TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Andmeteaduse õppekava

Annika Talvet

Laborianalüüside diskretiseerimine ja analüüs

Magistritöö (15 EAP)

Juhendaja: Sven Laur, PhD

Tartu 2024

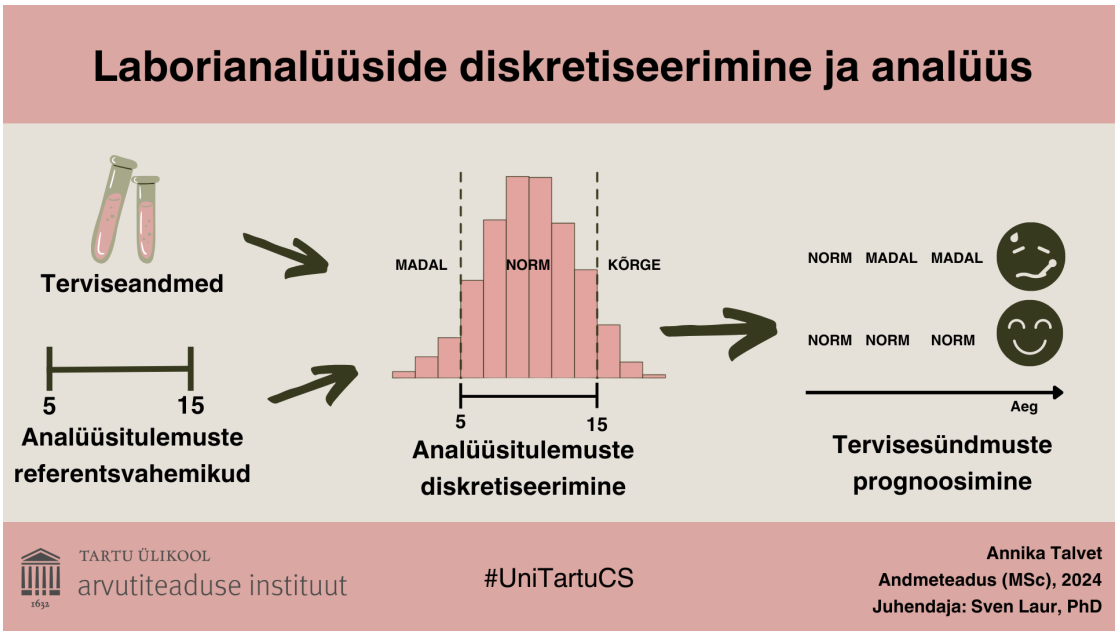
Laborianalüüside diskretiseerimine ja analüüs

Lühikokkuvõte:
Patsientide kliiniliste analüüside tulemuste interpreteerimisel on olulised referentsvahemikud, mis määravad vahemiku, kuhu mõõtmise tulemus võiks jääda terve indiviidi puhul. Need vahemikud võivad sõltuda vanusest ja soost, aga ka konkreetses laboris kasutatavast analüüsimetoodikast. Referentsvahemike abil diskretiseeritud analüüsitulemusi on lihtsam kasutada andmete analüüsimisel ning mudelite treenimisel. Analüüsitulemuste puhul võib aga olla probleemiks seotus vale LOINC koodi või mõõtühikuga.
Magistritöö eesmärk on tuvastada valesti grupeeritud või vale ühikuga analüüse ning referentsvahemikke. Lisaks uurida, kas referentsvahemiku abil diskretiseeritud tulemustest on kasu tervisesündmuste ennustamisel ning kas ennustuse täpsusel on vahe erinevate diskretiseerimismeetodite kasutamisel.
Valesti grupeeritud analüüsitulemuste tuvastamiseks klasterdati andmed kasutades Gaussi segumudelit. Diskretiseeritud tulemuste ennustusvõime hindamiseks uuriti seoseid mõõtmisfaktide ning erinevatel meetoditel diskretiseeritud mõõtmiste ja tervisesündmuse esinemise vahel, lisaks treeniti mudelid tervisesündmuse esinemise prognoosimiseks.
Töö tulemusena selgus, et erinevaid sisendeid kasutavate mudelite ennustustäpsustel ei ole olulist vahet. Selline tulemus viitab asjaolule, et tervisesündmuse ennustamisel on mõõtmise toimumise fakt võrdväärne referentsväärtuste abil diskretiseeritud analüüsitulemusega.

Võtmesõnad:
Laborianalüüsid, referentsvahemikud, klasteranalüüs

CERCS: B110 – Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Visuaalne kokkuvõte:



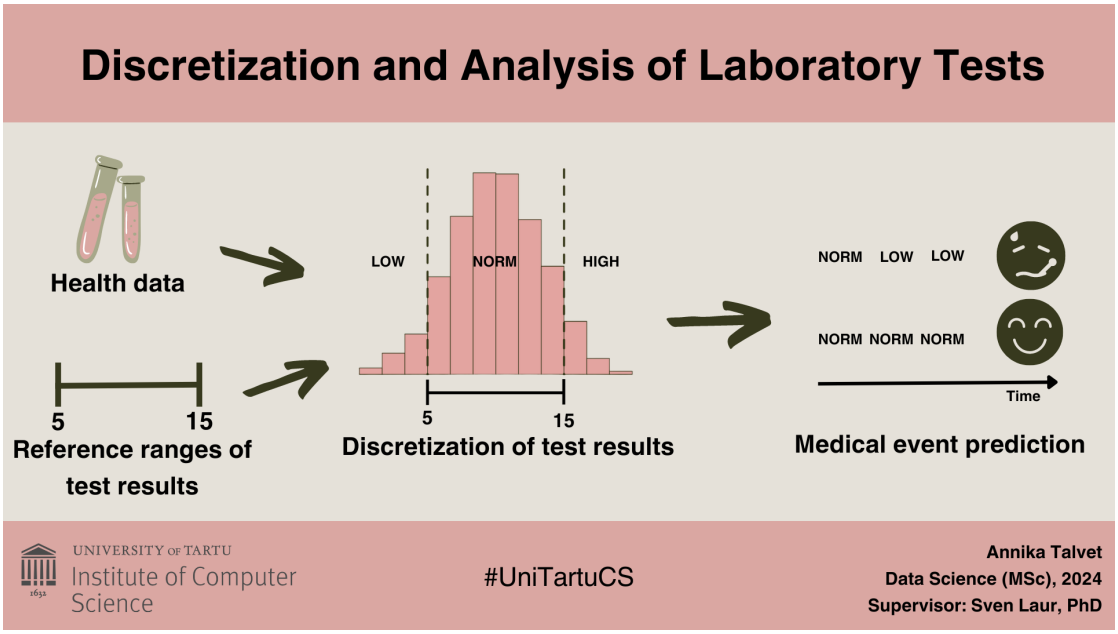
Discretization and Analysis of Laboratory Tests

Abstract:
When interpreting the results of patients' clinical analyses, reference ranges are important as they define the range within which a measurement result could fall for a healthy individual. These ranges can depend on age and gender, but may also vary depending on the methodology used in a particular laboratory. Using analysis results that are discretized based on reference ranges simplifies data analysis and model training. However, analysis results may be associated with incorrect LOINC codes or units of measurement.
The aim of this Master's thesis is to identify analyses and reference ranges grouped incorrectly or with incorrect units. Additionally, it aims to investigate whether discretized analysis results are beneficial for predicting medical events and if there is a difference in prediction accuracy using different discretization methods.
In order to identify incorrectly grouped analysis results, the data was clustered using a Gaussian mixture model. To assess the predictive capability of discretized results, dependencies between the occurrence of medical events and differently discretized measurements, as well as measurement facts, were examined and models were trained to predict the occurrence of medical events.
The results revealed that there is no significant difference in the prediction accuracy between models using different inputs. This suggests that in predicting medical events, the occurrence of measurement is as important as the discretized analysis result.

Keywords:
Laboratory analysis, reference range, cluster analysis

CERCS: B110 – Bioinformatics, medical informatics, biomathematics, biometrics

Graphical abstract:



