Elektrooniliste terviselugude analüüsimise võimalused Tartu perearstide infosüsteemi näitel

Sulev Reisberg[1,2,6], Raul Sirel[1,3], Ruth Kalda[5], Markko Merzin[1,4], Jaan Pruulmann[7], Jaak Vilo[1,2,6]

Eesti Arst 2013; 92(8):452–459

Saabunud toimetusse: 19.04.2013
Avaldamiseks vastu võetud: 14.06.2013
Avaldatud internetis: 30.09.2013

[1] Tarkvara Tehnoloogia Arenduskeskus OÜ,
[2] TÜ arvutiteaduse instituut,
[3] TÜ eesti ja üldkeeleteaduse instituut,
[4] TÜ infotehnoloogia osakond,
[5] TÜ polikliiniku ja peremeditsiini õppetool,
[6] Quretec OÜ,
[7] Commit OÜ

Kirjavahetajaautor:
Raul Sirel
rsirel@ut.ee

Võtmesõnad:
perearst, elektrooniline terviselugu, terviseinformaatika, andmeanalüüs

Taust, eesmärk. Artikli eesmärgiks on anda ülevaade analüüsidest, mida on võimalik teha Tartus asunud Maarjamõisa polikliiniku perearstide patsientide terviseandmetest moodustatud andmebaasi alusel. Tegemist on sissejuhatava uurimusega, mille eesmärgiks on tutvustada andmeanalüüsi potentsiaali elektrooniliste terviselugude uurimisel.

Metoodika. Aastatel 1995–2011 Tartu Maarjamõisa polikliiniku perearstide töövahendiks olnud patsientide anonüümseks muudetud terviseandmeid sisaldava infosüsteemi andmestikku analüüsiti erinevate jaotus- ja sagedusanalüüsi meetoditega.

Tulemused. Vaadeldud ajavahemikul toimus aastaaegade kaupa kõige enam arstivisiite sügisesel ja talvisel perioodil ning nädalapäevadest esmaspäeval. Visiidikirjed sisaldasid kokku 18 462 524 sõna, millest 14% moodustasid lühendid. Kokku kasutati 190 789 unikaalset algvormi ehk lemmat, millest ainult 78 909 lemmat (41,36%) oli kasutatud enam kui üks kord ning 25 437 lemmat (13,33%) oli kasutatud vähemalt 10 korda. Suurima esinemissagedusega sõnad või lühendid olid *rr, ravi, x, olema* ja *mg*. Kokku kasutati andmestikus 5389 erinevat RHK-10 diagnoosi, millest 425 (7,9%) sagedasemat unikaalset diagnoosi moodustasid 90% kõigist diagnoosidest. Andmestikus esinevad retseptid sisaldavad 718 erinevat toimeainet, millest 144 (20%) sagedasemat unikaalset toimeainet moodustasid 90% kõigist retseptidega väljakirjutatud toimeainetest.

Kokkuvõte. Analüüsi tulemusi on võimalik kasutada tervishoiukorralduslike otsuste tegemisel. Andmestik on erakordselt mahukas ja sel on seetõttu suur potentsiaal täiendavaks analüüsiks, sh andmekaevemeetodite rakendamiseks.

Viimase 10–15 aasta jooksul on nii Eesti kui ühtlasi kogu maailma meditsiinis toimunud oluline paradigmamuutus, mille tulemusel on suures osas loobutud paberipõhisest andmehaldusest ning arstide igapäevaste töövahendite hulka on lisandunud arvutid. Juba 1998. aastal kasutas arvuteid oma töös tervelt 76% Eesti perearstidest (1). Selline kvalitatiivne muutus on oluline nii patsiendi, tema raviarsti kui ka näiteks tervishoiu- või rahvastikuteadlaste vaatenurgast.

Patsiendile tähendab selline muutus näiteks ligipääsu oma haigusloole (Eestis on see teostunud alates 2009. aastast üleriigilise terviseinfosüsteemi kaudu), mis mitmete autorite hinnangul aitab patsientidel paremini toime tulla krooniliste haigustega, parandada ravikvaliteeti haiglast väljakirjutamise järgsel ajaperioodil ning loomulikult suurendada ka tervishoiuteenuste osutamise (sh arstide arutluskäikude) läbipaistvust (2). Samuti aitab elektrooniline terviselugu kaasa ka tõenduspõhise meditsiini lõplikule juurdumisele ning uute tarkvaratehnoloogiate (nn *decision support systems*) arendamisele, mis võimaldavad arstidel jõuda kiiremini informeeritud otsusteni, suurendavad patsiendi ohutust ning vähendavad kliinilistest eksimustest johtuvaid terviserikkeid (3).

Tervishoiu- ja rahvastikuteaduse seisukohalt annab elektrooniline terviselugu teadlastele aga pretsedenditu ligipääsu andmetele, millest on võimalik tuvastada ravimite varem kirjeldamata kõrval- ja

koostoimeid, iseloomustada ning ennustada rahvastiku tervist, seirata täpsemalt nakkushaiguste puhanguid jms (3).

Terviseuuring selle traditsioonilises tähenduses põhineb eesmärgipäraselt koostatud kohordi ja kindlate protseduuride alusel kogutud andmestiku analüüsil. Elektroonilise terviseloo (ETL, inglise keeles *electronic health record (EHR)*) andmete puhul on olukord teistsugune, sest valim hõlmab kogu (lokaalset) elanikkonda ning andmestik sisaldab väga suurt tunnuste komplekti (iga patsiendi kohta on erinevat tüüpi teavet eri arstidelt) kogutuna võrdlemisi pika ajavahemiku jooksul.

