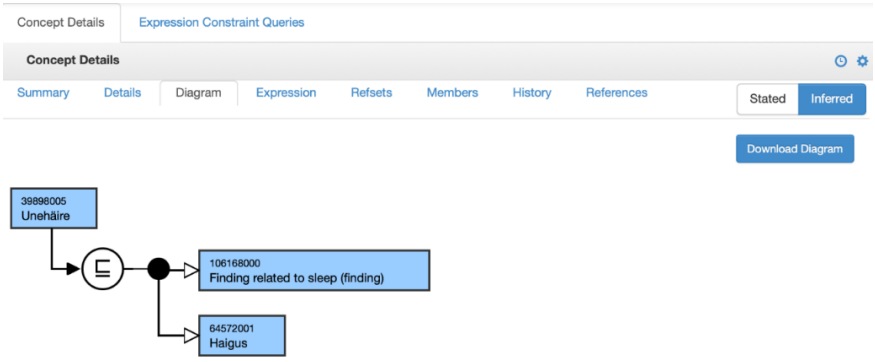
Veel on olemas „Favorites” ja „Refset” vahelehed, mis ei olnud kasutatud selles töös.

Brauseris on võimalik põhjalikult uurida iga konkreetse meditsiinilise kontseptsiooni detaile. Näiteks „Summary” vaheleht pakub kiiret ülevaadet põhiandmetest. Siit on näha iga kontseptsiooni unikaalne identifikaator SNOMED CT andmebaasis - SCTID. Ala „Parents“ näitab, millised teised kontseptid on selle kontsepti „vanemad”, ehk millistele ülemmõistetele see allub. Ala „Children“ loetleb „lapsed” ehk alamkontseptsioonid, mis on tuletatud uuritavast mõistest. Lisaks kuvatakse kontseptsiooni nimetuse inglise keeles. Kui kontseptsioonil on spetsiifilisi omadusi või täiendavaid andmeid, kuvatakse need atribuutide alla (Joonis 5).



Joonis 5. SNOMED CT Browser Summary vaade

Vaheleht „Diagram“ näitab graafilist kujutist kontseptsioonide suhetest ja seostest SNOMED CT andmebaasis (Joonis 6).



Joonis 6. SNOMED CT Browser Diagram vaade

Lisaks on olemas vahelehed „Details”, „Expression”, „Refsets”, „Members”, „History”, „References”, mis ei olnud käesolevas töös kasutatud.

2.2.2 ChatGPT

Kuna tehisintellekti valdkond, eriti ChatGPT platvorm [10], on viimasel ajal muutunud üha populaarsemaks ja laialdaselt kasutatavaks [11], otsustasin uurida selle tööriista potentsiaali oma uurimistöös. Proovisin ChatGPT abil automaatselt vastendada meditsiinilisi tekste SNOMED CT koodidega ning võrrelda neid käsitsi vastendamise tulemustega. Samuti kasutasin ChatGPT-d Pythoni skripti loomiseks, et koostada sagedussõnastikku ja tuvastada kõige sagedamini esinevaid sõnu ja fraase meditsiinilistes tekstides. Sooviks hinnata ja paremini mõista tehisintellekti võimekust ja piiranguid meditsiiniliste tekstide töötlemisel. Oma katsetuste jaoks ja tulemuste saamiseks kasutasin ChatGPT versiooni 4.0 [10].

2.2.3 Python

Minu diplomitöö käigus kasutasin Python programmeerimiskeele [12] sagedussõnastiku koostamiseks. Kuna mul ei ole kogemust Python skriptide kirjutamisel, otsustasin kasutada tehisintellekti abi skripti loomisel. Selleks koostasin selge ja täpse päringu, mille abil ChatGPT aitas mul luua vajaliku skripti.

See skript oli oluline tööriist sagedussõnastiku koostamisel, kus analüüsiti ja tuvastati sõnu ja fraase, mis esinesid kõige sagedamini vabatekstides meditsiiniliste ultraheliuuringute tulemuste kirjeldustes. Konkreetsemalt, skript töötas läbi üle 20 000 meditsiinilise tekstikirjelduse [13], eraldas neist individuaalsed sõnad ja fraasid ning koostas nende esinemissageduse nimekirja. Sagedussõnastiku koostamine võimaldas tuvastada võimalikke lünki ja probleeme SNOMED CT eestikeelses versioonis, näiteks sageli kasutatavate terminite tõlgete puudumine.

2.3 Protsess

Probleemi lahendamiseks kasutatakse juhtumianalüüsi metoodikat [14].

Töö algas praeguste teadusartiklite põhjaliku ülevaatega ja vabade meditsiiniliste tekstide vastendamise teemaga seotud uurimistööga. See võimaldas mul määratleda oma

uurimuse metoodilised alused ja mõista, millised lähenemisviisid ja vahendid oleksid kõige tõhusamad minu töö jaoks.

Järgmisena alustasin vastendamiseks vajalike lähteandmete kogumist. Oluliseks allikaks olid Tallinna Tehnikaülikooli professori dr Peeter Rossi tekstid. Need tekstid sisaldasid kõhu- ja vaagnapiirkonna ultraheliuuringute andmeid erinevatest vanuserühmadest ja soost patsientidelt.

Kogu andmekogumi hulgast valisin edasiseks analüüsiks välja 71 erinevat teksti [15]. Valim oli tehtud vanuserühmade kaupa [16] ja lisasin valikukriteeriumide hulka ka soo, et tagada kõige mitmekesisemate ja esinduslikemate meditsiiniliste tekstide kogum. Erinevad tekstid olid vajalikud selleks, et tuvastada võimalikult palju probleeme ja väljakutseid vastendamise käigus.

Pärast tekstide valimist alustasin vastendamise protsessi kahe erineva viisiga ja eraldi koostasin sagedussõnastiku.

2.3.1 Käsitsi vastendamise protsess

Käsitsi vastendamise eesmärgiks oli tuvastada ja analüüsida põhiprobleeme vastendamise protsessis. Selle jaoks olid sõnad ja fraasid väljavõetud vabadest meditsiinilistest tekstidest [16], mis sisaldasid kõhu- ja vaagnapiirkonna ultraheliuuringute tulemusi. Iga väljavõetud terminit otsiti SNOMED CT taksonoomia eestikeelses versioonis [6]. Kasutati spetsiaalset taksonoomia otsingumootorit, mis võimaldas leida mitte ainult sõnasõnalised vastavused, vaid ka sünonüümid või alternatiivsed tõlked.

Juhul, kui sobiv termin taksonoomias ei leitud, mille tõlkimine või tähendus oli ebaselge või ebatäpne, oli tähistatud spetsiaalse kategooriaga. See oli oluline edaspidise töö jaoks, et tuvastada lüngad taksonoomias ja võimalikud täiendused või parandused andmebaasis.

Vastendamise lõpus olid koostatud koondtabelid, mis sisaldasid kõiki andmeid terminite kohta, mida ei olnud leitud, ja nende kohta, mille tõlge oli ebakindel. Tabelid olid rühmitatud vastavalt vastendamise käigus tuvastatud erinevatele probleemidele.

Väljakutsete nimekirja valideerimiseks võrdlesin seda probleemidega, mida avastasin teaduskirjandusest. Samuti pöördusin doktor Peeter Rossi poole, kes oli mulle vajalikud tekstid edastanud, et saada tema professionaalset tagasisidet.

2.3.2 Tehisintellekti abil vastendamise protsess

Tehisintellekti [10] abil vastendamise eesmärgiks oli kontrollida selle tööriistu valmisoleku vabatekstide iseseisvaks vastendamiseks.

Selle jaoks, et teha vastendamist tehisintellekti abil, oli vaja välja selgitada, kas süsteem omab teavet SNOMED CT ja vabade meditsiiniliste tekstide vastendamise kohta.

Järgmisena oli valitud teksti spetsiifiliste lausete edastamine ChatGPT-le, palvega need vastendada vastavalt SNOMED CT taksonoomiale. Oluline on märkida, et tehisintellektile edastatud tekst ei sisaldanud patsiendi metaandmeid ehk sugu ja vanust. Oodati, et ChatGPT genereeriks vastavad SNOMED CT koodid iga meditsiinilise termini või väljendi kohta. Viimasena esitas ChatGPT vastendamise tulemused tekstiformaadis.

Järgnes tulemuste valideerimise etapp, kus võrdleti saadud SNOMED CT koode meditsiiniliste terminitega ning kontrolliti, kas esitatud koodid vastasid täpselt neile terminitele taksonoomias, et tagada vastendamise täpsus ja usaldusväärsus.

2.3.3 Sagedussõnastiku koostamine

Otsustati luua sagedussõnastiku, mis näitaks selgelt, milliseid sõnu arst kasutab kõige sagedamini kõhu- ja vaagnapiirkonna ultraheliuuringute kirjeldamisel. Selleks koostati skripti tehisintellekti abil, mis aitaks päringut lahendada. Kui skript oli kirjutatud ja töödeldud, see koostas vajaliku nimekirja kõige sagedamini kasutavate sõnade ja fraasidega.

Leitud sagedussõnastikku kasutati veel selleks, et tuvastada SNOMED CT eestikeelse versiooni tõlkimata termineid.

3 Peamised tulemused

See peatükk annab põhjaliku ülevaate kirjandusest, tekstide valiku protsessist, käsitsi ja tehisintellekti abil vastendamisest, sealhulgas leitud väljakutsetest ja valideerimisest, ning sagedussõnastiku koostamisest.

3.1 Kirjanduse ülevaade

3.1.1 SNOMED CT taksonoomia

SNOMED CT taksonoomia sisaldab üle 360 000 meditsiinilise termini, mis hõlmavad paljusid meditsiinilisi erialasid ja mõisteid. Taksonoomia on saadaval osaliselt tõlgitud versioonidena 18 keeles, mis teeb sellest väärtusliku ressursi rahvusvaheliste meditsiinikogukondade ja eri keelekontekstides tehtavate uuringute jaoks [17], [18].

Meditsiinilises taksonoomias on igal terminil unikaalne identifitseerimisnumber (SCTID). See tagab täpsuse ja järjepidevuse meditsiinilise informatsiooni vahetamisel erinevate süsteemide ja spetsialistide vahel.

Taksonoomia rakendab hierarhilist süsteemi, milles terminid on seotud „vanemate“ (parent) ja „laste” (children) suhetega. See võimaldab kasutajatel liikuda struktuuris ja paremini mõista erinevate meditsiiniliste mõistete vahelist konteksti ja seoseid.