Sisukord

1 Sissejuhatus

Patsiendi tervisest ülevaate saamiseks viiakse läbi erinevaid laboratoorseid analüüse, seejuures toetub analüüsitulemustele umbes 70% raviotsustest [1]. Laborianalüüside tulemusi on aga keeruline interpreteerida, sest need võivad sõltuda lisaks patsiendi tervislikule seisundile ka näiteks soost ning vanusest. Lisaks ei ole laborianalüüsi puhul tegelikult otseselt oluline selle konkreetne tulemus, vaid kas tulemus jääb tervislikku vahemikku. Seega selleks, et laborianalüüsidest korrektseid järeldusi teha, peavad need olema esitatud koos referentsvahemikega [2]. Referentsvahemikke kasutades diskretiseeritud tulemusi võiks olla lihtsam kasutada hilisemas kirjeldavas analüüsis ning mudelite loomisel.
Erinevad laborid kasutavad sama analüüsi korral erinevaid nimetusi, mis muudavad andmete analüüsimise keeruliseks. Probleemi lahendamiseks võeti Eestis 2016. aastal kasutusele LOINC süsteem, mis määrab igale analüüsile standardse koodi [3]. Tartu Ülikooli Terviseinformaatika uurimisrühma poolt loodud töövoog, mis ühendab muuhulgas iga mõõtmise LOINC koodiga, ei ole aga täielik ega veavaba, seega võivad ühe LOINC koodi alla sattuda erinevad analüüsid. Teine laborianalüüside andmetega seotud probleem on, et tulemused võivad olla märgitud vale ühikuga. Seega tuleb analüüsitulemuste kasutamiseks esmalt üles leida need vead.
Magistritöö eesmärk on tuvastada vigu laborianalüüside tulemuste andmestikus, et lihtsustada hilisemat andmete kasutamist ning parandada neil läbi viidavate analüüside kvaliteeti. Selleks on vaja tuvastada analüüse, mis on seotud kas vale LOINC koodiga või vale ühikuga ning kontrollida referentsvahemike õigsust. Töö teine eesmärk on uurida, kas disketiseeritud analüüsitulemustest on kasu tervisesündmuste ennustamisel ning kas erinevad diskretiseerimise viisid viivad erinevate tulemusteni. Selleks vaadeldakse kahte diskretiseerimise viisi ning kolme erineva tervisesündmuse esinemist tulevikus.
Töö koosneb neljast osast. Esimeses osas antakse ülevaade taustateadmistest, mõistetest ning andmetest. Teises osas kirjeldatakse töös kasutatud meetodeid. Kolmas osa keskendub analüüsitulemuste diskretiseerimisele, vaadeldakse erinevaid probleeme, mis tulenevad vigadest analüüsitulemuste ning referentsväärtuste kirjapanekul. Töö neljandas osas analüüsitakse diskretiseeritud analüüsitulemuste kasutamise eeliseid - uuritakse analüüsitulemuste seoseid tervisesündmuste esinemisega ning diskretiseeritud väärtuste kasutamise kasulikkust prognoosimudelite lihtsustamisel.

2 Taustateadmine ja andmete kirjeldus

Järgnev peatükk annab ülevaate edasises töös esinevatest terminitest ning analüüsimiseks kasutatavatest andmetest. Esimeses alapeatükis on selgitused järgnevatele mõistetele: LOINC, referentsvahemik, ICD ning ATC. LOINCi kasutatakse erinevate analüüside tulemuste eristamiseks, referentsvahemikud on olulised analüüsitulemuste interpreteerimisel ja diskretiseerimisel ning ICD ja ATC koode kasutatakse mudelite ennustatavate väärtuste valimiseks.

2.1 Taustainfo ja terminid

LOINC (*Logical Observation Identifiers Names and Codes*) on rahvusvaheline standardne terminoloogia tervisemõõtmiste, -vaatluste ja -dokumentide identifitseerimiseks, mis on loodud tervisealaste andmete vahetamise lihtsustamiseks ja ühtlustamiseks meditsiinilistes infosüsteemides [4]. Lisaks hõlbustab see terviseandmete analüüsi ja tõlgendamist. Eestis on LOINC süsteem kasutusel alates aastast 2016 [3].
LOINCi eesmärk on tekitada erinev kood igale testile, mõõtmisele või vaatlusele, millel on kliiniliselt erinev tähendus. Selleks arvestatakse iga mõõtmise puhul kuut osa: mõõdetav analüüt; analüüdi omadus; mõõtmise tegemise ajavahemik; materjal, millest proov võeti; kuidas tulemust väljendatakse (kvantitatiivne, järjestatud, nominaalne) ning mõõtmise tegemise meetod [5].

Referentsvahemikke kasutatakse analüüsitulemuste interpreteerimiseks, need määravad väärtused, mille vahele jäävad normaalsed mõõtmistulemused konkreetse analüüsi puhul. Vahemikust välja jääv mõõtmistulemus võib viidata mõnele terviseprobleemile. Referentsvahemikke hinnatakse terve elanikkonna põhjal ning vahemiku otspunktideks võetakse väärtused, mille vahele jäävad 95% tervete inimeste analüüsitulemustest. Seejuures võivad referentsvahemikud sõltuda patsiendi soost ja vanusest [2]. Analüüse läbi viivad laborid võivad määrata referentsvahemikud ise või kasutada kirjanduses esitatud vahemikke [6]. Kuna erinevates laborites kasutatakse erineva mõõtmistäpsusega seadmeid, siis võivad referentsvahemikud sõltuda ka laborist, kus mõõtmine läbi viidi.

ICD (*International Statistical Classification of Diseases and Related Health Problems*) on rahvusvaheline diagnooside, kaebuste ja seisundite standard, mis määrab igale haigusele unikaalse koodi, mis koosneb tähest ja kahest numbrist (paljudel juhtudel saab lisada täpsustuse). Seejuures on sarnased haigused on kokku grupeeritud, näiteks J00-J99 vastavad hingamiselundite haigustele. ICD alusel tehakse haigestumiste ja surma statistikat ning analüüsitakse terviseseisundeid ja -trende. ICD-10 on rahvusvahelise haiguste klassifikatsiooni 10. versioon. [7]

ATC (*Anatomical Therapeutic Chemical*) on rahvusvaheline ravimite klassifitseerimise süsteem, kus toimeained jaotatakse rühmadesse vastavalt nende farmakoloogilistele, terapeutilistele ja keemilistele omadustele ning elundile või süsteemile, millele nad mõjuvad. [8]

2.2 Kasutatavad andmed

Töös kasutatavad andmed pärinevad terviseandmete andmestikust (RITA-MAITT), kuhu on koondatud kolm riiklikku andmekogu - retseptikeskus (välja kirjutatud ja välja ostetud ravimid), Eesti Haigekassa andmekogu (raviarved) ning Tervise infosüsteem (epikriisid ja saatekirjade vastused). Andmestik sisaldab terviseandmeid aastatest 2012-2019 ning on moodustatud juhuvalimiga sisaldades 10% Eesti elanikkonnast (150824 patsienti) [9]. Täpsemalt kasutatakse töös vaid osa RITA-MAITT andmestikust - 8434 patsiendi laborianalüüside, diagnooside, välja kirjutatud ravimite ning haiglas viibitud päevade arvu andmeid. Mudelite testimiseks kasutatakse lisaks 8422 patsiendi andmeid. Seejuures on kõik isikuandmed pseudonüümitud kujul ning täpsete aegade asemel on relatiivsed ajad ehk päevade arv alates esimesest mõõtmisest andmestikus. Laborianalüüside andmestiku kuju on toodud Tabelis 1. Diagnooside andmestikus on iga patsiendi kohta tema diagnoosi saamise relatiivne aeg ning diagnoosi ICD-10 kood. Ravimite andmestikus on iga patsiendi kohta tema ravimi väljakirjutamise relatiivne aeg ning ravimi ATC kood. Haiglaravi andmestikus on iga patsiendi kohta tema haiglas viibimise aja algus (relatiivne aeg) ning haiglas viibimise aja pikkus päevades.

Tabel 1. Andmestiku kuju.

Välja nimi	**Kirjeldus**	**Näide**
Patsiendi id	Pseudonüümitud patsiendi identifikaator	1
Sugu	Patsiendi sugu	N
Vanus, aastad	Patsiendi vanus analüüsi tegemise hetkel, täisaastates	1
Vanus, kuud	Patsiendi vanus analüüsi tegemise hetkel, kuid lisaks aastatele	4
LOINC		731-0
E-labori T lühend	Analüüsi lühend	B-Lymph#
Analüüsi nimi		B-Hemogramm 5-osalise leukogrammiga
Parameetri nimi		Lymph
Parameetri ühik		g/L
		% kogu Hb-st
Standardiseeritud ühik		g/L
		%
Mõõtmise aeg	Relatiivne mõõtmise aeg päevades	403
Mõõtetulemus		8,41
		Kergelt vähenenud eGFR (75)
		Negatiivne (0,548) COI
Standardiseeritud arvuline mõõtetulemus		8,41
		75
		0,548
Standardiseeritud tekstiline mõõtetulemus		NA
		NA
		Negative
Referentsvahemik		(1,8;10)
		neg<0,8; pos>1,1
		M:<2,3;N:<2,8
Referentsvahemiku alguspunkt	Referentsvahemikust eraldatud alguspunkt	1,8
Referentsvahemiku lõpppunkt	Referentsvahemikust eraldatud lõppunkt	10

3 Metoodika

3.1 Gaussi segumudel müra komponendiga

Gaussi segumudelit kasutatakse andmete klasterdamiseks sobitades neile segu normaaljaotustest. Mudel on defineeritud järgmiselt:
Olgu Z juhuslik suurus, mille võimalikud väärtused on 1,...,K tõenäosusega $P\{Z = k\} = \pi_k, k = 1,...,K$. Juhuslik vektor X on juhuslike komponentide $X_1,...,X_K$ segu, kui selle tihedusfunktsioon avaldub järgmiselt

$$f(x;\theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k),$$

kus $\pi_1,...,\pi_k$ on komponentide kaalud, $\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$, f_k on komponendi X_k tihedusfunktsioon, θ_k on tiheduse parameetrite vektor ja $\theta = \{\pi_1,...,\pi_K,\theta_1,...,\theta_K\}$ tähistab segujaotuse kõigi parameetrite hulka.
Gaussi segumudeli puhul eeldatakse, et komponendid on mitmemõõtmelisest normaaljaotusest $X_k \sim \mathcal{N}(\mu_k,\Sigma_k)$, kus k-nda komponendi tihedusfunktsioon on

$$f_k\left(x \mid \theta_k\right) = f_k\left(x \mid \mu_k, \Sigma_k\right) = \frac{1}{\det\left(2\pi\Sigma_k\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right)\right)$$

μ_k on k-nda komponendi keskväärtuste vektor ja Σ_k on k-nda komponendi kovariatsioonimaatriks. [10]
Mudeli robustsemaks muutmiseks lisatakse sellele mürakomponent. See tähendab, et lisaks normaaljaotusega komponentidele lisatakse mudelisse ühtlase jaotusega komponent, mille eesmärk on kokku koguda erindid.