ETLiga seotud temaatika aktuaalsusele vaatamata on sellist tüüpi andmestikke seni uuritud siiski suhteliselt vähe. Jättes kõrvale publikatsioonid, mis uurivad ETLi eeliseid paberil terviseandmete ees (üksiku patsiendi või tervisekorralduse kulutõhususe seisukohast), on ETLi baasil enim uuringuid tehtud ravimite kõrvaltoimete kohta. Nii näitas Ameerika Ühendriikide Toidu- ja Ravimiameti töötaja David J. Graham ETLi andmeid analüüsides juba 2004. aastal, et mittesteroidne põletikuvastane ravim Vioxx suurendab võrreldes alternatiivsete ravimitega infarktiriski 50–370% (4). Samuti kasutasid ETLi andmeid Nicholas Tatonetti ja Russ Altman, kes otsisid 2011. aastal kinnitust hüpoteesile, et antidepressant paroksetiini ja kolesterooli alandava pravastatiini koostarvitamine tõstab veresuhkru taset (5, 6).

Eestis sellises mahus elektroonilisi terviselugusid autoreile teadaolevalt uuritud ei ole. Kõige sarnasemaks võib pidada Eesti Haigekassa raviarvete alusel tehtud uuringuid, kus analüüsi aluseks olid teenuste või protseduuride koodid. Nii näiteks on Eero Merilind uurinud perearstide töökoormuse dünaamikat aastatel 2005–2008 (7). Samuti on mitmeid uurimusi, kus on kombineeritud Eesti Haigekassa ja muude tervishoiuregistrite (näiteks müokardiinfarkti- ja rahvastikuregistri (8)) andmeid.

Artikli **eesmärgiks** on anda ülevaade analüüsidest, mida on tehtud (ja edaspidi võimalik teha) andmebaasi kogutud Tartu Maarjamõisa polikliiniku perearstide patsientide terviseandmete põhjal. Artiklis tutvustatakse andmestiku üldpilti: patsientuuri demograafilist profiili (sugu, vanus, päritolu), visiitide ajalist jaotust (nädalapäevade ja kuude kaupa) ning perearstide määratud diagnooside ning väljakirjutatud retseptide jaotust ning sageduslikke eripärasid. Tegemist on nimetatud andmestikul põhineva sissejuhatava uurimusega.

METOODIKA

Käsitletavaks andmestikuks on aastatel 1995–2012 Tartu Maarjamõisa polikliinikus kasutusel olnud infosüsteemi MIS+ andmed (9), mille kasutajate hulka kuulusid selles ajavahemikus selles asutuses töötanud perearstid (12 praksist) ning aastatel 1995–2001 polikliiniku registratuur ning röntgeni, sonograafia ja ehhokardiograafia kabinetid.

2012. aasta alguses infosüsteem suleti ning selles leidunud andmed muudeti anonüümseks TÜ inimuuringute eetika komitee loa 213/T-10 alusel. Anonüümseks muudetud andmeid on töödeldud Tarkvara Tehnoloogia Arenduskeskuse programmi alamprojekti WP 1.2 *„Biomedical data integration and mining"* töörühmas eesmärgiga analüüsida MIS+ infosüsteemis sisalduvaid kliinilisi andmeid lingvistilisest ja infotehnoloogilisest aspektist, rakendades neile erinevaid keeletehnoloogia, andme- ja tekstikaeve ning masinõppe meetodeid.

Infosüsteemist ekstraheeritud andmestik sisaldab mitmekülgset teavet patsientide haigusloo kohta, kuid samuti on näiteks demograafilisest aspektist teada patsiendi sugu, vanus ja elukoht omavalitsusüksuse täpsusega. Samuti sisalduvad seal andmed patsientidele pandud kliiniliste diagnooside (1995.–1996. aastal RHK-9, hiljem RHK-10 kodeeringus) ning neile määratud ravimite kohta. Lisaks kodeeritud informatsioonile sisaldab andmestik ka märkimisväärsel hulgal vabatekstilist teavet.

Andmestik ei sisalda infot selle kohta, milliseid raviasutusi on patsient külastanud lisaks Maarjamõisa perearstidele, samuti ei ole andmestikku rikastatud teiste andmekogude andmetega ega ole teada, millega seoses lõpevad konkreetse patsiendi andmed (patsientide nimistutevaheline liikumine, surm, arsti loobumine infosüsteemi MIS+ kasutamisest). Retseptide osas puuduvad andmed selle kohta, kas patsient ka tegelikult retsepti realiseeris ning millist preparaati tarvitas.

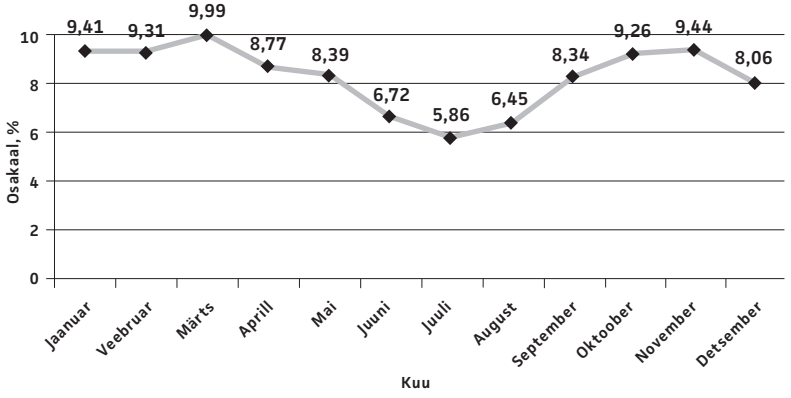
Käsitletava andmestiku kirjeldamiseks kasutatakse erinevaid jaotus- ja sagedusanalüüsi meetodeid.