SNOMED CT-s on määratletud kokku 19 põhikontseptsiooni millest igaüks kajastab konkreetset meditsiiniliste teadmiste valdkonda (Joonis 7):

1. Aine – jaotub erinevateks alamkategooriateks, mis kirjeldavad bioloogilisi aineid, ravimeid, diagnostilisi aineid, keha aineid ja muid materjale.

2. Erimõiste – see kontsept hõlmab erinevaid meditsiinilisi leiundeid või situatsioone, mis on tavalisest erinevad või haruldased, eriti seoses meditsiiniliste protseduuride või uuringutega. Näiteks anomaalsed bioloogilised, tsütoloogilised, histoloogilised leiud.

3. Farmatseutiline/bioloogiline toode – selline kontsept katab laia valiku meditsiinilisi ja bioloogilisi tooteid, mida kasutatakse tervishoius.

4. Füüsiline jõud – kontsept hõlmab mitmesuguseid füüsikalisi jõude ja nähtusi, mida kasutatakse meditsiinilises kontekstis nende mõju tõttu inimese kehale või tervisele.

5. Füüsiline objekt – kontsept käsitleb erinevaid esemeid, mida võib leida igapäevaelus või meditsiinilises keskkonnas ja mis võivad olla olulised tervishoiu kontekstis. Näiteks „seade“.

6. Kehaosa – kontsept sisaldab kategooriaid, mis on seotud inimkeha struktuuride ja nende anatoomiliste iseärasustega.

7. Keskkond või geograafiline asukoht – see kontsept hõlmab erinevaid kategooriaid ja alamkategooriaid, mis on seotud keskkondade ja geograafiliste asukohtadega.

8. Kliiniline leid – kontsept hõlmab laia valikut meditsiinilisi tulemusi ja vaatlusi, mis on seotud patsiendi tervisliku seisundiga. Näiteks „krooniline paranoidne psühhoos“ või „vanus alla 25 aasta“.

9. Liigitav väärtus – kontsept käsitleb erinevaid kategooriaid, mis on seotud tervishoius kasutatavate andmete klassifitseerimisega. Näiteks „määramatu“ ja „keskmine“.

10. Organism – kontsept keskendub elusolendite, peamiselt mikroorganismide süstematiseerimisele.

11. Proovimaterjal – kontsept viitab erinevatele bioloogilistele ja mittebioloogilistele näidistele, mida kasutatakse meditsiinilisteks uuringuteks ja analüüsideks. Näiteks „koematerjal“ või „tsütoloogiline materjal emakakaelast“.

12. Protseduur – kontsept kirjeldab erinevaid tegevusi ja sekkumisi, mida tervishoius kasutatakse haiguste diagnoosimiseks, ravi määramiseks või tervise jälgimiseks. Näiteks „ajutise südamestimulaatori paigaldamine“ ja „endoskoopia ja hemorraagia peatamine“.

13. Record artifact – kontsept käsitleb erinevaid meditsiiniliste ja tervishoiu dokumentide osi ja vorme. Näiteks hooldusplaanid.

14. Skaalad ja staadiumid – kontsept käsitleb erinevaid klassifikatsioonisüsteeme ja hindamisskaalasid, mida kasutatakse meditsiinis haiguste, seisundite ja sümptomite süstematiseerimiseks ja hindamiseks.

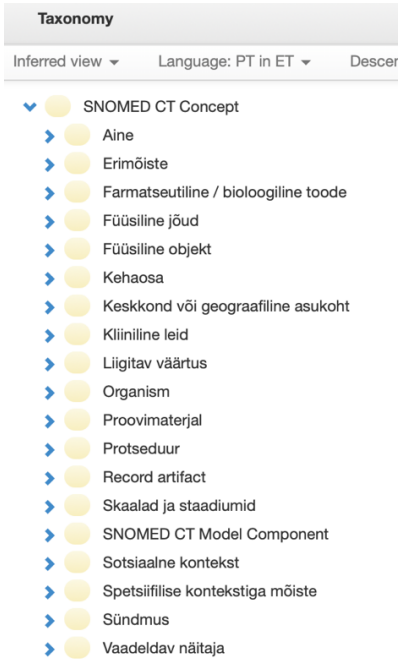
15. SNOMED CT Model Component – kontsept on oluline taksonoomia struktuuri ja haldamise jaoks. Ta pakub metadata struktuure ja määratlusi, mis aitavad klassifitseerida ja hallata erinevaid meditsiinilisi andmeid ja nende suhteid.

16. Sotsiaalne kontekst – hõlmab inimeste sotsiaalset staatust, eluviisi ja sotsiaalseid suhteid, mis võivad mõjutada nende tervist ja tervisekäitumist. Näiteks „eakad vanemad“ või „eluviis“.

17. Spetsiifilise kontekstiga mõiste – kontsept hõlmab mitmesuguseid stsenaariume, kus patsiendi tervislikku seisundit või meditsiinilist hoolitsust mõjutavad teatud spetsiifilised kontekstid või asjaolud.

18. Sündmus – kontsept sisaldab nii spetsiifilisi kui ka üldisi sündmusi, mis mõjutavad inimesi või keskkonda. Näiteks „surm“.

19. Vaadeldav näitaja - hõlmab erinevaid meditsiinilisi mõõdetavaid ja jälgitavaid näitajaid, mida kasutatakse diagnostikas ja ravi jälgimises.



Joonis 7. SNOMED CT taksonoomia kontseptid

3.1.2 Kasutuse võimalused

Süstemaatiline ülevaade [7] näitas, et SNOMED CT'd on võimalik kasutada erinevate ülesannete lahendamiseks:

- tekstidest vajaliku teabe ekstraheerimiseks,
- tekstide vastendamiseks SNOMED CT koodidega,
- terminite sünonüümide leidmiseks,
- meditsiiniliste dokumentide kvaliteeti hindamiseks,
- andmete sarnasuse hindamiseks, kasutades SNOMED CT terminite seoseid,
- tekstide klassifitseerimiseks,
- vastendamiseks teistesse terminoloogiatesse,
- muud.

3.1.3 Vastendamise meetodid

Läbivaadatud kirjanduses meditsiiniliste tekstide vastendamiseks SNOMED CT taksonoomiaga on kõige sagedamini kasutatud käsitsi vastendamine, loomuliku keele töötlemine (ehk Natural Language Processing, NLP), masinõpe (ML), reeglipõhine vastendamine (*rule based*) ja nende hübriidid. [7] annab põhjaliku ülevaade nendest meetodist ja nende tulemuslikkusest.

Käsitsi vastendamisega kodeeriti gastrektoomiaga patsientide meditsiiniliste tekste. Kuni 97% termineid oli vastendatud SNOMED CT terminitega [2].

[8] kirjeldab reeglipõhist lähenemist kopsuvähi staadiumi määramiseks. Selle töö raames suutsid autorid saavutada vastendamisel 73%-94% täpsust (*accuracy*).

NLP abil saavutasid autorid 67%-77% täpsust (*precision*) ja 73%-92% saagist (*recall*) [1].

3.1.4 Vastendamisel avastatud väljakutsed

Ülevaadatud kirjanduses on mainitud ka mõned vastendamisel avastatud väljakutsed. Tihti mainitakse järgmised:

- eituste tuvastamine [3], [8] — eitust võib väljendada mitu moodi: ’no’, ’no evidence of’, ’none of’, ’clear of’ [8],
- tekstides lühendite kasutamist [3] — näiteks, ’BMI’ tähendusega *body mass index*,
- sünonüümide kasutamist [1], [3] — näiteks, *heart attack* või *myocardial infarction*,
- sõnaliigi tuvastamist [1] — sama sõna võib kasutada nii tegusõnana, kui ka nimisõnana, näiteks *research*.

Võib ka juhtuda, et mingit terminit on võimalik vastendada osaliselt või ainult üldisema või täpsema terminiga [2].

Eraldi tähelepanu vajab SNOMED CT, kui taksonoomia, vasturääkivus. Näiteks, *GOT (Glutamic Oxoloacetic Transaminae)* ensüüm on saadaval ainult *Kliiniline leid* kategoorias (*“aspartate transaminase level (finding)“*), vaid GPT (Glutamic Pyruvic Transaminase) ensüüm on saadaval ainult *Protseduur* kategoorias (*“alanine aminotransferase measurement (procedure)”*) [2].

3.1.5 Vastendamise tööriistad

SNOMED CT terminite vastendamiseks on loodud mitu tööriista: Medtex [19], Medical Text Extraction Reasoning and Mapping System (MTERMS) [20], MetaMap [21], Mayo Clinic Vocabulary Server (MCVS) [22], clinical Text Analysis and Knowledge Extraction System (cTAKES) [23] ja muud.

Medtex, MTERMS ja MCVS ei ole avalikult kättesaadavad. MetaMap ja cTAKES on vabavaralised, aga ükski neist ei ole SNOMED CT taksonoomiaks spetsialiseeritud ega paku mitmekeelset tugi. Muude rakenduse puhul puuduvad andmeid selle kohta, kui hästi nende abil on võimalik tekste SNOMED CT terminitega vastendada [7].

3.1.6 Uurimise hetkeseis

Praegu on vähenenud uurimistegevus vabade meditsiiniliste tekstide ja SNOMED CT taksonoomia koodide vastavusse viimise valdkonnas. Selle põhjuseks on mitmed varasemates uuringutes tuvastatud keerukused, näiteks erinevused spetsialistide kirjeldustes, kooskõlastusprobleemid, kõrged valideerimiskulud ja keelepiirangud. Need raskused tekitavad teadlastele märkimisväärseid takistusi ja pidurdavad valdkonna arengut. Neist takistustest hoolimata, autorid arvavad et on vaja teha edasisi uuringuid, et töötada välja tõhusamad vastendamismeetodid ja ületada tuvastatud probleemid. See avab ukse tulevastele teadusuuringutele, et töötada välja uusi lähenemisviise ja vahendeid meditsiinilise teksti vastendamise protsessi parandamiseks SNOMED CT-s [1], [2], [8].

3.2 Tekstide valimine

Töö jaoks olid edastatud üle 20,000 teksti [13] kõhu- ja vaagnapiirkonna ultraheliuuringu tulemuse kohta, mis olid koondatud Exceli tabelisse. Tabel sisaldas informatsiooni protseduuri nimetuse, patsientide metaandmete, uuringu tellija ning ultraheli kirjelduste kohta (Joonis 8).