3.2 Bayesi informatsioonikriteerium

Bayesi informatsioonikriteeriumi (BIC) abil saab valida parima mudeli ehk klasterdamise puhul optimaalseima klastrite arvu. BIC sõltub mudeli logaritmitud maksimiseeritud tõepärast, millele on lisatud karistusliige, mis sõltub mudeli parameetrite arvust ja valimimahust.

$$BIC = -2\ln L(\widehat{\boldsymbol{\theta}}) + v\ln n,$$

kus $L(\widehat{\boldsymbol{\theta}})$ on mudeli maksimiseeritud tõepära, v on mudeli parameetrite arv ja n on valimi suurus. [11]
Bayesi informatsioonikriteeriumile lisatud karistusliige tähendab, et eelistatakse lihtsamaid mudeleid. Optimaalseima klastrite arvu leidmiseks arvutatakse BIC väärtus erinevate klastrite arvuga mudelite korral ning valitakse minimaalse BIC väärtusega mudel.

3.3 Otsustusmets

Otsustusmets (*Random Forest*) kuulub ansambelõppe (*ensemble learning*) meetodite alla, mille puhul lõppennustus moodustub mitme mudeli ennustuste kombinatsioonist. Otsustusmetsa moodustavad hulk otsustuspuid, millest kõik on treenitud erineval *bootstrap*'itud osal treeningandmestikust. *Bootstrap*'imine tähendab, et treeningandmestikust võetakse tagasipanekuga juhuslik valim. Mudelit saab kasutada nii klasifitseerimis- kui regressioonülesande lahendamiseks, klassifitseerimisel on lõppennustus kõigi otsustuspuude ennustuste hulgas enim esinenud väärtus, regressiooni korral kõigi otsustuspuude ennustuste keskmine. Otsustusmets tuleb toime erindite ja müraga andmestikus ning töötab hästi ka suure arvu tunnuste korral, *bootstrap*'imine vähendab mudeli ülesobitamist. Negatiivsest küljest on mudeli treenimine arvutuslikult kulukas ning võtab kaua aega, lisaks on selle tulemust keeriline interpreteerida. [12]

3.4 Logistiline regressioon

Logistiline regressioon on statistiline mudel, mida kasutatakse binaarsete tunnuste ennustamiseks. Mudel ennustab sündmuse esinemise tõenäosust ühe või mitme pideva või diskreetse tunnuse põhjal. Logistiline regressioon on populaarne, sest seda on kerge interpreteerida ning lihtne rakendada. [13]

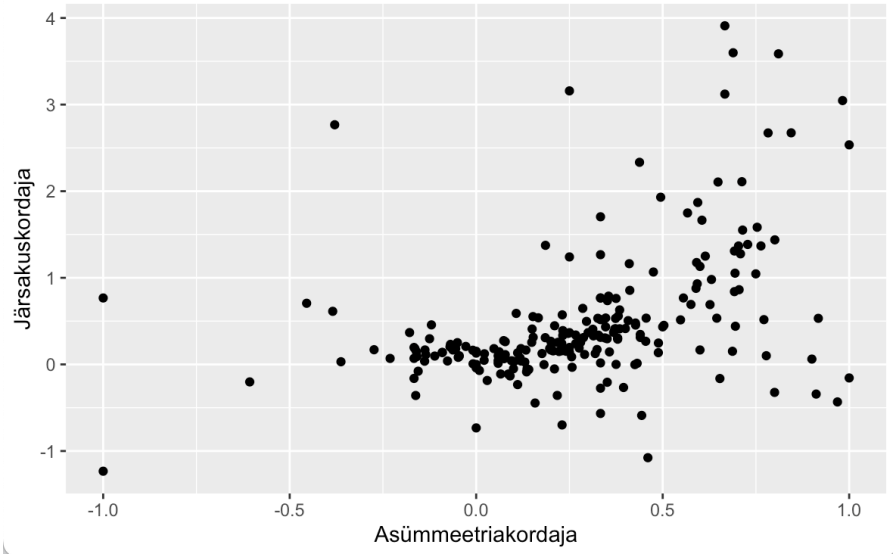
3.5 AUC

Binaarse tunnuse ennustusmudeli hindamiseks kasutatakse erinevaid mõõdikuid, üks neist on AUC (*Area Under ROC Curve*). AUC hindab, kui hästi suudab mudel eristada kahte klassi. AUC väärtused jäävad $0{,}5$ ning 1 vahele, kus $0{,}5$ tähendab, et mudeli ennustusvõime on sama hea kui juhuslikul mudelil ning 1 viitab ideaalsele vahe tegemisele kahe klassi vahel. [14]

4 Analüüsitulemuste diskretiseerimine

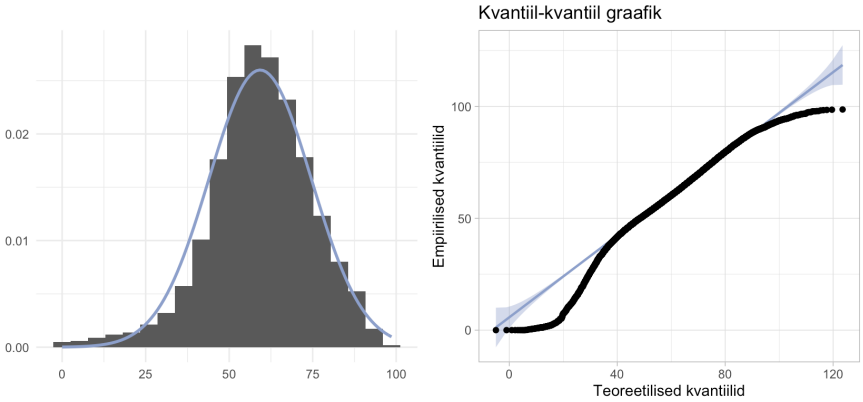
4.1 Analüüsitulemuste klasterdamine

Töös kasutatav terviseandmete andmestik sisaldab vigu, mida küll järjepidevalt parandatakse, aga mis teevad selle põhjal analüüside läbiviimise keeruliseks. Näiteks on ühe analüüsi tulemused märgitud valede mõõtühikutega või on tulemused seotud vale LOINC koodiga. Seetõttu võivad ka tervete inimeste tulemused omada pealtnäha mitmekordseid erinevusi. Selleks, et selliseid probleeme tuvastada, klasterdame analüüsitulemused.
Bioloogiliste mõõtmiste tulemused on enamasti normaaljaotusega, seega lähendame analüüsitulemusi normaaljaotusega. Lisaks vaatame lähendamist logaritmilise normaaljaotuse ning eksponentjaotusega. Seejuures kasutame robustset lähendamist, mis tähendab, et parameetrite hindamisel jäetakse välja 5% vähimatest ning 5% suurimatest väärtustest, see aitab kõrvaldada erindite mõju parameetrite hinnangule.
Joonisel 1 on toodud robustselt hinnatud mõõtmistulemuste järsakus- ja asümmeetriakordajad grupeerituna LOINC koodide kaupa, joonise selguse huvides on välja jäetud suure järsakuskordajaga erindid. Mõõtmistulemuste järsakuskordajad on enamasti nullilähedased. Et normaaljaotuse järsakuskordaja on null, saame öelda, et nullilähedaste jaotuste järsakus sarnaneb normaaljaotuse järsakusele. Osa järsakuskordajaid on tugevalt positiivsed, see viitab terava tipu ja raskema sabaga jaotusele. Asümmeetriakordajatest on enamik positiivsed, mis viitab sellele, et analüüsitulemuste jaotused on paremale kaldu.