TULEMUSED

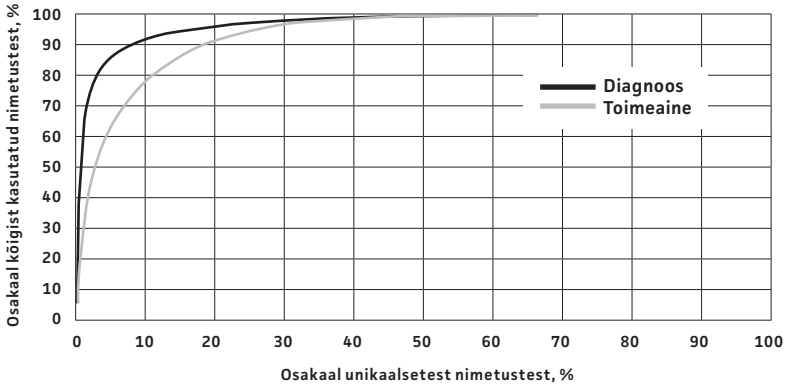
Andmebaasi kuulunud patsiente iseloomustavad demograafilised näitajad

Andmestik sisaldab infot 165 375 patsiendi kohta, kellest 44,5% olid mehed ning 53,1% naised, kuid 2,4% patsientide sugu ei olnud teada. Patsientide sünniaastad jäid vahemikku 1900–2012, kuid valdav osa neist (85,7%) oli sündinud aastatel 1930–1999. Pärast 2000. aastat sündinute osakaal oli vaid 1,6% ning 2,4% patsientide sünniaasta ei ole teada. Kümnendite kaupa on enim patsiente sündinud vahemikus 1970–1979 (15,8%), järgnevad 1980.–1989. aasta 13,47%-ga ning 1960.–1969. aasta 12,65%-ga.

Elukoha järgi oli enamik patsiente (61,3%) pärit Tartu maakonnast, kuid kuna polikliiniku röntgeni, sonograafia ja ehhokardiograafia kabinetid teenindasid patsiente ka teistest maakondadest, siis oli 5,4% patsientidest Jõgeva, 4,6% Valga, 4,2% Ida-Viru ning 4,1% Põlva ja 4,1% Viljandi maakonnast. Samuti oli patsiente Võrumaalt (3,6%), Lääne-Virumaalt (2,8%), Harjumaalt (2,1%), Järvamaalt (1,7%) ja Pärnumaalt (1,5%). Teiste Eesti maakondade esindatus andmebaasis jääb juba alla 1%.



Joonis 1. Aastatel 1995–2011 Tartu Maarjamõisa polikliinikus kasutatud perearstide infosüsteemis MIS+ sisalduvate arstivisiitide jaotumine kuude kaupa.



Joonis 2. RHK-10 diagnooside ja ravimite anatoomilis-terapeutilis-keemilise klassifikatsiooni (ATC) toimeainete unikaalsete koodide jaotumine aastatel 1995–2011 Tartu Maarjamõisa polikliinikus kasutatud perearstide infosüsteemis MIS+ andmestikus esinemise alusel.

Visiitide ajaline jaotus

Infosüsteem sisaldab 1 353 601 visiidikirjet aastatest 1995–2011. Visiitide jaotumine kuude kaupa on esitatud joonisel 1. Enim visiite toimus sügisesel (oktoobrist novembrini) ja talvisel (jaanuarist märtsini) perioodil ning oluliselt vähem oli perearstide vastuvõtte toimunud kevadel ja suvel.

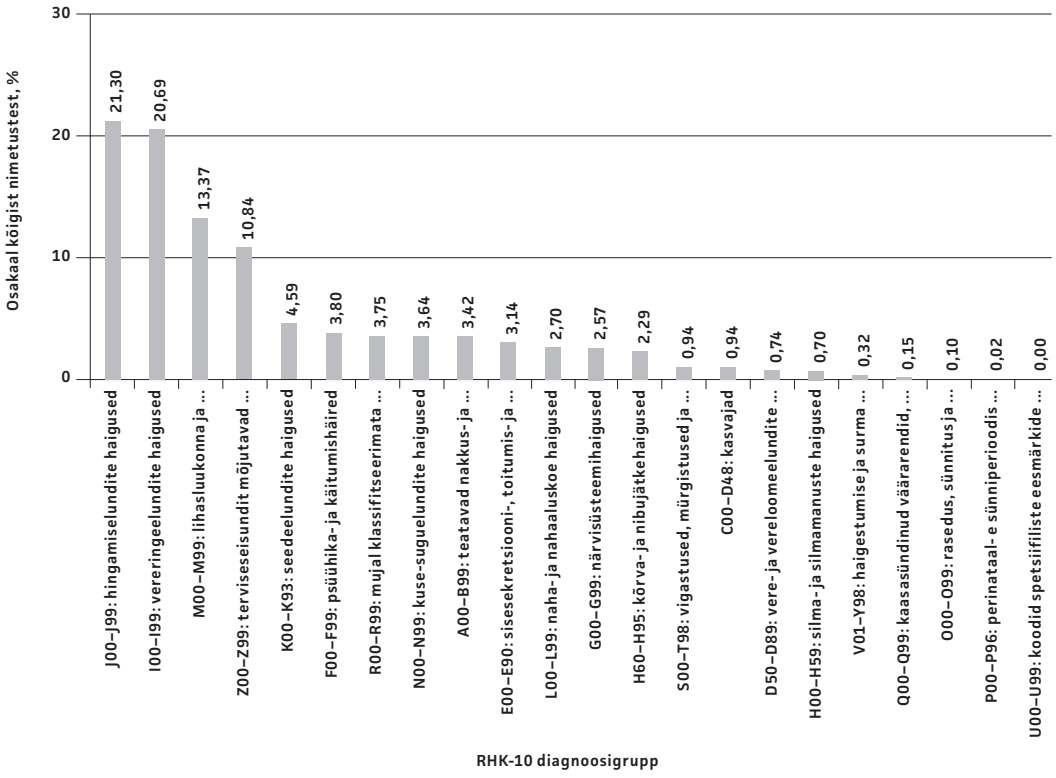
Nädalapäevadest toimus enim visiite esmaspäeviti (23,4%), järgnesid teisipäev (20,3%) ja reede (19,3%). Töönädala vältel toimus kõige vähem visiite kolmapäeviti (18,1%) ja neljapäeviti (17,9%). Umbkaudu 1% visiitidest oli andmestiku kohaselt toimunud nädalavahetustel.