	A	B	C	D	E	F
1	Uuring	ACK	PT vanus	PT sugu	Tellija	Vastus
2	7953 Kõhu- ja vaagnapiirkonna ultraheliuuring (UH)	HITKUH10 91231025	88a.	N	I Kardioloogia osakond	Kõhu- ja vaagnapiirkond Kliinilised andmed: kõhuvalu üla ja keskkõhus. Kõhukoopa UH uuring. LEID: Pankreas atroofiline, eakohase struktuuriga, mahulist muutust esile ei tule. Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta. Sapipõie seinad õhukesed, valendik vaba. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, koldeleiuta. Parem neer normaalse asendi, suuruse ja kujuga, paisu ei sedasta. Vasak neer opereeritud. Kusepõis vähe täitunud, sisaldis kajavaba. Kõhuaort nähtavas osas norm läbimõõduga, verevool tavaline. Retroperitoneaalselt suurenenud l/sõlmi ei näe. Vaba vedelikku kõhukoopas ei sedasta. Vasakul alakõhus, kus patsien kaebab valulikkust näha gaasiga täitunud sooleringe. Peensool laienenud ei ole, peritaltika jälgitav. KOKKUVÕTE: aktuaalse patoloogilise leiuta.
3	7953 Kõhu- ja vaagnapiirkonna ultraheliuuring (UH)	HITKUH10 91231021	78a.	N	Erakorralise meditsiini osakond Ravi 18	Kliinilised andmed: Neerupais? Valu vasakul pool kõhus. Kõhukoopa UH uuring. LEID: Pankreas mõõtmetelt normis, eakohase struktuuriga, mahulist muutust esile ei tule. Maks mõõtmetelt normis, veidi sõmerja struktuuriga, koldeleidu ei sedasta. Sapipõis ei tule selgelt nähtavale, annab tugeva kajavarju - valenikus ilmselt kivid. Sapiteed laienenud ei ole. Põrn mõõtmetelt kergelt suurenenud, koldeleiuta. Neerude suurus, asetsus ja struktuur tavaline. Parem neeruvaagen laiusega 1.5cm, karikad olulise laienemiseta. Vasaku neeru kogumissüsteem laienenud, karikad ~1.2cm. Ureeter ülaosas ~4mm, ureeteris konkrementi ei visualiseeru, põie poolselt ureeter ei visualiseeru. Neeru ülapooluses kahtlus 5mm kivile. Kusepõis vähe täitunud, sisaldis kajavaba. Kõhuaort nähtavas osas norm läbimõõduga, verevool tavaline. Vaba vedelikku kõhukoopas ei sedasta. KOKKUVÕTE: Maks kergelt sõmerja struktuuriga, sapipõies ilmselt kivid. Mõlema neeru kogumissüsteem kergelt laienenud sin>dex, vasaku neeru ülapoolusel kahtlus kivile.
4	7953 Kõhu- ja vaagnapiirkonna ultraheliuuring (UH)	HITKUH10 91231023	25a.	N	Erakorralise meditsiini osakond Ravi 18	Kliinilised andmed: Valu paremas küljes. Sapikivid? Kõhukoopa UH uuring. LEID: Pankreas mõõtmetelt normis, eakohase struktuuriga, mahulist muutust esile ei tule. Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta. Sapipõie seinad õhukesed, valendik vaba. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, koldeleiuta. Neerude suurus, asetsus ja struktuur tavaline. Neerudes paisu ei ole. Kusepõis vähe täitunud, sisaldis kajavaba. Emakas ja munasarjad visuaalselt iseärasusteta. Kõhuaort nähtavas osas norm läbimõõduga, verevool tavaline. Retroperitoneaalselt suurenenud l/sõlmi ei näe. Vaba vedelikku kõhukoopas ei sedasta. Apendiks nähtavale ei tule. KOKKUVÕTE: aktuaalse patoloogilise leiuta.
5	7953 Kõhu- ja vaagnapiirkonna ultraheliuuring (UH)	HITKUH10 91231019	88a.	N	II Kardioloogia osakond	Kõhu- ja vaagnapiirkond Kliinilised andmed: FA parox. Kõhuvalud. Kõhukoopa UH uuring. LEID: Pankreas mõõtmetelt normis, eakohase struktuuriga, mahulist muutust esile ei tule. Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta. Sapipõie seinad õhukesed, valendik vaba. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, koldeleiuta. Neerude suurus, asetsus ja struktuur tavaline. Neerudes paisu ei ole. Kummaski neerus on parapelvikaalsed tsüstid ja kortikaaltsüstid: paremal 3,1 cm ja 2,6 cm; vasakul 4,9 cm. Kusepõis vähe täitunud, sisaldis kajavaba. Kõhuaort nähtavas osas norm läbimõõduga, verevool tavaline. Retroperitoneaalselt suurenenud l/sõlmi ei näe. Vaba vedelikku kõhukoopas ei sedasta. UH kilpnäärmest. Leid: parem sagar 1,3 x 1,5 x 3,8 cm, maht 3,9 ml; vasak sagar 1 x 1,5 x 3,2 cm, maht 2,5 ml; isthmus on 3 mm paksusega. Kilpnäärme kogumaht 6,4 ml. Kilpnäärme mõlemas sagaras on mõned väikesed kajavaesed kolded 2...5 mm. Vaskularisatsioon normipärane. Suurenenud l/sõlmi kaelal esile ei tule. -Neerude tsüstid. -Kilpnäärme maht normis, mõlema sagara tsüstid.
6	7953 Kõhu- ja vaagnapiirkonna ultraheliuuring (UH)	HITKUH10 91231018	47a.	M	Ravikindlustamatu te osakond Magasini 34	Kõhu- ja vaagnapiirkond Kliinilised andmed: Muutusi pankreases? Sapikive? Kõhukoopa UH uuring. LEID: Pankrease peaosas mitmed 2- 2,6 cm tsüstjad struktuurid, kehaosas 1,5 cm tsüst ja pankreasest veidi ülevalt va 6 cm piklik lubistunud kapsliga tsüst . Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta. Sapipõie seinad õhukesed, valendik vaba. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, koldeleiuta. Neerude suurus, asetsus ja struktuur tavaline. Vasakus neerus ca 2 cm kortikaaltsüst.Neerudes paisu ei ole. Kusepõis vähe täitunud, sisaldis kajavaba.Prostata ii. Kõhuaort ning iliakaalsed veresooned nähtavas osas norm läbimõõduga, verevool tavaline. Retroperitoneaalselt suurenenud l/sõlmi ei näe. Vaba vedelikku

Joonis 8. Edastatud tekstid lähteformaadis

Esitatud tekstide seast valisin välja 71, määrates esmalt vanusekategooriad tekstide jaotamiseks vastavalt elukaarele [16]. Seejärel filtreerisin vanusekategooriad, alustades meestest ja seejärel naistest. Iga vanusekategooria ja soo kohta valisin välja neli teksti, erandiks olid 1-3-aastased lapsed, kellest valisin ainult kaks naissoost teksti ja 3-7-aastased lapsed, kellest valisin viis teksti kuna rohkem andmeid polnud saadaval. Püüdsin

valida mitmekesiseid tekste, et leida võimalikult palju erinevaid väljakutseid vastendamise käigus. (Joonis 9).

Tekstide jaotus:
1) imikuiga (0-1a) – 8 valitud teksti;
2) maimikuiga (1-3a) – 2 valitud teksti;
3) koolieelikuiga (3-7a) – 5 valitud teksti;
4) kainikuiga (7-12a) – 8 valitud teksti;
5) murdeiga (12-16a) – 8 valitud teksti;
6) noorukiiga (16-20a) – 8 valitud teksti;
7) täiskasvanuiga (20-35a) – 8 valitud teksti;
8) hiline küpsus (35-55a) – 8 valitud teksti;
9) elatanuiga (55-70a) – 8 valitud teksti;
10) vanuriiga (70-90a) – 8 valitud teksti.

Pärast nende valimist koostasin Exceli tabeli [15], kus oli selgelt näha, kuidas tekstid kategooriate järgi jaotuvad.