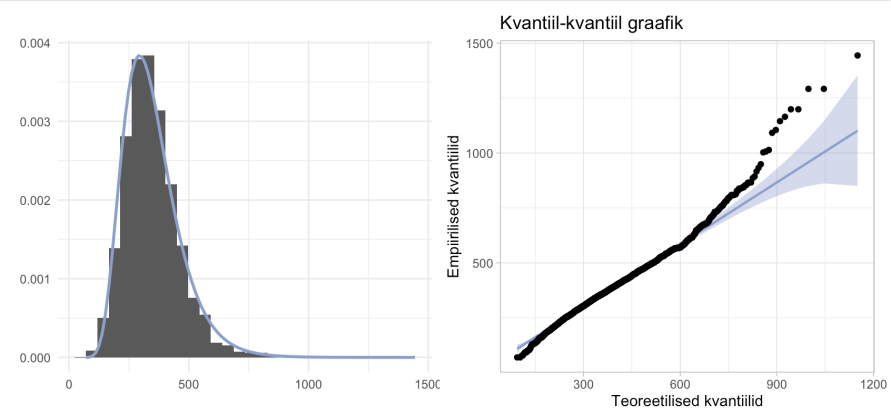


Joonis 1. Järsakus- ja asümmeetriakordajad

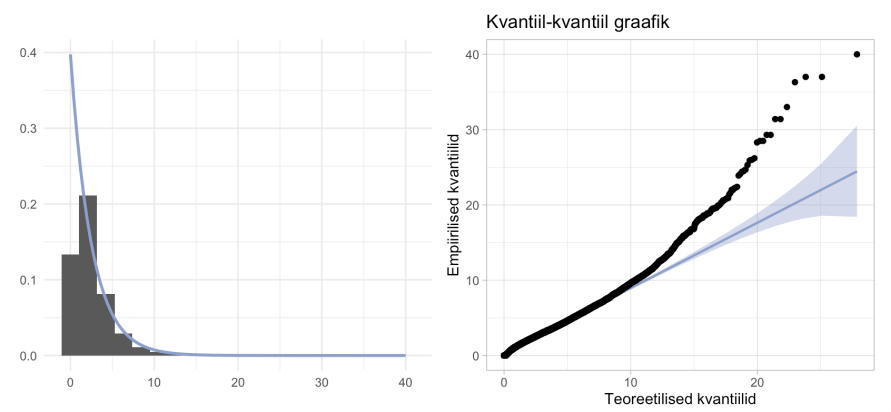
Kasutades robustset normaaljaotuse, logaritmilise normaaljaotuse ning eksponentjaotuse parameetrite hindamist, on tulemuseks, et 239 analüüdist 33 saab lähendada normaaljaotusega, 26 logaritmilise normaaljaotusega ning 180 eksponentjaotusega.
Joonistel 2-4 on kolme analüüdi mõõtetulemuste jaotuste histogrammid, mille peale on joonestatud sobitatud teoreetiliste jaotuste tihedusfunktsioonid. Histogrammi kõrval on valimi jaotuste võrdlemiseks teoreetiliste jaotustega kvantiil-kvantiil graafikud (*q-q plot*) koos 95% usaldusintervalliga. Kui valimi tegelik jaotus ning sellele sobitatav teoreetiline jaotus ühtivad, siis paiknevad punktid kvantiil-kvantiil graafikul ligikaudu ühel sirgel.
Näeme, et neutrofiilide protsendi jaotusele (joonisel 2) saab lähendada normaaljaotust, kusihappe jaotusele (joonisel 3) logaritmilist normaaljaotust ning eosinofiilide protsendi jaotusele (joonisel 4) eksponentjaotust. Kõigi analüütide puhul mahub histogramm peaaaegu teoreetilise jaotuse tihedusfunktsiooni alla ning kvantiil-kvantiil graafikul paiknevad puhktid ligilähedaselt sirgel.



Joonis 2. Neutrofiilide jaotus

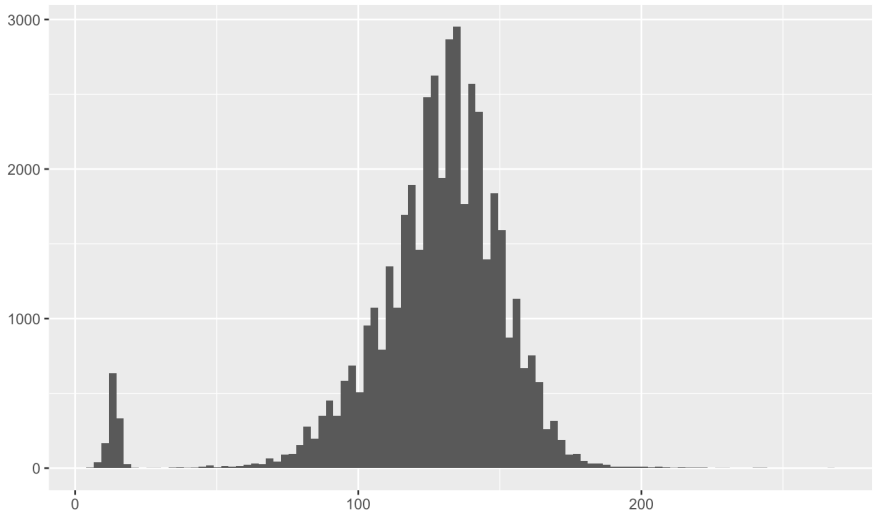


Joonis 3. Kusihappe jaotus



Joonis 4. Eosinofiilide jaotus

Osa analüütide puhul eristub mõõtetulemuste jaotustel mitu kühmu, mis võib viidata erinevate mõõtühikute kasutamisele. Näiteks hemoglobiini mõõtmistulemuste jaotuse puhul joonisel 5 on selgelt näha kahte kühmu.



Joonis 5. Hemoglobiini jaotus

Selliste jaotuste tuvastamiseks ning kühmude arvu määramiseks klasterdame tulemused kasutades Gaussi segumudelit. Et eemaldada erindite mõju, lisame mudelile mürakomponendi. Klasterdamisel kasutame R-i paketti *mclust*, mille puhul kasutatakse parameetrite hindamiseks EM-algoritmi. Optimaalse klastrite arvu valime kasutades Bayesi informatsioonikriteeriumi (BIC).

4.1.1 Klasterdamise tulemused

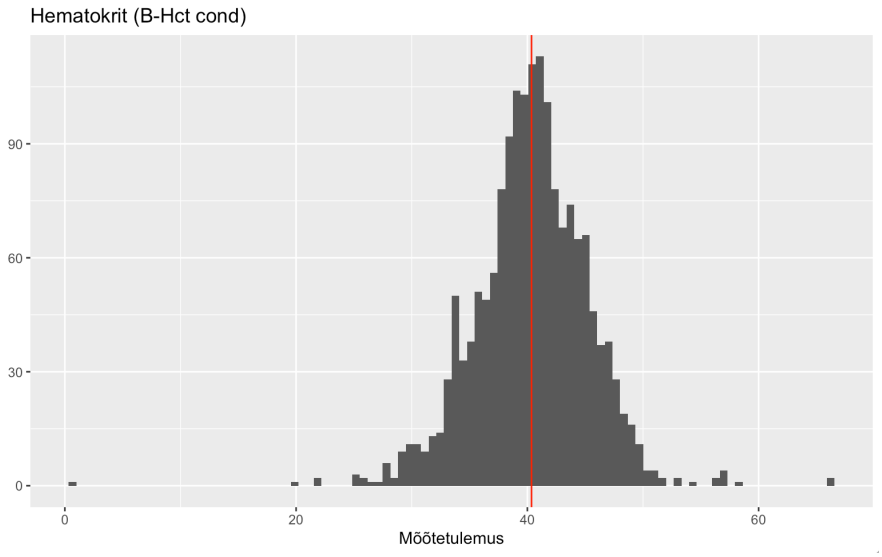
Klasterdamise eesmärk on tuvastada üksteisest erinevate analüüsitulemuste gruppe. Sellised grupid võivad põhjustada vead andmestikus ehk erineva mõõtühikuga märgitud tulemused või vale LOINC koodi alla sattunud mõõtmised. Teine võimalik gruppide tekkepõhjus on tervete ja haigete inimeste analüüsitulemuste erinevus. Soovime siinkohal tuvastada just vigu andmestikus.
Kui segumudeli optimaalseim klastrite arv on üks, siis üksteisest erinevaid analüüsitulemuste gruppe ei esine ning seega eeldadame, et tulemuste märkimisel vigu tehtud ei ole. Kui segumudeli optimaalseim klastrite arv on kaks või enam, siis see viitab teistest erinevate analüüsitulemuste gruppide olemasolule ning potentsiaalsete vigade esinemisele andmestikus.
Et tuvastada, kas gruppide tekkepõhjus on vead andmestikus või tervete ja haigete inimeste erinevused, vaatleme klastrite keskpunktide erinevusi ning võrdleme klastrite arvu erinevate mõõtühikute arvuga andmestikus.

Klasterdasime kõigi analüütide mõõtetulemused, mille puhul oli mõõtmisi vähemalt 300. Tabelist 2 võib näha, et üks klaster moodustus vaid üheksa analüüdi puhul. Suurem osa analüüte klasterdati kaheks klastriks ning kõige suurem optimaalne klastrite arv oli seitse.

Tabel 2. Optimaalsed klastrite arvud.

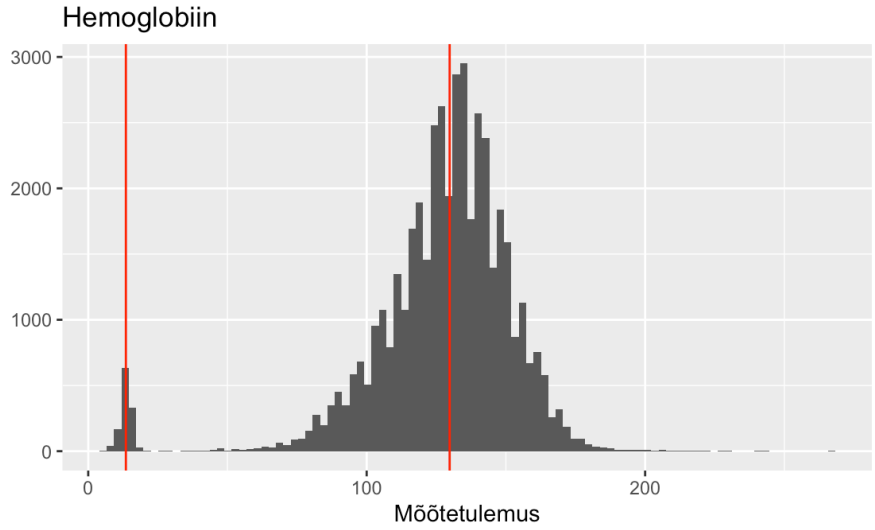
Optimaalne klastrite arv	**Analüütide arv**
1	9
2	192
3	18
4	5
5	1
6	2
7	2

Hematokriti mõõtmiste juures on optimaalseim klastrite arv üks. Selle analüüdi puhul paistab ka jooniselt 6 üks klaster ning andmestikus esineb sellel üks mõõtühik.