Erinevate diagnooside jaotus andmestikus

Andmestik sisaldas 766 886 RHK-10 kodeeringuga diagnoosi ja 33 970 muud koodi (kaasa arvatud RHK-9 koodid), mis ei kvalifitseeru RHK-10 diagnoosiks. Jooniselt 2 ilmneb, et patsientidele pandud diagnooside sagedused olid üsnagi erinevad. Enamik unikaalseid diagnoose (kokku 5389 erinevat) oli väga väikse esinemissagedusega ja ainult väike rühm diagnoose moodustas suurema osa kogu andmestikust: näiteks 425 (7,9%) unikaalset diagnoosi moodustas 690 216 ehk 90% kõigist andmestikus esinenud diagnoosidest, sarnaselt moodustas 908 (16,8%) unikaalset diagnoosi enam kui 95% (728 543) kõigist esinenud diagnoosidest.

RHK-10 diagnoosigruppide esinemissagedus

Kui vaadelda andmestikus kasutatud RHK-10 koode diagnoosigruppide kaupa (vt joonis 3), on näha, et andmestikus kasutati enim diagnoose J00–J99 (hingamiselundite haigused) ja I00–I99 (vereringeelundite haigused), vastavalt 21,3% (163 367) ning 20,7% (158 737) kõigist diagnoosidest. Samuti võis märgata diagnoosigruppide M00–M99 (luulihaskonna- ja sidekoehaigused) ja Z00–Z99 (terviseseisundit mõjustavad tegurid ja kontaktid tervise-



Joonis 3. RHK-10 diagnoosigruppide jaotumine aastatel 1995–2011 Tartu Maarjamõisa polikliinikus kasutatud perearstide infosüsteemi MIS+ andmestikus.

teenistusega) mõneti laialdasemat kasutust. Muid diagnoosigruppe kasutati juba oluliselt väiksemal määral – ükski neist ei moodustanud üle 5% kasutatud RHK-10 koodide üldarvust.

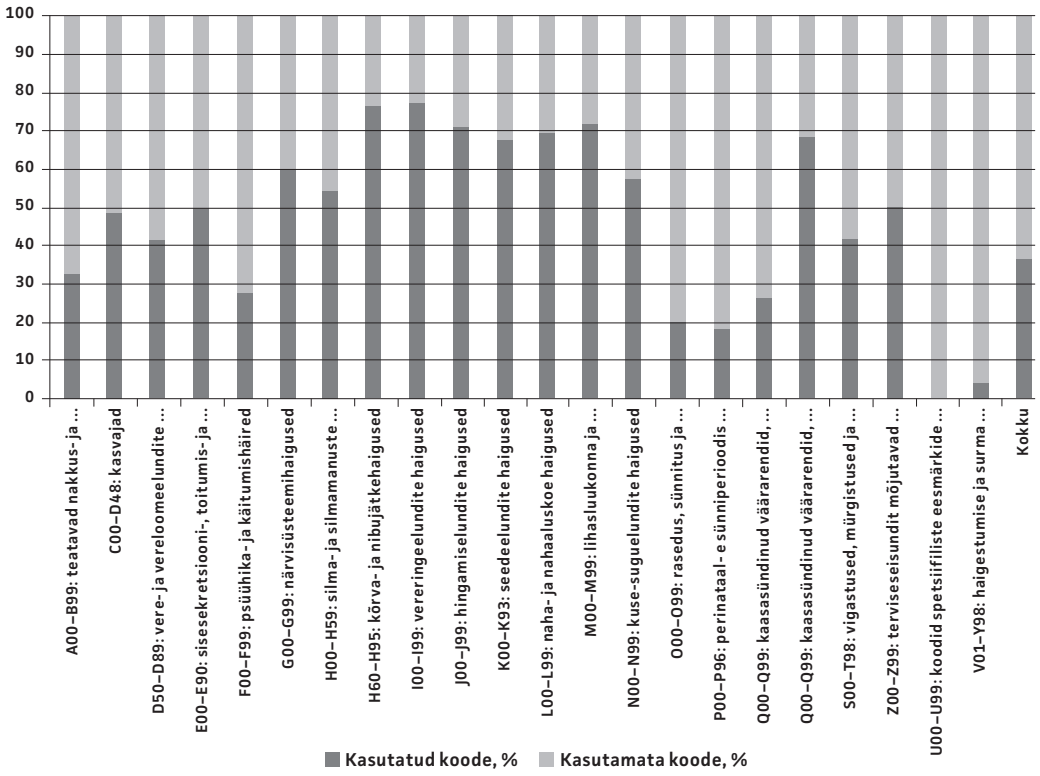
RHK-10 klassifikatsiooni kasutuse ulatus