Teksti number	Vanus	Sugu	Tekst
	Imikuiga (0-1a)		
1	0k.	M	Kliinilised andmed: Antenataalselt nähtud neeruvaagnate laienemist. Kõhukoopa UH uuring. LEID: Maks ealiste mõõtmetega 46mm, sapipõis täitunud. Sapiteed laienenud ei ole. Põrn ealise suuruse ja struktuuriga. Neerude suurus ja struktuur tealine. Paremas neerus paisu ei ole, vasakul karikad 4mm ja vaagen 7mm. Kusepõis täitunud, sisaldis kajavaba. Retroperitoneaalselt suurenenud l/sõlmi ei näe.
2	0k.	M	Kliinilised andmed: Paremal kubemesong? Dünaamika. Kõhukoopa UH uuring. LEID: Maks ealiste mõõtmetega, ühtlase struktuuriga, 43 mm. Sapipõie seinad õhukesed, sapiteed laienenud ei ole. Põrn ealise suuruse ja struktuuriga. Neerude suurus, asetsus ja struktuur ealine, parem neer 49mm, vasak neer 44 mm, vaagnate laienemist nähtavale ei tule. Kusepõis mõõdukalt täitunud. Vaba vedelikku kõhukoopas ei sedasta. PINDMISTE KUDEDE UURINGUL paremal kubemesong dünaamikata. Songakott 1,7 x 0,5 cm. Songavärat 6mm. Testist kubemekanalis nähtavale ei tule.
3	6p.	M	Kliinilised andmed: Loote UH-s vas. neeruvaagna laienemine? Kõhukoopa UH uuring. LEID: Pankreas esile ei tule. Maks ealiste mõõtmetega (58mm) Sapipõis kontraktsioonis. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, 36 mm. Neerude suurus, asetsus ja struktuur tavaline. Neerudes paisu ei ole. Parem 41 mm, vasak 45mm. Kusepõis vähe täitunud, sisaldis kajavaba. Retroperitoneaalselt suurenenud l/sõlmi ei näe. Vaba vedelikku kõhukoopas ei sedasta.
4	1k.	M	Kliinilised andmed: 1-l uuringul neeruvaagnate laienemine. Dünaamika. Kõhukoopa UH uuring. LEID: Pankreas ega kõhuaort ei visualiseeru (soolegaas). Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta. Sapipõie seinad õhukesed, valendikus veidi kajarikkamat sisaldist- pigem sade. Sapiteed laienenud ei ole. Põrna pikimõõde 5,1 cm, koldeleiuta. Neerude suurus, asetsus ja struktuur tavaline. Kummaski neerus on kogumissüsteemi laienemine: paremal karikad a 2-4 mm, vaagen 1,3 cm, vasakul karikad a 4 -5 mm, vaagen 1,6 cm. Samas ureeterite ülaosad veenvalt ei visualiseeru. Kusepõis siledaseinaline, sisaldis kajavaba. Vaba vedelikku ei tuvasta. Võrdluseks 20.04. UH uuring- nüüd neerude kogumissüsteemid mõnevõrra laiemad. ---Mõlema neeru kogumissüsteemi laienemine- võrreldes varasema uuringuga ebasoodne dünaamika. Samas ureetereid veenvalt ei erista. Leiu alusel ei välista PU stenoose. Vajaks uroloogi konsult. Põrn normi ülapiiril.
5	0k.	N	Kliinilised andmed: Neeruvaagnade laienemine. Kõhukoopa UH uuring. LEID: Maks mõõtmetelt normis (parema sagara pikimõõt ca 5 cm). Maks ühtlase struktuuriga, koldeleidu ei sedasta. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole (pikimõõdus 4,8 cm), koldeleiuta. Neerude suurus, asetsus ja struktuur tavaline. Paremal neeruvaagen 2 mm. Vasemal neeruvaagen 7 mm, karikad laienenud ei ole, samuti pole ureeter nähtav. Kusepõie sisaldis kajavaba. KOKKUVÕTE: Vasema neeruvaagna laienemine.
6	3k.	N	Kliinilised andmed: Sünni järel kahtlus vasakpoolsele dupleksneerule - diagnoosi täpsustamiseks. Sünnijärel mõlema külgvatasakese eessarve osas. Kahtlus septile- dünaamika hindamiseks. Kõhukoopa UH uuring. LEID: Maks ealiste mõõtmetega, ühtlase struktuuriga. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, 43mm. Neerude suurus, asetsus ja struktuur tavaline. Neerudes paisu ei ole. Veenvat dupleksneeru vasakul ei erista. Parem 43mm, vasak 46mm. Kusepõis vähe täitunud, sisaldis kajavaba. Aju külgvatsakeste eessarvad ümarate tippudega, tsüste nähtavale ei tule.Parem 6mm, vasak 6mm, III vatsake 4mm. Plexused ühtlase kajaga, sümmeetrilised. IV vatsake 6 mm. Corpus callosum hästi väljendatud. Periventrikulaarne kajalisus on tavaline.Taalamused ühtlase kajaga. Fissuur 6mm, subarahnoidaalruum 4mm.
7	22p.	N	Kliinilised andmed: 3 nädalane tütarlaps, kellel harv roojamine. Kõht puhitunud, väljaheide tihe, lihtjas. M.Hirschsprungi? Kõhukoopa UH uuring. LEID: Pankreas mõõtmetelt normis, eakohase struktuuriga, mahulist muutust esile ei tule. Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta. Sapipõie seinad õhukesed, valendik vaba. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, koldeleiuta. Neerude suurus, asetsus ja struktuur ealised. Neerudes paisu ei ole. Kusepõis vähe täitunud, sisaldis kajavaba. Kusepõie seinad suhteliselt paksud - tühja põie korral mõõdetav ca 7 mm seina paksus. Vaba vedelikku kõhukoopas ei sedasta. Rohkelt soolegaasi üle kogu kõhu. Kokkuvõte: Rohkelt soolegaasi üle kogu kõhu.
8	3p.	N	Kliinilised andmed: Antenat. nähtud bilat. neeruvaagnate laienemist. Kõhukoopa UH uuring. LEID: Pankreas normaalsekuju,kajalisuse ja suurusega. Maks mõõtmetelt normis, ühtlase struktuuriga, normaalse kajalisusega. Sapipõie seinad õhukesed, valendik vaba. Maksa veenides ja portas normaalne õige suunaga vool. Sapiteed laienenud ei ole. Põrn mõõtmetelt normaalne (5cm pikkus). Neerude suurus (mõlemal pool umbes 5cm), asetsus ja struktuur tavaline. Neerudes paisu ei ole. Põikisuunas neeruvaagen paremal 5mm ja vasakul 3mm, karikad ei ole eristatavad. Leid vastab ealisele normile aga vähese ebasümeetria pärast soovitav kontrolluuring umbes 1 kuu pärast. Kusepõis tühi. Neerupelised tavalised. Kõhuaordist nähtav vaid ülaosa, läbimõõt ja vool tavaline. Vaba vedelikku kõhukoopas ei sedasta. - Neerude kogumissüsteem vastab ealisele normile aga vähese ebasümeetria pärast soovitav kontrolluuring umbes 1 kuu pärast.

Joonis 9. Näide imikuea kategooriast valitud tekstidest

3.3 Käsitsi vastendamine

Pärast tekstide valiku tegemist alustasin nende käsitsi vastendamisega. Selleks tuli esmalt tekstid sõnadeks või fraasideks jaotada. Lõin Wordi dokumendi [24], kus koostasin kõigist 71 valitud tekstist tabelid kolme veeruga: esimeses veerus paiknes tekst ise, mille mina seejärel sõnadeks ja väljenditeks jaotasin. Teist veergu kasutasin SNOMED CT taksonoomiast pärinevate kontseptsioonide ja koodide märkimiseks. Kolmandas veerus hoidsin ruumi täiendavatele kommentaaridele või märkustele, mis mul vastendamise käigus tekkisid. Sinna sai lisada ka märksõnu, mille alusel hiljem vastendamise käigus ilmnenud probleeme grupeerisin.

3.3.1 Teksti vastendamise näide

Seejärel tegelesin otseselt käsitsi vastendamisega. Otsisin konkreetsetele sõnadele ja väljenditele vastavaid koode taksonoomiast ja täitsin tabelid. Siin on konkreetse teksti vastendamise näide (vt Tabel 1).

Tabel 1: Teksti vormistamine vastendamiseks

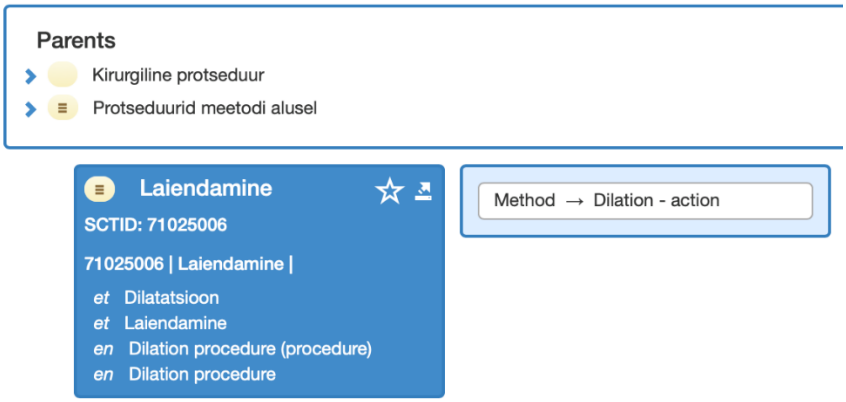
Originaalne tekst 3:	**SNOMEDi koodid+kontsept**	**Probleemid/märkused**
Kliinilised andmed: Loote UH-s vas. neeruvaagna laienemine? Kõhukoopa UH uuring. LEID: Pankreas esile ei tule. Maks ealiste mõõtmetega (58mm) Sapipõis kontraktsioonis. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, 36 mm. Neerude suurus, asetsus ja struktuur tavaline. Neerudes paisu ei ole. Parem 41 mm, vasak 45mm. Kusepõis vähe täitunud, sisaldis kajavaba. Retroperitoneaalselt suurenenud l/sõlmi ei näe. Vaba vedelikku kõhukoopas ei sedasta.		
Kliinilised andmed:	404684003 Kliiniline leid	Kas sama?
Loote UH-s vas. neeruvaagna laienemine?		
Loote	83418008 Loode, tervik	kääne
UH-s	16310003 Ultraheliuuring	lühend
vas.	7771000 Vasak	lühend
neeruvaagna laienemine?	362221007 25322007 Neeruvaagen, tervik; Laiendamine	kääne+kirjavahemärk

3.3.2 Leitud väljakutsed

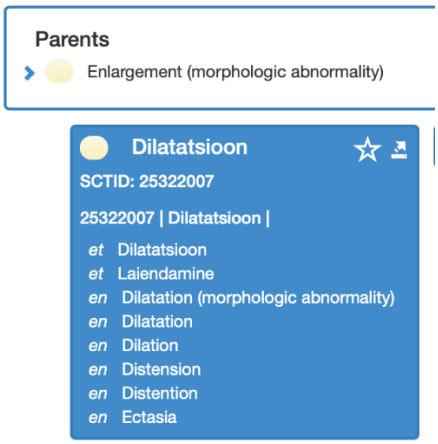
Vabade meditsiiniliste tekstide käsitsi vastendamisel SNOMED CT taksonoomiakoodidega puutusin kokku mitme raskusega.

Näiteks sõna „laiendamine“ võib mõista kui millegi laiendamist kirurgilise protseduuri mõttes, siis on selle kood asjakohane SCTID: 71025006 (Joonis 10). Ja seda sõna võib kasutada ka kirjeldamaks mõnda normist kõrvalekaldumist, ehk siis SCTID: 25322007 ja sünonüümne terminiga Dilatatsioon (Joonis 11).

Sisuliselt kaks identset sõna, kuid SNOMED-taksonoomia eripära tõttu võib vastendamisel tekkida raskusi, olenemata kontekstist. Ilma selge arusaamata kontekstist võib sõna olla seostatud kas kliinilise leiuga või kirurgilise protseduuriga.