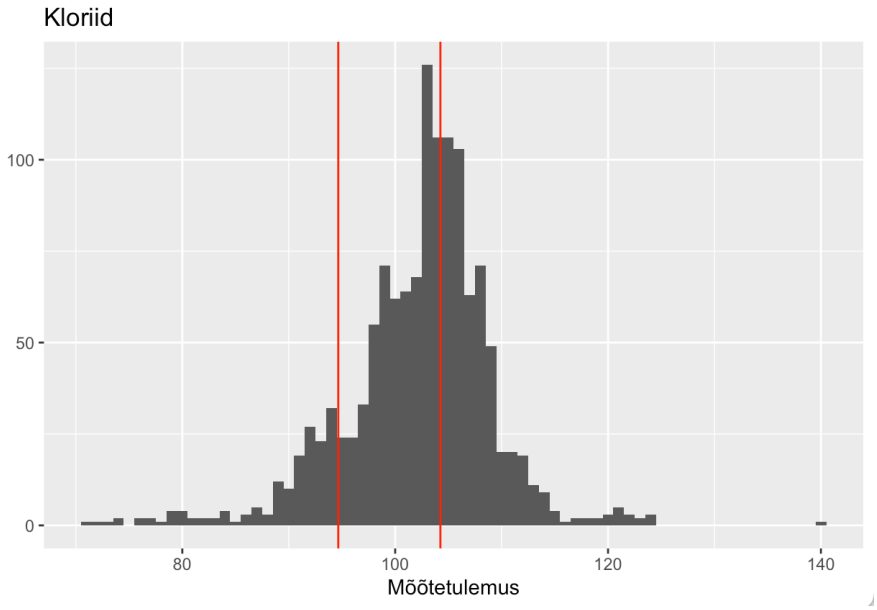
Joonis 6. Hematokriti klasterdus.

Hemoglobiini puhul (joonis 7) on optimaalne klastrite arv kaks, mille keskpunktid on 13,53 ja 129,80 ehk keskpunktide vahe on ligikaudu kümnekordne. Erinevate mõõtühikute vahe on enamasti samuti kümnekordne ehk siin viitab klastrite keskpunktide vahe erinevate mõõtühikute kasutamisele bioloogiliste erinevuste asemel. Saadud tulemuse valideerimiseks vaatame andmestikus esinevaid mõõtühikuid hemoglobiini mõõtmistulemuste juures. Mõõtmisi on märgitud kahe mõõtühikuga: g/dL ja g/L. Et ka andmestikus esinevate mõõtühikute vahe on kümnekordne, saame järeldada, et kahe klastri tekkimise põhjus ongi erinevate mõõtühikute kasutamine. Segumudeli sobitamine aitas meil tuvastada eri mõõtühikutega mõõtmisi isegi siis, kui me tõenäosuslikku mudelisse sisse ei kodeerinud ühikute erinevust.



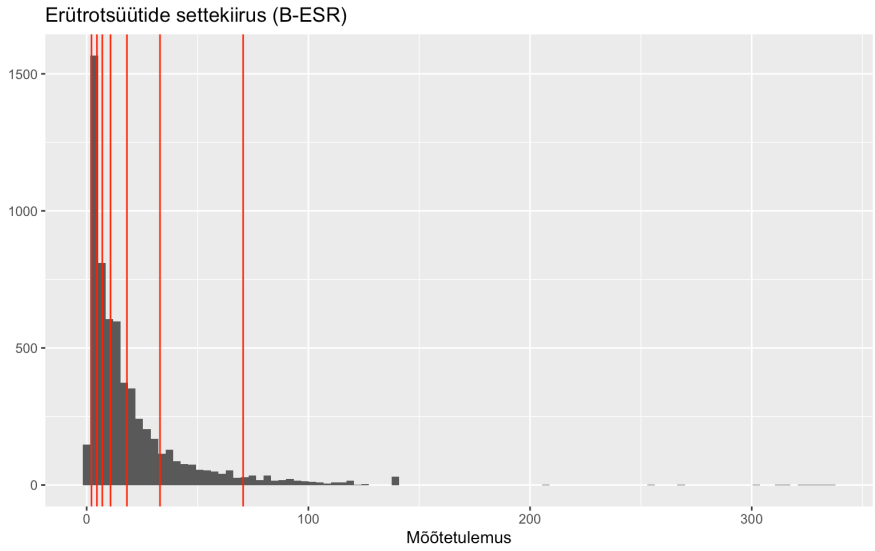
Joonis 7. Hemoglobiini klasterdus.

Jooniselt 8 paistab, et kloriidi mõõtetulemuste korral on parim klastrite arv samuti kaks ning klastrite keskpunktid on 94,65 ja 104,25. Selline vahe on liiga väike, et viidata mõõtühikute erinevusele ning ka andmestikus on selle analüüdi mõõtetulemuste märkimisel kasutatud ühte ühikut. Seega on võimalik, et mudel eristab siinkohal tervete ning haigete inimeste tulemusi.



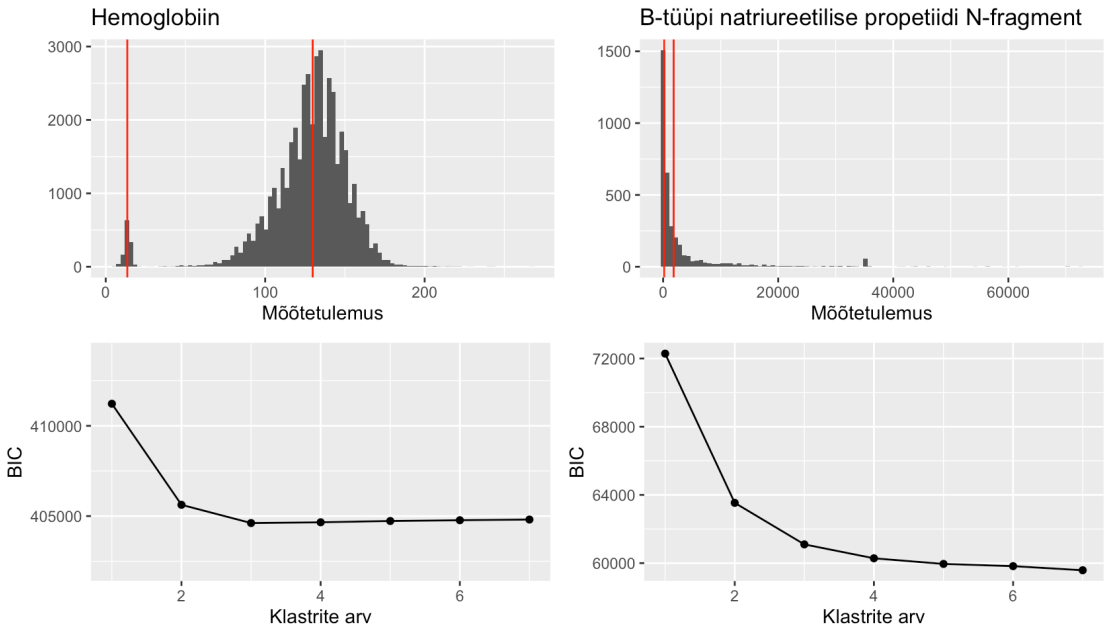
Joonis 8. Kloriidi klasterdus.

Klasterdused, mille puhul on optimaalne klastrite arv suurem kui kaks, tõid esile pigem Gaussi segumudeli puuduseid. Näiteks erütrotsüütide settekiiruse mõõteandmete optimaalseim klastrite arv on seitse, jooniselt 9 on aga näha probleem, et selle analüüdi andmed ei ole mitte normaal- vaid eksponentjaotusega. Seega peaks proovima eksponentjaotusega analüütide klasterdamist läbi viia segumudeliga, mille komponendid on eksponentjaotusest. Sellise mudeli parameetreid saab hinnata näiteks kasutades Pythoni *TensorFlow Probability* paketti [15].



Joonis 9. Erütrotsüütide settekiiruse klasterdus.

Jooniselt 10 paistab, et näiteks normaaljaotusega hemoglobiini puhul saame erinevate klastrite arvudega mudelite BICe võrreldes suhteliselt kindlalt väita, et kahe klastriga mudel on kõige optimaalsem. Eksponentjaotusega B-tüüpi natriureetilise propetiidi N-fragmendi puhul jätkab klastrite arvu kasvades BIC langemist ehk nii selget optimaalset klastrite arvu ei ole.
Siinkohal ei ole seega Gaussi segumudel optimaalne. Tuvastasime, et suure osa analüütide mõõtetulemused on normaaljaotuse asemel eksponentjaotusega ning nende puhul oleks otstarbekam kasutada segumudelit, mille komponendid on eksponentjaotusega. Lisaks võib olla probleemiks see, et tervete ja haigete inimeste analüütide jaotused on erinevad ning Gaussi segumudel tuvastab andmestiku probleemide asemel erinevas tervislikus seisus olevate inimeste gruppe.