RHK-10 klassifikatsiooni 14 780 elemendist leidis rakendust 5389 diagnoosikoodi ehk 36,46% ning ülejäänud 9391 koodi ei ole kordagi kasutatud. Joonisel 4 on esitatud diagnoosikoodide kasutus diagnoosigrupiti. Enim erinevaid diagnoose on olnud kasutusel diagnoosigrupis I00–I99 (vereringeelundite haigused) ja H60–H95 (kõrva- ja nibujätkehaigused), vastavalt 77,4% (kokku 455 koodi, kasutatud 352) ja 76,3% (kokku 135, kasutatud 103) kõigist grupis olevatest diagnoosikoodidest. Ka gruppide M00–M99 (luulihaskonna- ja sidekoehaigused) ja J00–J99 (hingamiselundite haigused) puhul oli kasutusel enam kui 70% koodidest (vastavalt 623 koodist 446 ja 277 koodist 195). Vähim rakendust leidsid grupid U00–U99 (koodid spetsiifiliste eesmärkide jaoks) ja V01–Y98 (haigestumise ja surma välispõhjused), vastavalt 0% (2 koodist 0) ja 3,94% (3681 koodist 145). Tehtud analüüsis ei arvestatud RHK-10 koodide omavahelist sõltuvust ega asjaolu, et klassifikaator pole MIS+ kasutuse ajal püsinud muutumatuna.

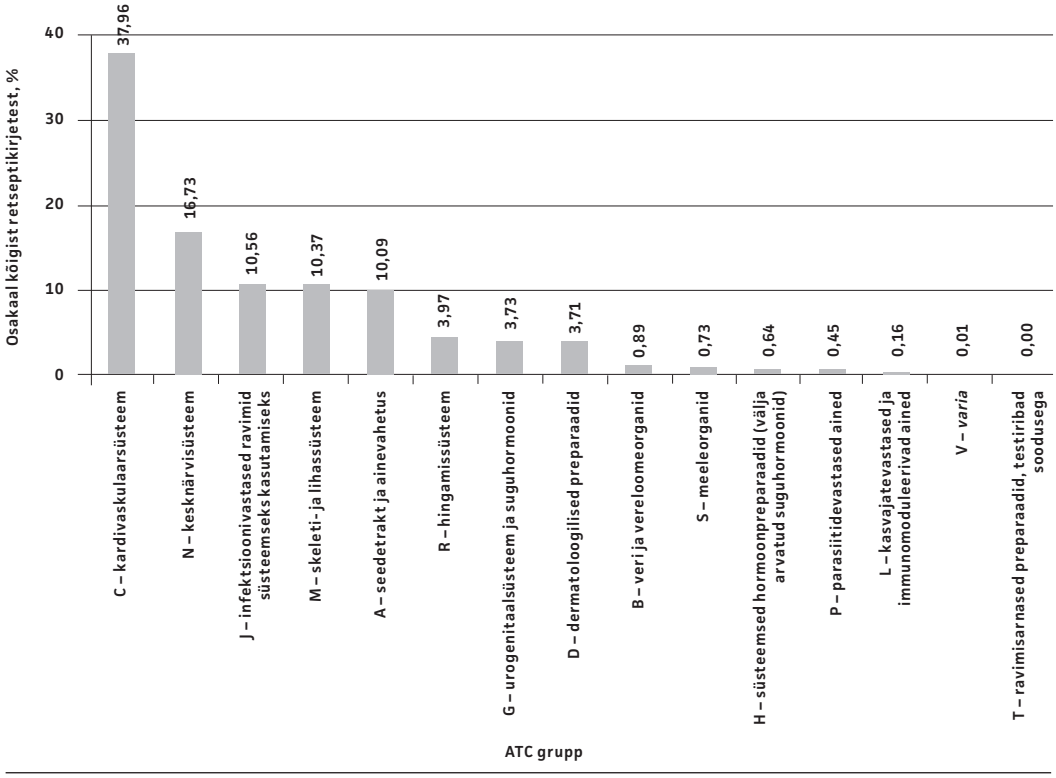
Retseptide toimeainete esinemine ja jaotus

Andmestik sisaldas 451 353 retsepti, mille toimeained on esitatud ravimite anatoomilis-terapeutilis-keemilise klassifikatsiooni (ATC) kodeeringus. Erinevaid toimeaineid leidus andmestikus 718. Et algandmestik sisaldas oluliselt väiksemal määral toimeainepõhiseid retsepte (toimeaine lisandus retseptile kantavate andmete hulka alles 2003. aastal (10)), siis enamikul juhtudest lisati ATC-kood automaatselt analüüsi käigus preparaadi nimetuse põhjal.

Väljakirjutatud retseptide toimeainete jaotust kogu andmestikus kirjeldab joonis 2. Toimeainete jaotuskõver sarnanes mõneti diagnooside kõveraga, s.t 72 (10%) unikaalset toimeainet sisaldus koguni 77,8%-s (351 102) kõikidest väljakirjutatud retseptidest ning 144 (20%) unikaalset toimeainet 90,1%-s (410 584) kõikidest väljakirjutatud retseptidest. Andmestikus sisaldunud retseptide jaotumine nende



Joonis 4. RHK-10 diagnooside kasutus diagnoosigruppide kaupa aastatel 1995–2011 Tartu Maarjamõisa polikliinikus kasutatud perearstide infosüsteemi MIS+ andmestikus.