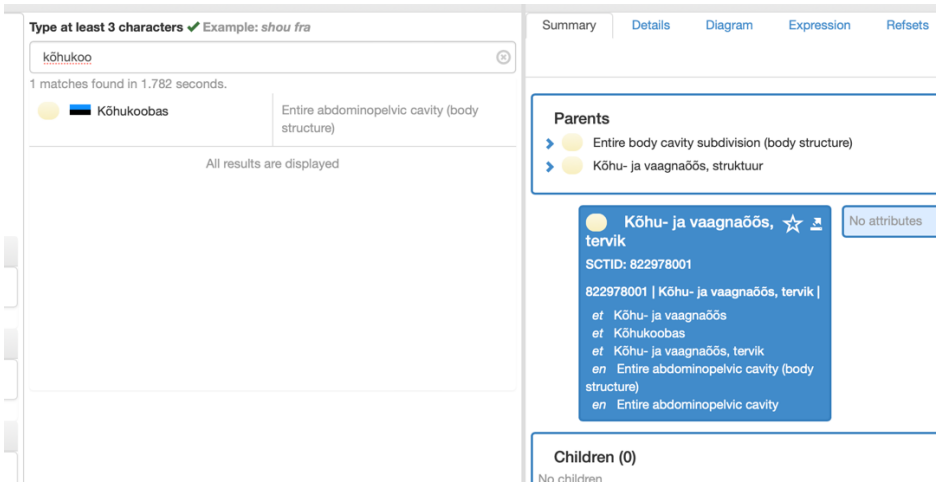


Joonis 10. Laiendamine kui kirurgiline protseduur



Joonis 11. Laiendamine kui normist kõrvalekaldumine

Kui otsida terminit „kõhukoobas“, annab otsingumootor tulemuseks „Kõhu- ja vaagnaõõs, tervik“. SNOMED CT-s on need terminid tähistatud sünonüümidena, kuid nad on erinevad meditsiinilisest vaatenurgast (Joonis 12).



Joonis 12. Kõhukoopa sünonüüm

Näiteks fraas „Kõhukoopa UH uuring“ koosneb kolmest sõnast, mida tuleb vaadelda eraldi. Kõhu- ja vaagnapiirkonda tähistatakse SCTID: 822978001, kuid otsides seda terminit taksonoomias täpselt nii, nagu see on tekstis kirjutatud, ei ole võimalik seda leida (Tabel 2). On vajalik lihtsustada „kõhukoopa“ sõna „kõhukoo“-ni, et leida selle nimetavas käändes vorm „Kõhukoobas“. Paljud sõnad meditsiinilistes tekstides on algsest vormist erinevas käändes. SNOMED CT andmebaasis otsivate arvutisüsteemide jaoks tekitas see raskusi, sest optimaalne otsing eeldab sõnade kasutamist nende algkujul - nimetavas käändes.

Fraasis „UH uuring“ on lühend „UH“, mis tähendab ultraheli, seega pean otsima „ultraheliuuring“ taksonoomiast, mille SCTID on 16310003 (Tabel 2). Otsingumootorid ei tunne enamasti ära meditsiinilistes tekstides sageli kasutatavaid lühendeid. Selle põhjuseks on see, et SNOMED CT algoritmid ei ole alati võimelised lühendeid ilma kontekstita õigesti tõlgendama.

Nii saab kokku kaks SCTID-d, mida on võimalik ühiselt vastendada.

Tabel 2: Kõhukoopa ultraheliuuringu vastendamine.

Originaalne tekst 3:	**SNOMEDi koodid+kontsept**	**Probleemid/märkused**
Kõhukoopa UH uuring		
Kõhukoopa	822978001 Kõhu- ja vaagnaõõs, tervik	kääne
UH uuring	16310003 Ultraheliuuring	lühend

Näiteks fraas „ealiste mõõtmetega“ esineb tihti, kuid seda ei leia SNOMED CT taksonoomiast. Katsetuste käigus õnnestus mul leida sünonüümid „Eakohane kasv ja areng“, millele vastab SCTID: 23397005. Kuna ma ei olnud täiesti kindel, et selline vastendus on parim võimalik vastavus, lisasin selle „Kas sobib?“ kategooriasse, et näidata, et valik võib vajada edasist kinnitust või ülevaatamist (Tabel 3).

Tabel 3: Leidmatu fraasi sünonüümiga asendamine.

Originaalne tekst 3:	**SNOMEDi koodid+kontsept**	**Probleemid/märkused**
Maks ealiste mõõtmetega (58mm)		
Maks	181268008 Maks, tervik	
ealiste mõõtmetega	23397005 Eakohane kasv ja areng	Kas sobib?

Näiteks, fraas „neerude“ esineb sageli kõhu- ja vaagnapiirkonna ultraheliuuringute kirjeldustes. SNOMED CT taksonoomias ei ole olemas üldist terminit, mis viitaks mõlemale neerule korraga, mistõttu oli vajalik fraas „neerude“ jagada kaheks eraldi terminiks: paremaks ja vasakuks neeruks. Selline jaotus võimaldab täpsemat vastendamist SNOMED CT süsteemis, kus parema neeru kood on 362208000 ja vasaku neeru kood on 362209008 (Tabel 4). See lähenemine tagab, et ultraheliuuringute andmed on selgelt määratletud ja vastavad rahvusvahelistele meditsiinilistele standarditele.

Tabel 4: Neerude vastendamine.

Originaalne tekst 3:	**SNOMEDi koodid+kontsept**	**Probleemid/märkused**
Neerude suurus, asetsus ja struktuur tavaline.		
Neerude	362208000 362209008 Parem neer, tervik; Vasak neer, tervik	kääne
suurus, asetsus ja struktuur tavaline	30389008 Normaalne kude	Kas sobib?

Näiteks mõned sõnad ja fraasid, nagu „karikad“, „sisaldis kajavaba“, „mahulist muutust esile ei tule“ ja „soolegaas“, ei ole esindatud eestikeelses versioonis SNOMED CT-s. Selle tulemusena tekkis vajadus grupeerida sellised terminid, mida SNOMED CT ei hõlma, eraldi kategooriasse. Ma nimetasin selle kategooria „Puudub SNOMEDist“, et selgelt väljendada, et need spetsiifilised terminid ei leia vastet eestikeelses versioonis SNOMED CT-s (Tabel 5). See raskendab meditsiiniliste fraaside ja väljendite täpset võrdlemist standardsete koodidega, mis võib mõjutada vastendamise kvaliteeti.

Tabel 5: Näited puuduvatest sõnadest ja fraasidest taksonoomias.

Väljendid originaalsetest tekstidest	**SNOMEDi koodid+kontsept**	**Probleemid/märkused**
karikad		Puudub SNOMEDist
sisaldis kajavaba		Puudub SNOMEDist
mahulist muutust esile ei tule		Puudub SNOMEDist
soolegaasi varjust		Puudub SNOMEDist

Mõned ülaltoodud näited olid üksikud juhud. Tekstide vastendamisel rühmitasin eraldi sõnad ja fraasid kõige sagedamini esinevate probleemide alusel ning koostasin koondtabelid, mis sisaldasid kõiki tekkinud küsimusi, probleeme ja märkusi vabadest meditsiinilistest tekstidest:

1) Eitused. Sellesse kategooriasse kuuluvad sõnad ja väljendid, millel on eessõna „ei“ või mille lõpuosa on „-ta“.

Näiteks: „*mahulist muutust esile* ***ei tule***“, „*mõõtmetelt* ***suurenenud ei ole***“, „*kolde****leiuta***“, „*paisu* ***ei ole***“, „***ei*** *sedasta*“.

2) Lühendid. Sellesse kategooriasse kuuluvad sõnad, mis kontekstist lähtudes on täissõna mingi variandi ilmselge lühend.

   Näiteks: „*l/sõlmi*“, „*vas.*“, „*PU*“, „*v.porta*“.

3) Kääned. Sellesse kategooriasse kuuluvad sõnad kõigis võimalikes eesti keele käändes.

   Näiteks: „*kõhukoopa*“, „*kõhukoopas*“, „*neerude*“, „*tsüste*“, „*vasakul*“, „*vasakus*“.

4) Kirjavahemärgistused. Sellesse kategooriasse kuuluvad kõik sõnad, milles esineb kirjavahemärk, mis võiks mõjutada konteksti.

   Näiteks: „*neerupais?*“, „*LEID:*“.

5) Ladina keel. Sellesse kategooriasse kuuluvad kõik meditsiinilised terminid, mis on kirjutatud ladina keeles.

   Näiteks: „*Coecum*“, „*Dolor abdominis*“, „*Corpus callosum*“.

6) Inglise keel. Sellesse kategooriasse kuuluvad kõik meditsiinilised terminid, mis on kirjutatud inglise keeles.

   Näiteks: „*Appendicitis??*“, „*Pancreatitis*“.

7) Trükivead. Sellesse kategooriasse kuuluvad kõik sõnad, milles on esinenud trükiviga.

   Näiteks: „***T****ealine*“, „*App****e****nditsiit?*“.

8) Kontekstist sõltuvad fraasid. Sellesse kategooriasse kuuluvad kõik fraasid, mis vajavad selgitamist kontekstis.

   Näiteks: „***1-l uuringul*** *– kas ultraheliuuringul või teisel?*“, „***võrreldes varasema uuringuga*** *– missuguse uuringuga? UH või midagi muud?*“, „***kõhuvalu*** *- missugune kõhuvalu? Kas konkreetne valu tüüp?*“.

9) Suuruste kirjutamine. See kategooria hõlmab kõiki võimalikke suuruste kirjeldusi.

   Näiteks: „*44 mm*“, „*1,7 x 0,5 cm*“, „*80 x 32 x46 cm*“.

10) Kas sama? Sellesse kategooriasse kuuluvad kõik väljendid, mille õiges vastendamises ma ei olnud kindel, kuid mille sünonüümid ma taksonoomiast üles leidsin.

    Näiteks: „***homogenne** = normaalne kude (SCTID: 30389008)*“, „***Düsuurilised vaevused?** = Düsuuria (SCTID: 49650001)*“, „***Grav in hebd. 36** = Sünnituseelne 36. nädala läbivaatus (SCTID: 169721009)*“.

11) Puudub SNOMEDist. Sellese kategooriasse kuuluvad kõik sõnad ja fraasid, mis puuduvad SNOMED CT eestikeelsest versioonist.

    Näiteks: „*kubemesong*“, „*Dünaamika*“, „*nähtavale ei tule*“, „*mõõdukalt täitunud*“, „*Songavärat*“, „*Kubemekanal*“, „*kontraktsioonis*“, „*ei visualiseeru*“.

12) Sarnased terminid. Sellesse kategooriasse kuuluvad kõik terminid, mis kirjutatakse samamoodi, kuid neil on kaks või enam tähendust.

    Näiteks: „*Laiendamine*“.

13) Vigane tõlge. Sellesse kategooriasse kuuluvad kõik terminid, millele mille puhul otsingumootor annab tulemuseks sünonüümid, kuid need erinevad meditsiinilisest vaatenurgast.

    Näiteks: „*kõhukoobas*“ ei ole sama „*Kõhu- ja vaagnaõõs, tervik*“.

3.3.3 Valideerimine

Tulemuste valideerimiseks, saadud väljakutsete nimekiri oli võrreldud kirjanduses leitud väljakutsetega ning saadetud doktor Peeter Rossile ülevaatamiseks.