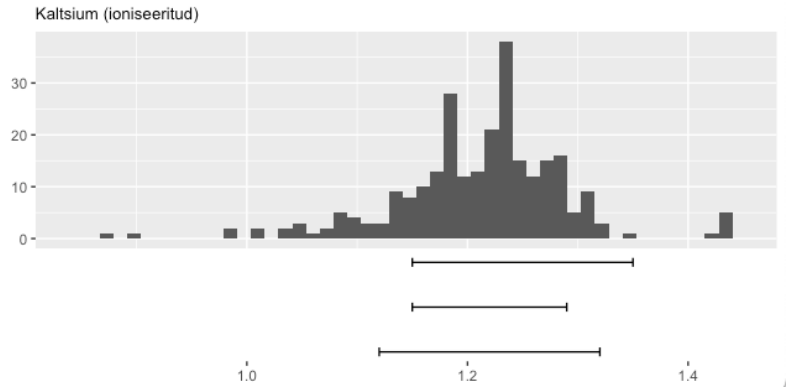
Joonis 10. Erinevate klastrite arvuga mudelite BIC väärtused.

4.2 Referentsväärtused

Analüüsitulemused võivad lisaks tervislikule seisundile sõltuda ka soost ja vanusest ning kohati ka laborist, kus analüüs läbi viidi. Seetõttu ei ole piisav vaadata vaid tulemust ennast, vaid tulemust koos referentsväärtustega. Labor edastab analüüsitulemused koos referentsväärtustega, seega peaks ka töös kasutatavas andmestikus iga analüüsitulemuse juures olema selle referentsvahemik. Probleemiks on aga suur hulk puuduvaid vahemikke ning kontrollida tuleks olemasolevate vahemike õigsust.

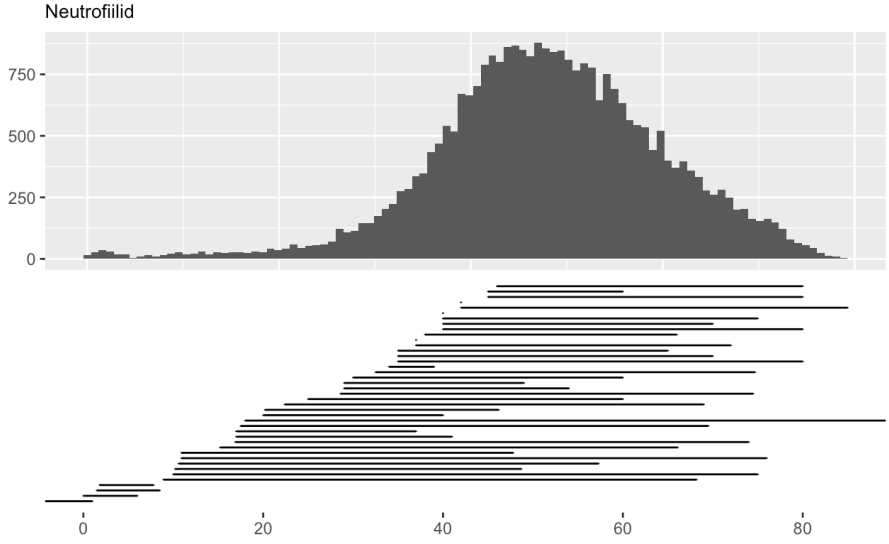
4.2.1 Mitme referentsvahemiku probleem

Ühel analüüdil võib olla mitu erinevat referentsvahemikku. Joonisel 11 on ioniseeritud kaltsiumi referentsvahemikud, mille otspunktid veidi erinevad, aga vahemikevaheline kattuvus on suur. Selline erinevus viitab laboritevahelisele mõõtetäpsuse kõikumisele.



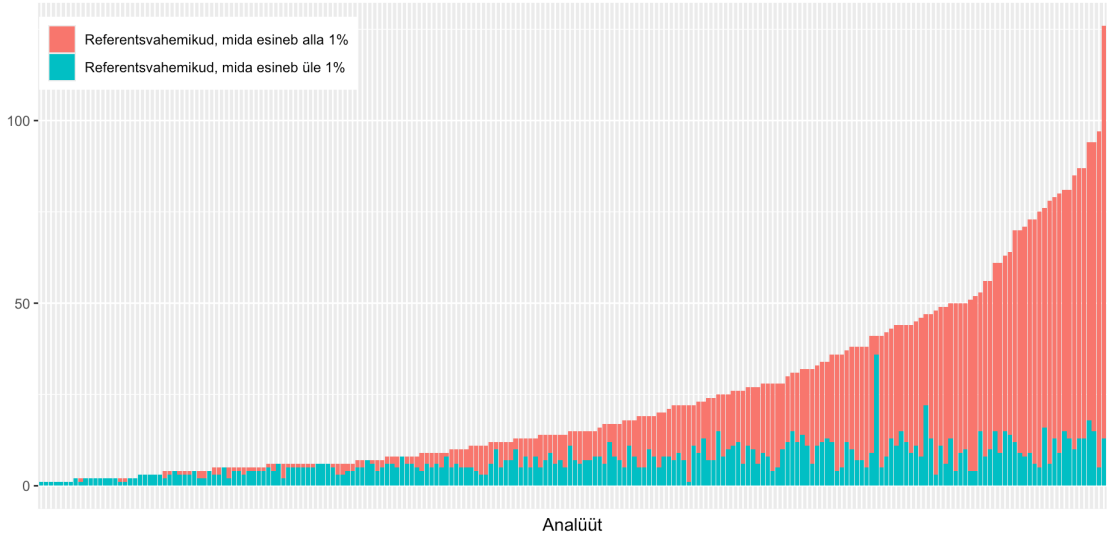
Joonis 11. Ioniseeritud kaltsiumi referentsvahemikud.

Kohati on referentsvahemikke ühe analüüdi puhul palju erinevaid, näiteks neutrofiilide mõõtetulemusel on erinevaid referentsvahemikke 41, need on toodud joonisel 12. Lisaks paistab, et referentsväärtuste varieeruvus on suur ning osa vahemikke ei ole isegi kattuvad.



Joonis 12. Neutrofiilide referentsvahemikud.

Joonisel 13 on referentsvahemike arvud analüütide kaupa ning erinevate värvidega on tähistatud vahemike arvud, mida esineb vähem või rohkem kui 1% andmestikust. $96{,}8$ protsendil analüütidest esineb andmestikus vähemalt kaks erinevat referentsvahemikku ning erinevate referentsvahemike arvud küündivad 126-ni. Seejuures on vahemikke, mida esineb rohkem kui 1% analüüdi kohta maksimaalselt 36 ehk referentsvahemike arvu tõstavad palju vahemikud, mida esineb vaid üksikult.



Joonis 13. Erinevate referentsvahemike arv analüüdi kohta.