Joonis 5. Toimeainete väljakirjutamine ravimite anatoomilis-terapeutilis-keemilise (ATC) klassifikatsiooni anatoomilise tasandi gruppide kaupa aastatel 1995–2011 Tartu Maarjamõisa polikliinikus kasutatud perearstide infosüsteemi MIS+ andmestikus.

toimeaine ATC-koodi kõige üldisema taseme alusel on esitatud joonisel 5, kus ilmneb, et ülekaalukalt kirjutati kõige enam välja kardiovaskulaarsüsteemi (C-rühm – 171 354 ehk 37,96% kõigist retseptidest) ja kesknärvisüsteemi (N-rühm – 75 527 ehk 16,73%) toimivaid ravimeid. Samuti on teistest enam välja kirjutatud süsteemseks kasutamiseks mõeldud infektsioonivastaseid ravimeid (J-rühm – 47 653 ehk 10,56%), skeleti- ja lihassüsteemile (M-rühm – 46 822 ehk 10,37%) ning seedetraktile ja ainevahetusele (A-rühm – 45 521 ehk 10,09%) mõeldud ravimeid.

Perearstide visiite kirjeldavad vabatekstilised märkmed

Andmestikus sisalduvad vabatekstilised andmed hõlmavad 1 080 757 perearstide kirjapandud visiidikirjet (visiitide vältel tehtud märkmed) 51 767 patsiendi kohta. Lisaks sellele on andmestikus ka umbkaudu 113 000 radioloogilist kirjeldust (keskmine pikkus 16 sõna) enam kui 45 000 patsiendi kohta ning 43 000 sonograafilist kirjeldust (keskmise pikkusega 26 sõna) enam kui 25 000 patsiendi kohta. Siinses artiklis on neist tekstitüüpidest põgusalt käsitletud vaid esimest.

Perearstidele tehtud visiite puudutavate märkmete keskmiseks pikkuseks on 27 sõna ning neis on enamasti kirjeldatud patsiendi kaebusi, objektiivseid leide, eelnevat ravi, manustatud ravimeid jms. Need tekstid teeb eriti huvitavaks asjaolu, et tegemist ei ole toimetatud tekstidega (nagu terviseloo epikriise sisaldavate tekstide puhul), vaid toorandmetega, mis sisaldavad arstide märkmeid iseendale. See tähendab, et tekstides sisalduvad eelkõige faktid, mida arstid on enda jaoks oluliseks pidanud.

Tekstide keeleliseks analüüsiks kasutati morfoloogilist analüsaatorit ja ühestajat ning kõikidest tekstisõnedest on leitud lemmad ehk algvormid. Tabelis 1 on toodud tekstides sisaldunud sagedamad lemmad, mille hulgast on eemaldatud arvulised väärtused, lausemärgid ja muud sümbolid (nt –, /, ?, %, >, <). Samuti on eemaldatud spetsiifilise tähenduseta sõnad nagu *ja, ei* jt.

Tabel 1. Arstide vabatekstiliste märkmete sagedasemad lemmad aastatel 1995–2011 Tartu Maarjamõisa polikliinikus kasutatud perearstide infosüsteemi MIS+ andmestikus

Lemma ehk algvorm	Absoluutne sagedus	Suhteline sagedus, %	Mitmes visiidis esines	Mitmes visiidis esines, %
rr	294987	1,598	269983	24,981
ravi	291623	1,580	269307	24,918
x	259479	1,405	190355	17,613
olema	230012	1,246	181381	16,783
mg	200748	1,087	147053	13,606
n	165685	0,897	125437	11,606
valu	160230	0,868	133887	12,388
köha	142049	0,769	132668	12,275
fr	136282	0,738	130412	12,067
t	132764	0,719	114786	10,621
kops	128837	0,698	125923	11,651
tvl	119444	0,647	115159	10,655
mmhg	115878	0,628	111722	10,337
cor	109782	0,595	109515	10,133
l	108284	0,587	93519	8,653
pulm	101089	0,548	100680	9,316
vesik	95819	0,519	95456	8,832
parem	91322	0,495	80075	7,409
mgx	88660	0,480	66952	6,195
puhas	86869	0,471	77538	7,174
nohu	83931	0,455	79552	7,361
pt	81691	0,442	71572	6,622
toon	78560	0,426	77928	7,211
päev	77386	0,419	71629	6,628
palavik	75407	0,408	71159	6,584
rp	74610	0,404	67105	6,209
s	73966	0,401	64675	5,984
ii	72224	0,391	65471	6,058
palp	72069	0,390	62305	5,765

ARUTELU JA JÄRELDUSED

Kui vaadelda visiitide jaotumist kuude kaupa (enim oli neid sügistalvisel perioodil), siis on see ühelt poolt selgitatav infektsioonhaiguste suure sagedusega, mis vaibub suvisel-varasügisesel ajal, kuid teiselt poolt nii arstide kui ka patsientide puhkuseperioodiga. Nädalapäevade kaupa on esmaspäevaste visiitide suur arv põhjendatav lõppenud nädalavahetusega, sest üldjuhul sel ajal perearstide vastuvõttu ei toimu. Nädalavahetuse visiidid (umbes 1%) pärinevad 1990. aastate lõpust, kui Maarjamõisa polikliiniku perearstidel oli töö korraldatud selliselt, et valvegraafiku alusel toimus patsientide vastuvõtt ka laupäeviti. Artiklis toodud statistikat visiitide jaotumisest on võimalik kasutada vastuvõtutundide (personali) planeerimiseks – näiteks planeerida esmaspäevaks, teisipäevaks ja reedeks rohkem tunde kui kesknädalaks.