Doktor Peeter Ross tõi välja, et täpseks kodeerimiseks on oluline mõista sõnade konteksti, meditsiinilise tausta ja täpset tähendust. Näiteks rõhutas doktor Ross, et ultraheliuuringu puhul võib „vaagen“ viidata erinevatele struktuuridele, nagu „neeru vaagen“ ja mitte ainult üldisele anatoomilisele piirkonnale. Samuti juhtis ta tähelepanu, et „laienemine“ ja

„laiendamine“ ei ole sünonüümid, kuna „laiendamine“ viitab tegevusele, samas kui „laienemine“ on leid. Need näited illustreerivad, kuidas konteksti arvestamata jätmine võib viia vigaste koodide määramiseni, mis omakorda mõjutab tulemuste täpsust ja usaldusväärsust.

Enamus väljakutseid leidis aset ka kirjanduse ülevaatuse käigus (peatükk 3.1.4). Olid ka uued väljakutsed (nt. inglise- ja ladinakeelset sõnad, eesti keele kääned ning tõlkevead SNOMED CT eestikeelses versioonis). Kuna need on eesti keele ja eestikeelse SNOMED CT spetsiifilised, neid ei ole võimalik kirjandusega kinnitada. Aga on selge, et eesti keel ning eestikeelne SNOMED CT versioon kindlasti mõjutavad vastendamisprotsessi.

Kokkuvõtteks võib öelda, et saadud väljakutsete nimekiri on korrektne, kuigi ei pruugi olla täielik.

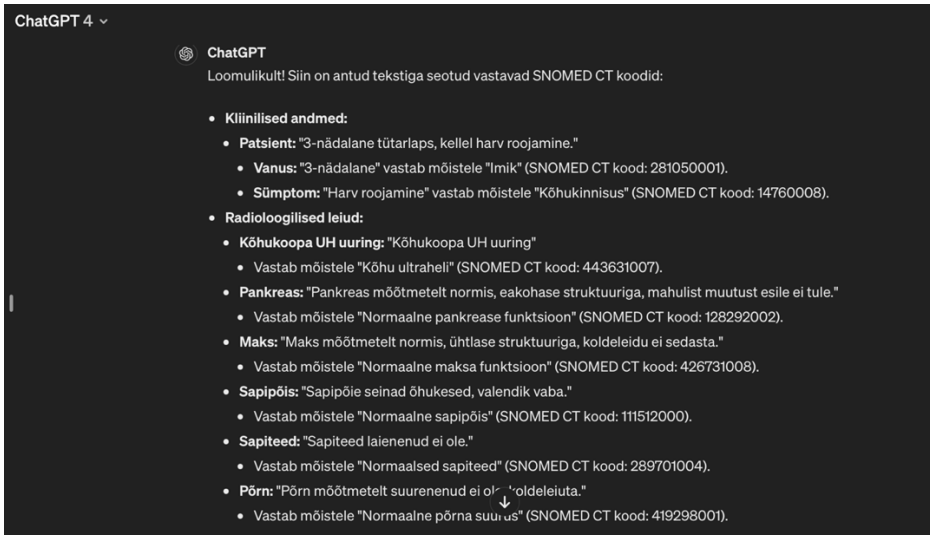
3.4 Vastendamine tehisintellektiga

Selleks, et katsetada tekstide vastendamist tehisintellekti ChatGPT [10] abil, koostasin detailse päringu. Päringus selgitasin täpselt, mida soovin tulemuseks saada. Eesmärgiks oli see, et ChatGPT suudaks vastendada ühe meditsiinilise vabateksti vastavate SNOMED CT taksonoomia koodidega (Joonis 13).

3.4.1 Näide ChatGPT vastendamisega

Katsetamise jaoks tehti päring järgmise teksti vastendamiseks:

„Kliinilised andmed: 3 nädalane tütarlaps, kellel harv roojamine. Kõht puhitunud, väljaheide tihe, lihtjas. M.Hirschsprungi? Kõhukoopa UH uuring. LEID: Pankreas mõõtmetelt normis, eakohase struktuuriga, mahulist muutust esile ei tule. Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta. Sapipõie seinad õhukesed, valendik vaba. Sapiteed laienenud ei ole. Põrn mõõtmetelt suurenenud ei ole, koldeleiuta. Neerude suurus, asetsus ja struktuur ealised. Neerudes paisu ei ole. Kusepõis vähe täitunud, sisaldis kajavaba. Kusepõie seinad suhteliselt paksud - tühja põie korral mõõdetav ca 7 mm seina paksus. Vaba vedelikku kõhukoopas ei sedasta. Rohkelt soolegaasi üle kogu kõhu. Kokkuvõte: Rohkelt soolegaasi üle kogu kõhu.“.

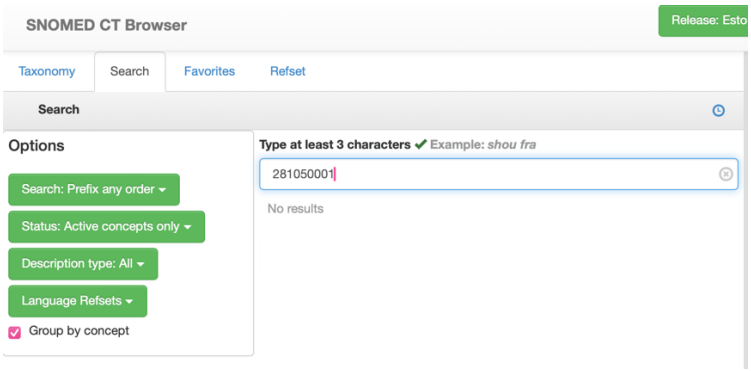


Joonis 13. Tehisintellektiga vastendamise tulemus.

3.4.2 Valideerimine

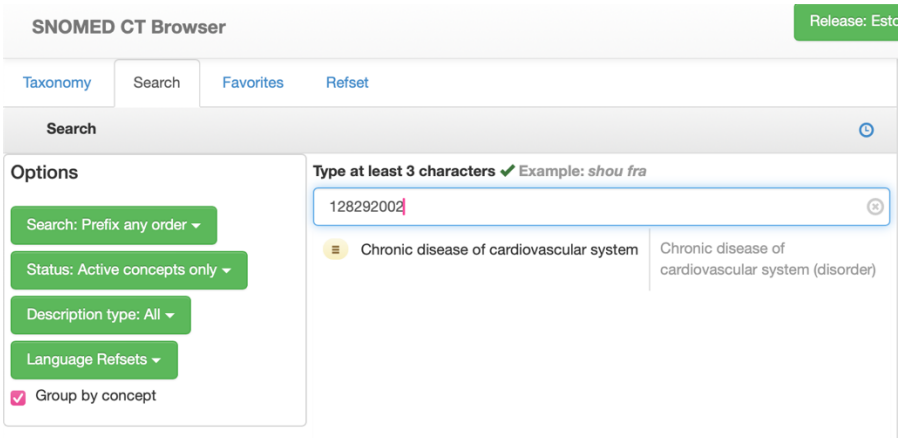
Pärast seda, kui olin võrrelnud ChatGPT poolt esitatud tulemusi SNOMED CT taksonoomia tegelike koodidega, selgus, et ChatGPT oli mõne kontseptsiooni puhul koodid valesti määranud ja üldistanud.

Näiteks ei ole SNOMED CT taksonoomias konkreetset koodi „Imik“ tähistamiseks, ja tehisintellekti poolt edastatud SCTID 281050001 puudub praeguses taksonoomia versioonis (Joonis 14).



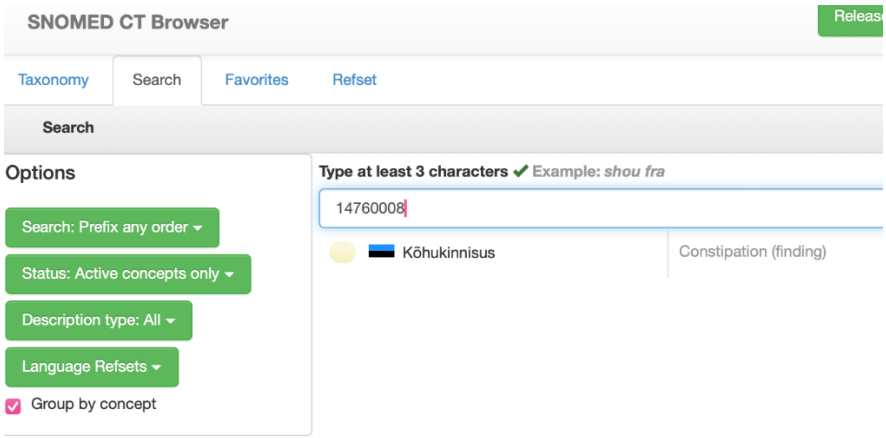
Joonis 14. SCTID 281050001 puudub SNOMED CT-st.

Samuti, „Normaalne pankrease funktsioon“ vastendas ChatGPT koodiga SCTID: 128292002, mis praeguses SNOMED CT versioonis tähistab kroonilisi kardiovaskulaarsüsteemi haigusi (Joonis 15). See ei ole kuidagi seotud pankrease ega selle funktsioonidega.



Joonis 15. SCTID 128292002 vastav termin

„Kõhukinnisus“ (SCTID: 14760008) on ainus termin, mille ChatGPT õigesti mõistis ja vastendas selle vaba meditsiinilise teksti jaoks (Joonis 16).



Joonis 16. Kõhukinnisus on õigesti vastendatud

Samuti jättis tehisintellekt tähelepanuta olulised aspektid, näiteks et tegemist oli naissoost patsiendiga. Ta ei kirjeldanud ka üksikasjalikult iga leidu maksast (lähtefraas: „Maks mõõtmetelt normis, ühtlase struktuuriga, koldeleidu ei sedasta“), vaid asendas selle ühe üldise fraasiga: „Normaalne maksa funktsioon“. Kirjeldades põrna (lähtefraas: „Põrn mõõtmetelt suurenenud ei ole, koldeleiuta“), jättis ChatGPT mainimata, et põrnas ei olnud koldeid, märkides vaid, seda suurust fraasiga: „Normaalne põrna suurus“.

3.5 Sagedussõnastiku koostamine

Sagedussõnastiku koostamiseks oli tehtud Python skript, mis loeb kõiki tekste sisse, tükeldab neid tühiku järgi, ning loeb sõnade kahe-, kolme- ja neljapikkuste väljendite arvu kõikides tekstides.