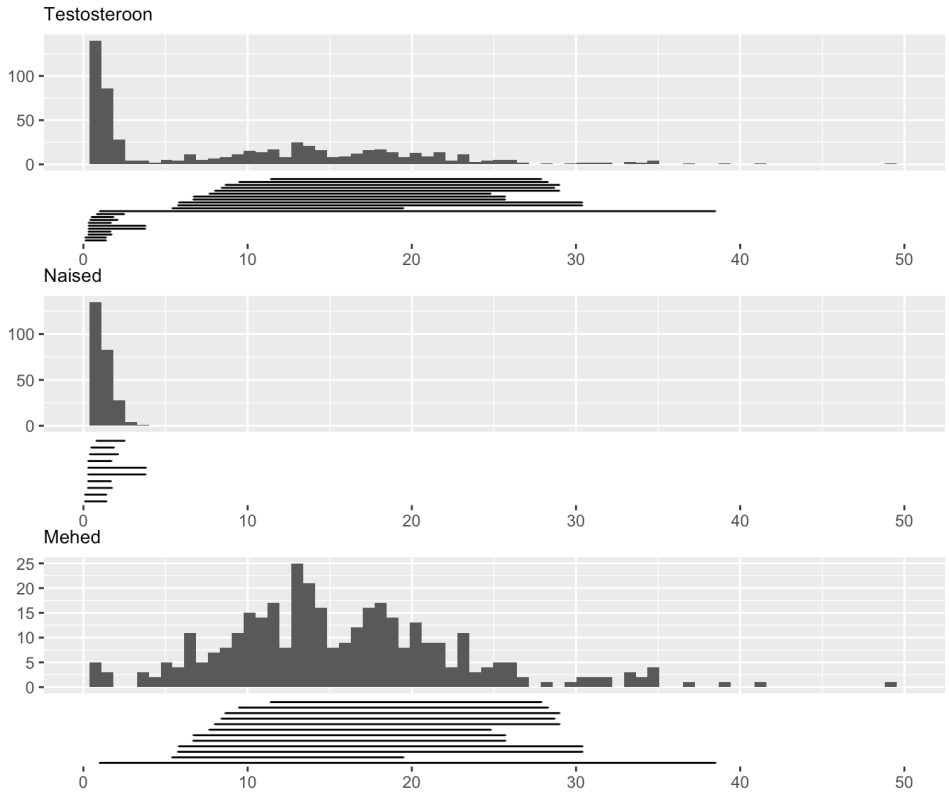
Referentsvahemike rohkuse ja varieeruvuse põhjuseks võib olla nende sõltuvus soost ja vanusest või vead andmestikus. Vigade põhjuseid võib olla mitu: analüüdid on seotud vale LOINCiga, analüüsid on esitatud erinevate mõõtühikutega või on tehtud laboripoolne viga allikandmetes.
Esiteks võime eeldada, et vahemikud, mida esineb üksikult ehk analüüdi kohta vähem kui 1%, on vead. Selle eelduse tegemisel väheneb referentsvahemike arv analüüdi kohta kohati märkimisväärselt, nagu selgus jooniselt 13.
Edasi uurime allesjäänud referentsvahemike puhul, milliste analüütide referentsvahemike varieeruvuse põhjus võib olla nende sõltuvus soost ja vanusest.
Soost sõltuvuse uurimiseks võrdleme naiste ja meeste referentsvahemike keskpunkte. Kasutame Wilcoxoni astaksummatesti, et teha kindlaks, kas kahe grupi keskpunktide vahel on statistiliselt oluline erinevus.
16 analüüdil on andmestikus vaid üks erinev referentsvahemik või on referentsvahemikke märgitud ainult ühe soo puhul, seetõttu meeste ja naiste vahelist erinevust nende juures uurida ei saa. Ülejäänud analüütidest 91 puhul meeste ja naiste referentsvahemike keskpunktide vahel olulist erinevust ei esine (Wilcoxoni astaksummatesti FDR korrektsiooniga p-väärtus suurem kui $0{,}05$). Oluline erinevus esineb 117 analüüdil.
Võrdleme saadud tulemusi Eesti suurima meditsiinilabori Synlabi referentsvahemike soolise jaotusega [16]. Analüüte, mis esinevad Synlabi andmetes ning millel on andmestikus mitu referentsvahemikku ja mõõtmisi tehtud kahe soo puhul, on kokku 107. Tabelis 3 on võrdlus Synlabi soolise jaotuse ning andmestikus esinenud soolise jaotuse vahel. 67 analüüdi korral on Wilcoxoni testi tulemus ning Synlabi andmestikus esinenud jaotus samad. 38 analüüti, mille puhul Synlabi kohaselt sugude referentsvahemikel vahet ei ole,

klassifitseeris Wilcoxoni test soost sõltuvaks.

Tabel 3. Referentsvahemike sõltuvus patsiendi soost

		Synlab	
		Sõltub soost	**Ei sõltu soost**
Wilcoxoni test	**Sõltub soost**	32	38
	Ei sõltu soost	2	35

Testosterooni puhul on Wilcoxoni testi kohaselt sugude referentsväärtuste keskpunktidel oluline vahe, sugude referentsvahemikel tehakse vahet ka Synlabi andmestikus. Testosterooni referentsvahemikud on joonisel 14. Selgub, et referentsvahemike keskpunktide võrdlemine vahemikevaheliste erinevuste tuvastamiseks on tulemuslik, sest see vähendab laborite erinevast täpsusest tulenevat müra.



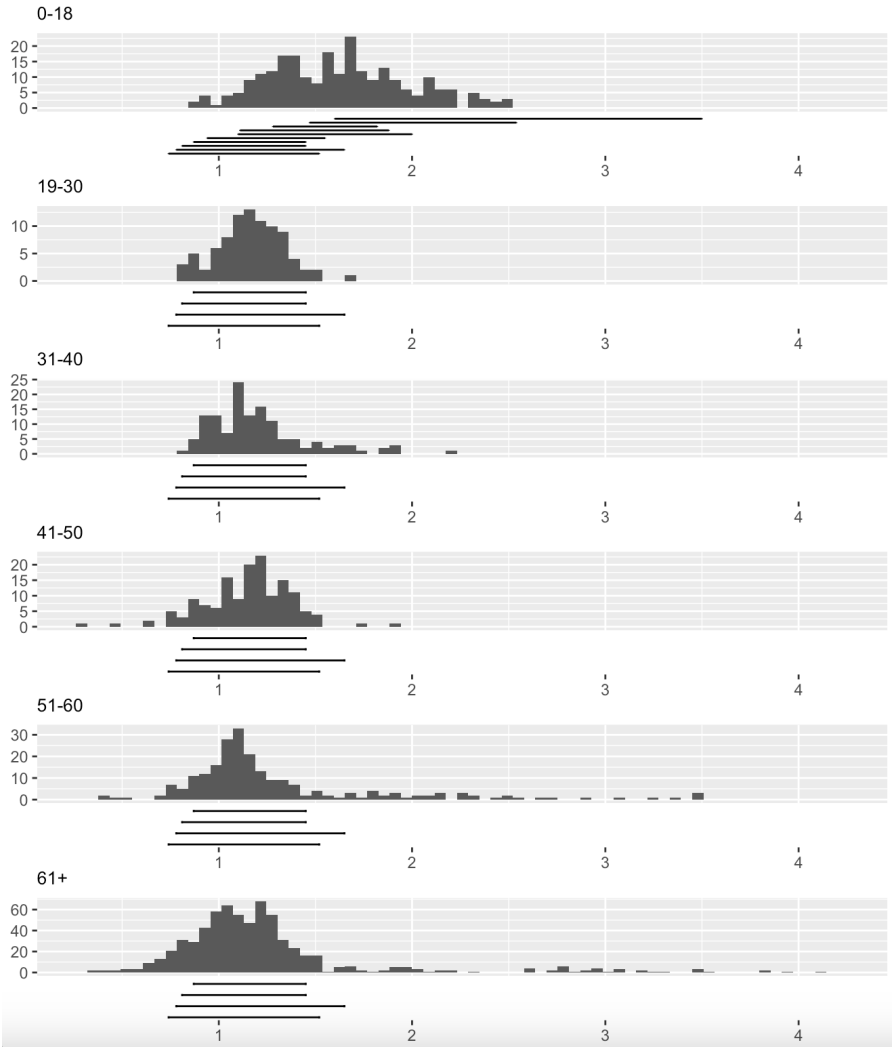
Joonis 14. Testosterooni referentsvahemikud grupeerituna soo järgi.

Et uurida, kas referentsvahemike keskmistel on erinevate vanusegruppide vahel statistiliselt oluline erinevus, kasutame Kruskal-Wallise testi, mis testib erinevust vähemalt kahe grupi vahel. Selleks jagame patsiendid vanuse järgi gruppidesse järgmiselt: 0-18; 19-30; 31-40; 41-50; 51-60; 61+.
Vaatame analüüte, mille puhul on andmestikus vähemalt kaks referentsvahemikku. Nendest 40 Kruskal-Wallise testi FDR korrektsiooniga p-väärtus on suurem kui 0,05 ehk vanusegruppide vahel ei ole statistiliselt olulist erinevust. 173 puhul on erinevus oluline. Tabelis 4 on erinevused Synlabi ning Kruskal-Wallise testi vanusejaotuste vahel.

Tabel 4. Referentsvahemike sõltuvus patsiendi vanusest.

		Synlab	
		Sõltub vanusest	**Ei sõltu vanusest**
Kruskal-Wallise test	**Sõltub vanusest**	57	39
	Ei sõltu vanusest	2	13

Statistiliselt oluliselt erinevad on näiteks eri vanusegruppide fosfaadi referentsvahemike keskpunktid (joonisel 15). Paistab, et siin erinevad oluliselt laste ning täiskasvanute referentsväärtused.

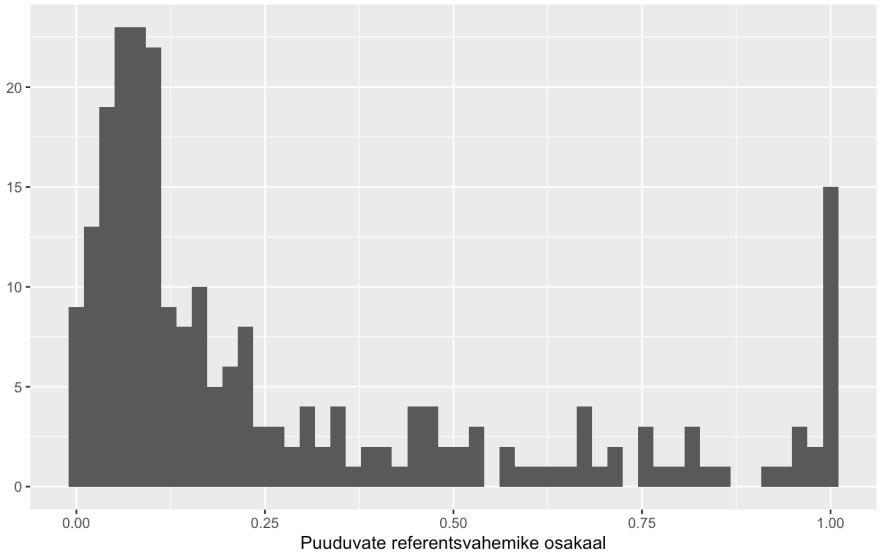


Joonis 15. Fosfaadi referentsvahemikud vanusegruppide kaupa.