Kui võrrelda andmestikus leiduvate diagnooside esinemissagedust ning võrrelda seda Eurocommunication-II uuringus (11) esitatud tulemustega Eesti perearstide pandud sagedaste diagnooside kohta, siis ilmneb, et mõlemas uuringus on olulisemate diagnoosigruppide kasutussagedused väga sarnased. Näiteks on Eurocommunication-II uuringu raportis välja toodud, et 20,1% kõigist Eesti perearstide pandud diagnoosidest olid seotud hingamisteedega (J00–J99) ning 15,6% seotud luulihaskonna ja sidekoega (M00–M99). Mõneti suurem on erinevus vereringeelundite haigustega seotud diagnooside (I00–I99) osas (Eurocommunication-II raportis 16,7% diagnoosidest, MIS+ süsteemis 20,69%), kuid sama raporti andmete kohaselt olid patsientide visiidi põhjuseks 20,6%-l juhtudest just vereringeelunditega seonduvad probleemid. Niisugune sarnasus osutab ühest küljest meie andmestiku õigsusele, kuid samas viitab ka võimalusele kasutada andmeid ulatuslikumates analüüsides.

Arstide kirjapandud märkmeid analüüsides on võimalik teha üldistusi selle kohta, mida peavad arstid vajalikuks kirja panna. Visiidikirjelduste sõnasagedusi vaadeldes väärib märkimist, et suurima esinemissagedusega sõnad ja lühendid on *rr, ravi, x, olema* ja *mg*. Neist esimest on kasutatud 294 987 korral, kusjuures see on esinenud peaaegu 25%-l visiidikirjetest. See osutab, et sagedasemaks protseduuriks, mida patsientidele tehakse ja mille tulemus ka üles märgitakse, on vererõhu mõõtmine. Ligi veerandi visiitide korral kasutatud ka lemmat *ravi* (kasutatud 291 623 korda) ning 17,6% visiitide puhul lühendit *x* (259 479 korda). *Ravi, mg* ning *x*-i sage kasutamine viitab sellele, et väga tihti arutletakse farmakoterapeutilise ravirežiimi üle. Sõnaühendiga *mg x* märgitakse ravimidoosi, sagedasti esineb tekstides ka tühikuta varianti *mgx*. Lemma *olema* sage kasutamine on eesti kirjakeele sõnastiku järgi tavaline – tegemist on sagedasima lemmaga eesti keeles (12). Laialdaselt on kasutust leidnud ka mitmesugused kaebused ja objektiivsed leiud: valu (12,4%-s visiidikirjetest), *köha* (12,3%-s visiidikirjetest), *vesik* (8,8%-s visiidikirjetest) jt.

Lisaks on oluline tekstide sagedusanalüüsist välja tuua ka mitmesuguste lühendite rohkust. Soomekeelseid haiguslugude epikriise uurides on selgunud, et umbkaudu 7% kõigist tekstisõnedest on lühendid-akronüümid (soomekeelsetes ajalehetekstides kõigest 0,4%), kusjuures nende hulka kuuluvad nii üldkeeles tuntud lühendid nagu *ml* või *kg* kui ka domeenispetsiifilised akronüümid nagu *MRT, TU* jts (13). MIS+ süsteemi kogutud andmeid analüüsides aga selgus, et lausa 14% kõigist tekstisõnadest olid mitmesugused lühendid, võrdlusena on eesti kirjakeeles lühendeid kõigest 1,66% (14).

Samuti ilmnes sõnavara sageduslikust analüüsist, et kokku oli visiidikirjetes kasutatud 190 789 unikaalset lemmat. Väärib märkimist, et neist ainult 78 909 lemmat (41,36%) oli kasutatud enam kui üks kord ning 25 437 lemmat (13,33%) oli kasutatud vähemalt 10 korda. Sellest on järeldatav tõsiasi, et tekstides esineb väga palju trüki- ja kirjavigu ning sarnaselt diagnooside ja ravimitega on väga suur osa tekstidest kirjeldatav väga väikese hulga lemmadega.

Esitatud tulemused on siiski vaid põgus osa sellest, milliseid analüüse on kasutusel olnud andmestiku põhjal võimalik teha. Ühest küljest saab andmestikku kasutada perearstide endi töö edaspidisel korraldamisel, näiteks personali koormuse planeerimisel, näitlike õppematerjalide koostamisel, patsiendi kohta kiire tervikpildi saamise lahenduste loomisel jm. Teisest küljest aga annavad elektroonilise terviseloo kujul talletatud andmed võimaluse analüüsida terviseandmeid näiteks tervishoiukorralduslike otsuste tegemise eesmärgil (raviteenuste vajadus rahvastikurühmiti, vanuserühmiti jms), epidemioloogilistes uurimustes (probleemide esinemine esmatasandi patsientuuri hulgas). Samuti on võimalik teha kaebuste, haiguste kulu ja ravi analüüsi. Oluline oleks ravimite kõrval- ja koostoimete uurimine.

Olgugi et ETL pakub mitmekülgseid võimalusi, on sellel ka mõningaid puudusi. Nii näiteks ei ole andmete tekkimine rangelt kontrollitud protsess, kuna puuduvad ajas muutumatud ühtsed reeglid andmete kodeerimiseks, andmed võivad olla lünklikud või sisaldada süstemaatilisi, kuid analüütikule teadmata vigu. Lisaks on ETLi andmete negatiivseks iseärasuseks asjaolu, et kuigi andmestik võimaldab kirjeldada väga erinevaid tunnuseid (konkreetsed ravimid, diagnoosid, visiidid, analüüsid jms), on üldjuhul iga patsiendi kohta täidetud andmestik siiski väga hõre ja enamiku tunnuste kohta teave puudub (3). Keeruliseks teeb analüüsi see,

et andmete osaline puudumine ei tähenda info välistatust, näiteks ei tähenda diagnoosi puudumine patsiendi anamneesis, et sellega tähistatud haigust pole tal esinenud.