Sagedussõnastik on esitatud Exceli tabeli formaadis [25], kus ühes veerus on eraldi välja toodud sõnad ja väljendid ning teises veerus on näidatud, mitu korda need korduvad kõikides 20 000 tekstides (Joonis 17).

Sagedussõnastikus on aga liiga palju sõnu ja väljendeid, mis teeb selle töötlemise keeruliseks. Näiteks tabeli koostamisel ei ole arvestatud eesti keele käänetega, mis tähendab, et sama sõna võib esineda mitmes erinevas vormis, muutes analüüsi keerukamaks. Samuti on kirjavahemärgid jäetud tähelepanuta, mis võib segada täpset tulemuste tõlgendamist.

Selline sagedussõnastiku tabel vajab väga põhjalikku ja pikaajalist käsitsi töötlemist. Esiteks ilmnes, et skript oli võimeline käsitsema praktiliselt identseid sõnu või väljendeid kui erinevaid elemente, mis põhjustas andmete duplitseerimist, mis tuleb eemaldada. See viis olukorrani, kus analoogsed terminid registreeriti ekslikult kui unikaalsed sõnad, raskendades seeläbi analüüsi tulemuslikkust ja täpsust. Teiseks tuleb kõrvaldada numbrid ja arvud, mis ei oma vastendamise seisukohalt tähendust ja võivad segadust tekitada. Lisaks on oluline tähelepanu pöörata kirjavahemärkidele, mis võivad mõjutada sõnade ja väljendite eristamist ja mõistmist.

67	suurenenud ei	17348
68	suurenenud ei ole	17348
69	ole. põrn	17301
70	ei ole. põrn	17301
71	laienenud ei ole. põrn	17286
72	kõhukoopas	17282
73	mõõtmetelt suurenenud ei	17269
74	mõõtmetelt suurenenud ei ole	17264
75	põrn mõõtmetelt suurenenud ei	17235

Joonis 17. Näide sagedussõnastiku tabelist

Kõige parem oleks, kui sagedussõnastik vaadataks üle koos meditsiinivaldkonna spetsialisti või selliste tekstide autoriga. Autor suudab paremini hinnata, millised sõnad ja väljendid on tõesti olulised ja millised mitte. See koostöö aitaks tagada, et sagedussõnastiku tulemused oleksid täpsemad ja usaldusväärsemad, muutes seeläbi kogu vastendamise protsessi efektiivsemaks ja täpsemaks.

Ülaltoodud probleemidele vaatamata, aitas sagedussõnastik tuvastada tõlkimata, aga tihti ultraheli uuringutes kasutatavad sõnad ja väljundid. Näiteks, liigitavad väärtused (*qualifiers*) ja atribuudid (*attributes*): „suurenenud“, „mõõtmetelt“, „nähtav“ ja muud. Arusaadav, et need on ultraheli uuringute spetsiifilised ning tõenäoliselt ei olnud kõige prioriteetsemad tõlkimiseks. Seega kui tahta teha vastendamismudelit ultraheli uuringute jaoks, tuleks need sõnad ka tõlkida SNOMED CT eestikeelses versioonis.

4 Analüüs ja järeldused

See peatükk keskendub kasutatud tööriistade-, protsesside kirjeldamise- ja saadud tulemuste analüüsimisele, tulemuste valideerimisele ja edasise töö võimalustele.

4.1 Tööriistad

SNOMED CT kasutamine vahendina vastendamiseks oli minu töös suurepärane valik. Süsteemi kasutajasõbralik liides pakkus kogu vajalikku informatsiooni kiiresti ja hõlpsasti kättesaadavalt. Eriti meeldis mulle terminite filtreerimise võimalus, mis võimaldas tõhusalt leida vajalikud terminid. Lisaks positiivsele kasutajakogemusele brauseris sain palju väärtuslikku teavet, mis oli minu uurimistöö jaoks äärmiselt oluline. SNOMED CT on väga hea tööriist vastendamise eesmärgil, kuid kahjuks on eestikeelne versioon oluliselt piiratum võrreldes ingliskeelse versiooniga, mis omakorda loob piirangud uurimistööle.

Katse kasutada vastendamiseks ChatGPT näitas, et tehisintellekt võib meditsiinilisi tekste valesti tõlgendada. See tõi kaasa ebaõigeid tulemusi, mis rõhutab teabe kontrollimise tähtsust. Kuna SNOMED CT on litsentseeritud, ei tohitud talle SNOMED CT identifikaatoreid ise edastada. Mõnesid õigesti vastendatud identifikaatoreid tõenäoliselt õppis ta avalikult saadavatest allikatest.

Pythoni kasutamine tööriistana minu uurimistöös osutus kasulikuks. Pythoni skripti abil suutsin töödelda ja analüüsida suuri tekstikogumeid, mis sisaldasid meditsiiniliste ultraheliuuringute tulemusi. Skript võimaldasid mul kiiresti ja täpselt koostada sagedussõnastikku, mis aitas tuvastada kõige sagedamini esinevaid sõnu ja fraase vabatekstides.

4.2 Protsess

Minu uuringu tulemused tõid esile vajadust meditsiiniliste terminite mõistmine ja nende korrektne sobitamine SNOMED CT koodidega ilma meditsiinilise taustata. Paljudel juhtudel vajas see täiendavat uurimist, et tagada terminite õige valik. Samal ajal meditsiinilise tausta puudumine ja vastendamisel vigade tegemine aitas just potentsiaalsed väljakutsed kaardistada.

Kokku vastendasin täielikult 38 teksti ja osaliselt 33, mis oli piisavalt peamiste probleemide tuvastamiseks SNOMED-i eestikeelse versiooni vastendamisel ning sagedasti korduvate sõnade ja fraaside kindlakstegemiseks esitatud vabades meditsiinitekstides.

4.3 Tulemused

Töö põhitulemuseks on tekkinud vastendamisel tekitavate väljakutsete nimekiri: eitused, lühendid, eesti keele kääned, kirjavahemärgistus, ladina ja inglise keele kasutamine tekstides, trükivead, kontekstist sõltuvad väljendid, suurused, sünonüümid, tõlkimata ja valesti tõlgitud terminid eestikeelses SNOMED CT versioonis ja sõnaliigi tuvastamine.

Kirjanduses mainitakse ka SNOMED CT, kui taksonoomia, vasturääkivust ning seda, et mõnesid termineid on võimalik vastendada ainult osaliselt või ainult üldisema või täpsema SNOMED CT terminiga.

Kokkuvõtteks võib öelda, et eestikeelsete meditsiiniliste tekstide vastendamist SNOMED CT taksonoomiaga mõjutavad järgmised aspektid:

- SNOMED CT, kui taksonoomia, vasturääkivus,
- igas keeles esinevad aspektid, nagu eitused, sünonüümid ja sõnaliigi tuvastamine,
- eesti keele spetsiifika (nt. kääned ja paindlik süntaktiline struktuur),
- tekstide mitmekeelsus (eesti arstid kasutavad ka inglise ja laadina keelt oma tekstides),
- kontekst (laiemas mõttes nagu meditsiiniline valdkond, konkreetse arsti valdkond nagu ultraheli uuringud, kui ka konkreetse teksti kontekst nagu viitamine mingile teisele dokumendile või uuringule),
- eestikeelse SNOMED CT versiooni seis (tõlkimata või valesti tõlgitud terminid),
- inimfaktor tekstide kirjutamisel (kirjavead ja lühendite kasutamine).

Kuigi ülaltoodud väljakutsete ja aspektide nimekirjad ei pruugi olla täielikud, need on kindlasti abiks edasises töös vastendamise protsessi automatiseerimisel. Need annavad ülevaadet selle kohta, mida tuleb vastendamise protsessi automatiseerimisel jälgida.

Kuigi doktor Peeter Ross tõi välja, et käsitsi vastendamisel saadud SNOMED CT terminid olid enamasti valed ja nendest temale kasu ei ole, aitas see väljakutseid kaardistada. Töö väärtus tema jaoks on see, et näitab sellise tegevuse perspektiivitust ja vajadust enne sõnade kodeerimist tegeleda lausete sisuga.

Huvitav on ka see, et isegi tekstide autorina doktor Peeter Rossil ühe teksti jaoks oli ka raske ühe termini tähendusest aru saada ja õigesti seda tõlgendada, kuna tekstide kirjutamise ajast möödus tükk aega.

Teiseks tulemuseks oli see, et ChatGPT abil ei ole mõistlik tekste SNOMED CT taksonoomiaga vastendada. Enamus SNOMED CT identifikaatoreid mõtles ta välja. Seda võib tõlgendada niimoodi, et SNOMED CT on litsentseeritud.

Kolmandaks tulemuseks oli sagedussõnastiku koostamine doktor Peeter Rossi tekstide põhjal. See aitas tuvastada eestikeeleses SNOMED CT versioonis tõlkimata termineid.

Kui kasutada sagedussõnastiku koostamiseks loomuliku keele töötlemise edasijõudnud tehnikaid ja kaasata selle puhastamiseks meditsiinilised töötajad, võib selle abil koostada arsti isikliku vastendamismudelit.

Kuigi mõned vastendamisel tuvastatud väljakutsed on tihti esinevad iga teksti töötlemisel, selle töö uudsus on just eestikeelsete tekstide vastendamine ja eestikeelse SNOMED CT versiooni probleemid. Eestikeelset SNOMED CT versiooni vastendamist ei ole veel põhjalikult uuritud ning eestikeelse SNOMED CT versiooni seis piirab oluliselt selle taksonoomia uurimistööd ja praktilist rakendamist eestikeelses meditsiini kontekstis.

Hindan oma tööd positiivselt, sest olen saavutanud kõik eesmärgid, mis endale püstitasin. Uurimistöö käigus suutsin tuvastada ja uurida probleeme, mis on seotud eestikeelsete meditsiiniliste tekstide vastendamisega SNOMED CT taksonoomias. Minu töö toimus minu juhendajate tähelepaneliku ja tundliku juhtimise all. Lisaks kontrollis tulemusi kogenud radioloog, mis lisab uuringu järeldustele kaalu ja usaldusväärsust. Mul on väga hea meel, et sain oma lõputöö kaudu meditsiinimaailma puudutada. Olen rahul sellega, kuidas mu töö edenes ja millised ülesanded sellega kaasnesid, samuti olen rahul, et suutsin

tuvastada vabade meditsiinitekstide vastendamisel tekkivad probleemid, mis oligi selle töö peamine eesmärk.

4.4 Valideerimine

Valideerimiseks kasutasin mitmeid meetodeid, et tagada tulemuste täpsus ja usaldusväärsus. Esimeseks meetodiks oli võrrelda saadud tulemusi olemasolevate allikatega, mis hõlmas teadusartiklite ja meditsiiniliste andmebaaside analüüsi. Teiseks konsulteerisin dr Peeter Rossiga, kes on tegutsev radioloog ja kes andis tagasisidet vastendamise täpsuse kohta. Tema kogemus ja teadmised aitasid tuvastada võimalikke vigu. Kolmandaks kontrollisin termineid SNOMED CT-s, et kinnitada ChatGPT poolt tehtud vastendamise õigsust. See hõlmas SNOMED CT-s leiduvate terminite ja nende vastavuste uuesti läbivaatamist, et veenduda, et kõik vastendatud terminid olid korrektsed. Kõik need meetodid osutusid tõhusaks ja asjakohaseks selle uurimistöö jaoks.

4.5 Edasine töö

SNOMED CT on võimas taksonoomia, mida on võimalik kasutada paljude ülesannete lahendamiseks (peatükk 3.1.2). Samuti eestikeelse SNOMED CT versiooni vastendamist ei ole põhjalikult uuritud. Selle töö tulemused annavad palju võimalusi edasiseks tööks.

Kirjanduse ülevaade käigus tuli välja, et SNOMED CT terminitega vastendamiseks kasutatakse erinevaid meetodeid: käsitsi vastendamine, reeglipõhine vastendamine, vastendamine kasutades loomuliku keele töötlemise tehnikaid, vastendamine kasutades masinõpe meetodeid ja nende hübriidid. Järgmisena võib eestikeelsete tekstide vastendamist SNOMED CT terminitega proovida automatiseerida, arvestades selle töö tulemusena avastatud vastendamise väljakutseid ja vastendamist mõjutavaid aspekte.

Üks perspektiivne võimalus meditsiinilise teksti vastendamise parandamiseks on luua eri spetsialistide jaoks kohandatud „sõnastikud“, mis võimaldaks võtta arvesse nende erialase keele ja märkmete koostamise stiili eripärasid. See võiks oluliselt parandada vastendamise täpsust ja lihtsustada protsessi konkreetsete meditsiinivaldkondade puhul.

Selle töö raames sagedussõnastik oli koostatud ultraheli uuringute tekstide põhjal. See aitas tuvastada selle valdkonnaga seotud tõlkimata SNOMED CT termineid. Sama eesmärgi saavutamiseks võib kasutada ka teiste valdkondade tekste.

5 Kokkuvõte

Käesolevas töös uuriti meditsiiniliste vabatekstide ehk kõhu- ja vaagnapiirkonna ultraheliuuringute vastendamist SNOMED CT taksonoomiaga. Probleem seisneb selles, et vabade meditsiiniliste tekstide töötlemine arvutitega on keeruline, kuna sama nähtust saab kirjeldada erinevalt ja sama sõna võib olenevalt kontekstist tähendada erinevaid asju. Töö eesmärk oli tuvastada ja analüüsida probleeme, mis tekivad eestikeelsete vabade meditsiiniliste tekstide vastendamisel SNOMED CT taksonoomiakoodidega. Metoodikas kasutati peamiselt käsitsi vastendamist, kus valiti välja 71 erinevat teksti, jaotati need vanuse- ja soogruppidesse ning koostati vastavad tabelid. Seejärel viidi läbi vastendamine, kus määrati SNOMED CT koodid sõnadele ja fraasidele.

Tulemuste kohaselt leiti, et käsitsi vastendamise abil oli võimalik määrata osa SNOMED CT koodidest, kuid esines ka probleeme: näiteks kontekstist sõltuvad tähendused, eitused ja lühendid. Saadud tulemused näitasid, et tõlke kvaliteet ja vastendamise täpsus on omavahel tihedalt seotud. Samuti ilmnes, et mõned terminid ja kontseptsioonid ei olnud SNOMED CT-s üldse esindatud, mis raskendas vastendamise protsessi.

Vastendamise tehisintellekti abil leiti, et ChatGPT-le ei saa tugineda, kuna mõnede kontseptsioonide puhul määras ta koodid valesti ja üldistas need liialt.

Järeldustes leiti, et üldise vastendamismudeli loomine kõigi erialade arstidele ja meditsiinitöötajatele on keeruline, mistõttu keskenduti konkreetse arsti individuaalsele sõnastikule. Selle töö jaoks olid edastatud vabatekstide näited, mis võimaldasid luua sagedussõnastiku, kus kajastusid kõige sagedamini kasutatavad sõnad ja fraasid. Lisaks tuvastati vastendamise probleemid ja koostati kokkuvõtvad tabelid, mis võimaldavad tulevikus arutelusid ja tulemuste täpsustamist. Kokkuvõttes on töö keerukus keskmine, arvestades vajadust käsitsi vastendamise ja kontekstitundliku analüüsi järele. Viidatud allikad on asjakohased ja ajakohased, tagades töö teoreetilise ja praktilise aluse.

Kasutatud kirjandus

[1] S. Pakhomov, J. Buntrock, ja P. Duffy, „High Throughput Modularized NLP System for Clinical Text“, *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, M. Nagata ja T. Pedersen, Toim, Ann Arbor, Michigan: Association for Computational Linguistics, juuni 2005, lk 25–28. doi: 10.3115/1225753.1225760.

[2] E.-Y. So ja H.-A. Park, „Mapping medical records of gastrectomy patients to SNOMED CT“, *Stud Health Technol Inform*, kd 169, lk 764–768, 2011.

[3] J. Patrick, Y. Wang, ja P. Budd, „An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology“.

[4] „Unified Medical Language System (UMLS)“. Vaadatud: 19. mai 2024. [Online]. Available at: https://www.nlm.nih.gov/research/umls/index.html

[5] „SNOMED CT“. Vaadatud: 19. mai 2024. [Online]. Available at: https://www.nlm.nih.gov/healthit/snomedct/index.html

[6] „SNOMED CT - Home“. Vaadatud: 19. mai 2024. [Online]. Available at: https://browser.ihtsdotools.org/?

[7] C. Gaudet-Blavignac, V. Foufi, M. Bjelogrlic, ja C. Lovis, „Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review“, *J Med Internet Res*, kd 23, nr 1, lk e24594, jaan 2021, doi: 10.2196/24594.

[8] A. N. Nguyen *et al.*, „Symbolic rule-based classification of lung cancer stages from free-text pathology reports“, *J Am Med Inform Assoc*, kd 17, nr 4, lk 440–445, 2010, doi: 10.1136/jamia.2010.003707.

[9] E. Chang ja J. Mostafa, „Cohort Identification from Free-Text Clinical Notes Using SNOMED CT’s Hierarchical Semantic Relations“, *AMIA Annu Symp Proc*, kd 2022, lk 349–358, apr 2023.

[10] „ChatGPT“. Vaadatud: 19. mai 2024. [Online]. Available at: https://chatgpt.com

[11] T. Wu *et al.*, „A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development“, *IEEE/CAA Journal of Automatica Sinica*, kd 10, nr 5, lk 1122–1136, mai 2023, doi: 10.1109/JAS.2023.123618.

[12] „Welcome to Python.org“, Python.org. Vaadatud: 19. mai 2024. [Online]. Available at: https://www.python.org/about/

[13] „Kõhu- ja vaagnapiirkonna ultraheliuuring (UH)_7953_01.07.2009-01.07.2011 2.xlsx — Microsoft Excel Online“. Vaadatud: 25. mai 2024. [Online]. Available at: https://onedrive.live.com/view.aspx?resid=B9CDD07B19BCF00D!420311&cid=b9cdd07b19bcf00d&authkey=!AAn3DEGLNuYdFuU&CT=1716659626368&OR=ItemsView

[14] P. Runeson, M. Höst, A. Rainer, ja B. Regnell, *Case Study Research in Software Engineering: Guidelines and Examples*, 1. tr. Wiley, 2012. doi: 10.1002/9781118181034.

[15] „Valitud tekstide nimekiri.pdf — Microsoft Word Online“. Vaadatud: 25. mai 2024. [Online]. Available at: https://onedrive.live.com/view.aspx?resid=B9CDD07B19BCF00D!420307&authkey=!AAn3DEGLNuYdFuU

[16] „Inimese elukulg - ELUKAAR“. Vaadatud: 19. mai 2024. [Online]. Available at: https://www.hariduskeskus.ee/opiobjektid/elukulg/?INIMESE_ARENG___ELUKAAR

[17] „5-Step briefing“, SNOMED International. Vaadatud: 19. mai 2024. [Online]. Available at: https://www.snomed.org/five-step-briefing

[18] „Overview of SNOMED CT“. Vaadatud: 19. mai 2024. [Online]. Available at: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html

[19] A. N. Nguyen, M. J. Lawley, D. P. Hansen, ja S. Colquist, „A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text“.

[20] L. Zhou *et al.*, „Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to Process Medication Information in Outpatient Clinical Notes“, *AMIA Annu Symp Proc*, kd 2011, lk 1639–1648, 2011.

[21] A. R. Aronson, „Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.“, *Proc AMIA Symp*, lk 17–21, 2001.

[22] P. L. Elkin *et al.*, „Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists“, *Mayo Clinic Proceedings*, kd 81, nr 6, lk 741–748, juuni 2006, doi: 10.4065/81.6.741.

[23] G. K. Savova *et al.*, „Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications“, *J Am Med Inform Assoc*, kd 17, nr 5, lk 507–513, 2010, doi: 10.1136/jamia.2009.001560.

[24] „Käsitsi vastendamine.pdf — Microsoft Word Online“. Vaadatud: 25. mai 2024. [Online]. Available at:

https://onedrive.live.com/view.aspx?resid=B9CDD07B19BCF00D!420310&authkey=!AAn3DEGLNuYdFuU

[25] „Sagedussõnastik.xslx — Microsoft Excel Online“. Vaadatud: 25. mai 2024. [Online]. Available at: https://onedrive.live.com/view.aspx?resid=B9CDD07B19BCF00D!420309&cid=b9cdd07b19bcf00d&authkey=!AkJfl1CjZGfqn8c&CT=1716676746040&OR=ItemsView

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Milena Suvorova

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Eestikeelsete meditsiinitekstide vastendamine SNOMED CT-ga“, mille juhendaja on Bahdan Yanovich
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

26.05.2024

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.