Vaatamata nimetatud puudustele on ETLi andmete kasutamine andmeanalüüsi seisukohalt siiski väga perspektiivikas, kuna suur andmemaht lubab rakendada mitmeid uudseid analüüsialgoritme nii teadaolevate seoste kinnitamiseks kui ka seni teadmata seoste tuvastamiseks. Tarkvara Tehnoloogia Arenduskeskuse uurimisrühm plaanib jätkata MIS+ infosüsteemi andmestiku uurimist mitmes nimetatud suunas. Arstkonnalt on teretulnud relevantsed meditsiinilised probleemid ja hüpoteesid, mida käsitletud andmestiku pealt kontrollida.

VÕIMALIKU HUVIKONFLIKTI DEKLARATSIOON

Artikli autoritel ei ole mingeid konkureerivaid või isiklikke huve teemaga seonduvalt.

SUMMARY

Analysis of the electronic health record dataset of the general practitioners of Tartu

Sulev Reisberg[1,2,6], Raul Sirel[1,3], Ruth Kalda[5], Markko Merzin[1,4], Jaan Pruulmann[7], Jaak Vilo[1,2,6]

Introduction. The aim of the current paper was to introduce the initial data analyses conducted on the basis of the electronic health records of the patients of general practitioners in Tartu. It is a preliminary study demonstrating the potential of the data analysis of such datasets. Estonian electronic health record databases have not been researched on such a large scale earlier.

Methods. The health data of patients were collected to create a database in 1995–2011 and were examined in an anonymised format using different methods of distribution and frequency analysis.

Results. The majority of visits to the general practitioners were made during the autumn and winter seasons, also Monday was the busiest weekday. The visit records contain a total of 18 462 524 words, 14% of them being abbreviations. Altogether 190 789 unique lemmas (base forms) were used in the visit texts; only 78 909 (41.36%) of them were used more than once and 25 437 (13.33%) were used at least10 times. The most frequent words/abbreviations in the visit records were *rr, ravi* (treatment), *x, olema* (to be), and *mg*. The dataset included 5389 different ICD-10 diagnostic codes. The 425 (7.9%) most frequent items accounted for 90% of all diagnoses. The prescriptions comprised 718 different pharmaceutical substances (ATC codes), 144 (20%) most frequent of them accounted for 90% of all prescribed substances.

Conclusion. The results of the study can be used in organizational/regulative decision-making in the healthcare area. The existing dataset is extremely large and has a great potential for further analysis, including usage of data mining tools.

KIRJANDUS/REFERENCES

1. Kalda R, Lember M. Setting national standards for practice equipment. Presence of equipment in Estonian practices before and arter introduction of guidelines with feedback. Int J Qual Health Care 2000;12:59–63.
2. Wiljer D, Urowitz S, Apatu E, et al. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Internet Res 2008;10:e34.
3. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Gen 2012;13:395–405.
4. Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. Lancet 2005;365:475–81.
5. Tatonetti NP, Denny JC, Murphy SN, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther 2011;90:133–42.
6. Tatonetti NP, Fernald GH, Altman RB. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. JAMA 2012;19:79–85.
7. Merilind E. Perearstide töökoormuse dünaamika aastatel 2005–2008. Eesti Arst 2011;90:70–4.
8. Blöndal M, Ainla T, Marandi T, Baburin A, Eha J. Sex-specific outcomes of diabetic patients with acute myocardial infarction who have undergone percutaneous coronary intervention: a register linkage study. Cardiovasc Diabetol 2012;11:96.
9. Pruulmann J, Karu A. Reaalaja statistika praktilises meditsiinis – kogemusi ja probleeme. Meditsiinistatistika ja –registrid. ESS Teabevihik 1997;9:24–9.
10. Sotsiaalministri 26.11.2002 määrus „Ravimite väljakirjutamise ja apteekidest väljastamise kord ning retsepti vorm". RTL 2002, 134, 1964. https://www.riigiteataja.ee/akt/223958.
11. van den Brink-Muinen A, van Dulmen AM, Bensing JM, et al. Eurocommunication II. A comparative study between countries in Central and Western-Europe on doctor-patient communication in general practice. NIVEL: Utrecht, Netherlands; 2003.
12. Eesti kirjakeele sagedussõnastik. http://www.cl.ut.ee/ressursid/sagedused1/.
13. Laippala V, Danielsson-Ojala R, Aantaa K, Salakoski T, Salanterä S. Vocabulary in discharge summaries – the patients and the nurses' perspective. In: Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2013): 4th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2013), 11–12 Feb 2013 Sydney.
14. Sõnaliikide sagedusloend ning käändsõna grammatiliste kategooriate sagedusloendid tasakaalus korpuse põhjal. http://www.cl.ut.ee/ressursid/.

[1] Software Technology and Applications Competence Centre Ltd, Tartu, Estonia
[2] Institute of Computer Science, University of Tartu, Tartu, Estonia
[3] Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia
[4] Information Technology Office, University of Tartu, Tartu, Estonia
[5] Department of Family Medicine, University of Tartu, Tartu, Estonia
[6] Quretec Ltd, Tartu, Estonia
[7] Commit Ltd, Tartu, Estonia

Correspondence to:
Raul Sirel
rsirel@ut.ee

Keywords:
general practitioner, electronic health record, health informatics, data analysis