Referentsvahemike grupeerimine soo ja vanuse järgi vähendab vahemike varieerivust. Vigaste vahemike tuvastamiseks, mida esineb rohkem kui 1% analüüdi kohta, tuleks vaadata referentsvahemike varieeruvust ühes soo- ja/või vanusegrupis. Kui seal esineb teistest märkimisväärselt erinevaid referentsvahemikke, on nende puhul alust oletada, et tegemist on veaga andmestikus.

4.2.2 Puuduvad referentsvahemikud.

Referentsvahemike puudumine on probleem, sest ilma vastava vahemikuta on analüüsitulemust keeruline interpreteerida. Töös kasutatavas andmestikus on referentsväärtused puudu $12{,}9\%$ mõõtmistulemustest. Joonisel 16 on histogramm puuduvate referentsvahemike osakaaludest. Probleemsed on analüüdid, millel on referentsväärtused puudu rohkem kui 10% analüüsitulemustest, selliseid analüüte on $59{,}8\%$, kusjuures 15 analüüdi puhul on vahemikud puudu kõigi mõõtmistulemuste juurest.



Joonis 16. Puuduvad referentsvahemikud.

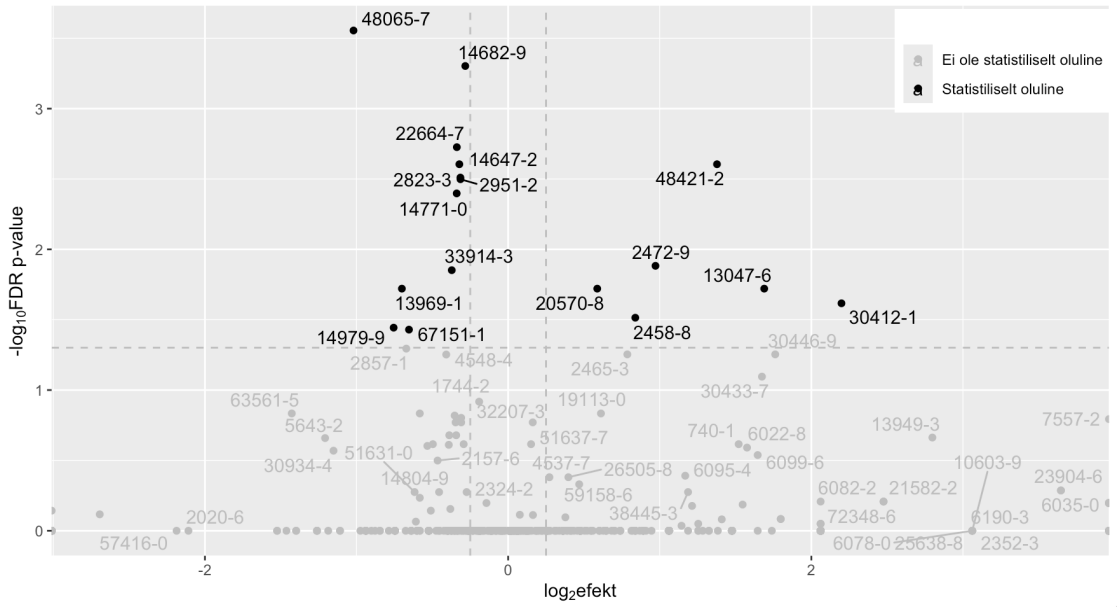
5 Diskretiseeritud tulemuste kasutamine

Selle asemel, et tegeleda erinevate terviseprobleemide tagajärgedega, on otstarbekas keskenduda nende ennetamisele. See on kasulik nii patsiendile kui tervisesüsteemile tervikuna, vähendades selle koormust ja toetades tervema elanikkonna loomist. Kui oleks võimalik prognoosida patsiendi tulevasi terviseprobleeme, saaks rakendada ennetavaid meetmeid, et terviseprobleemi tekkimine välistada või leevendada selle tõsidust. Prognoosimine annaks arstidele parema ülevaate patsiendi tervise seisust ning innustaks patsiente probleemi ennetamiseks tegutsema. Laborianalüüsid on üldiselt hea viis patsiendi tervislikust seisust ülevaate saamiseks ning üks võimalus tervisesündmuste prognoosimiseks on laborianalüüside tulemuste abil.
Terve inimese näitajad võivad mingil määral kõikuda, tervist hakkavad need mõjutama, kui jäävad alla või ületavad teatud piiri. Sellised piirid on määratud referentsväärtustega. Uurime, kas ja milline seos on analüüsitulemustel erinevate tervisesündmustega ning kas analüüside referentsväärtuste kasutamine lihtsustab ennustusmudelite loomist ja analüüsi.
Vaatame kolme tervisesündmuse esinemist. Esiteks uurime haigestumist ülemiste hingamisteede ägedatesse nakkustesse ehk haigustesse ICD-10 koodiga J06. Teiseks uurime antibiootikumiravi toimumist kasutades beetalaktaamantibiootikume ja penitsilliine ehk ravimeid ATC koodiga J01C. Kolmandaks uurime haiglaravile sattumist kauem kui nädalaks. Täpsemalt uurime kolme aasta laborianalüüside tulemusi ning nende seost sellega, kas neljandal aastal tervisesündmus toimub või ei toimu.
Laborianalüüse tehakse vahel kontrolliks tervetele inimestele (näiteks töötervishoiu tervisekontrollid), ent enamasti tehakse analüüse siis, kui esineb kahtlus mõne terviseprobleemi esinemiseks. Ainika Adamsoni magistritöö "*Assessment of the suitability of the Estonian Health Record data for the prediction of ischemic stroke*" tulemus oli, et insuldi ennustamisel on mõõtmisfakt ning analüüsitulemuse absoluutväärtus samaväärsed mudeli sisendid. See tähendab, et analüüsi läbiviimise fakt on isheemilise insuldi ennustamiseks piisavalt hea alus [17]. Seega uurime esmalt seoseid analüüsi läbiviimise ning tervisesündmuse toimumise vahel. Seejärel vaatame, kas kolme aasta jooksul läbi viidud analüüsi tulemusel on seos neljandal aastal haigestumise, antibiootikumiravi või haiglaravile sattumisega. Kui kolme aasta jooksul on tehtud ühte analüüsi mitu korda, võtame arvesse viimase tulemuse.
Analüüsitulemuste täpsete väärtuste asemel kasutame diskretiseeritud tulemusi ehk jagame tulemused kolme gruppi: normaalne, alla normaalse ning üle normaalse. Seejuures vaatame kahte diskretiseerimise viisi. Üks võimalus on võtta normaalseks analüüsitulemuseks keskmine tulemus ning sellest standardhälbe võrra madalamad ning kõrgemad tulemused. Kasutatavas andmestikus on aga nii tervete kui haigete inimeste tulemused ehk keskmine tulemus on nihkega. Lisaks võib analüüsitulemusele mõju olla peale tervisliku seisundi ka sool ning vanusel. Seega on teine ning ilmselt parem võimalus diskretiseerimiseks kasutades etteantud referentsvahemikke. Need on hinnatud tervete

inimeste baasil ning võtavad vajadusel arvesse sugu ja vanust.
Uurime seoseid mõõtmise ja tervisesündmuse toimumise vahel kasutades hii-ruut testi ning prognoosimiseks kasutame otsustusmetsa (*Random Forest*) ning logistilise regressiooni mudeleid. Seejuures ei ole eesmärk luua parima ennustusvõimega mudeleid, vaid eelkõige võrrelda ennustusvõime muutumist erinevate sisendite korral.

5.1 Mõõtmise toimumise seos tervisesündmustega

Mõõtmise toimumise ja tervisesündmuste vaheliste seoste selgitamiseks on joonistel 17-19 kasutatud efekt-olulisusgraafikuid, mille x-teljel on kahendlogaritm efektist ning y-teljel on negatiivne kümnendlogaritm hii-ruut testi p-väärtusest. Efekt, mis on võrdne ühega tähendab, et tervisesündmuse esinemise tõenäosusel ei ole vahet patsientidel, kellel on mõõtmine toimunud ning patsientidel, kellel mõõtmist toimunud ei ole, kahendlogaritmitud graafikul asub see kohal null. Efekt, mis on suurem kui üks (graafikul suurem kui null), tähendab, et selle mõõtmise toimumisel on suurem tõenäosus tervisesündmuse esinemiseks kui mõõtmise mittetoimumisel. Hii-ruut testi p-väärtuste juures on rakendatud FDR (*False Discovery Rate*) korrektsiooni, mis vähendab esimest liiki vea tegemise tõenäosust mitmese testimise puhul. Graafiku punktide juures on analüüdi LOINC kood ning halliga on kujundatud analüüdid, mille hii-ruut testi FDR korrektsiooniga p-väärtus on väiksem kui $0{,}05$ ehk mille seos tervisesündmuse toimumisega ei ole statistiliselt oluline.
Joonisel 17 on seosed analüüside toimumise ning ülemiste hingamisteede ägedatesse nakkustesse haigestumise vahel. Selgub, et 936-st analüüdist 17 puhul on seos statistiliselt oluline ning 392 analüüdi efekt on suurem kui üks. Kuue statistiliselt olulise seosega analüüdi efekt on ühest suurem ning 11 statistiliselt olulise analüüdi efekt on väiksem ühest.



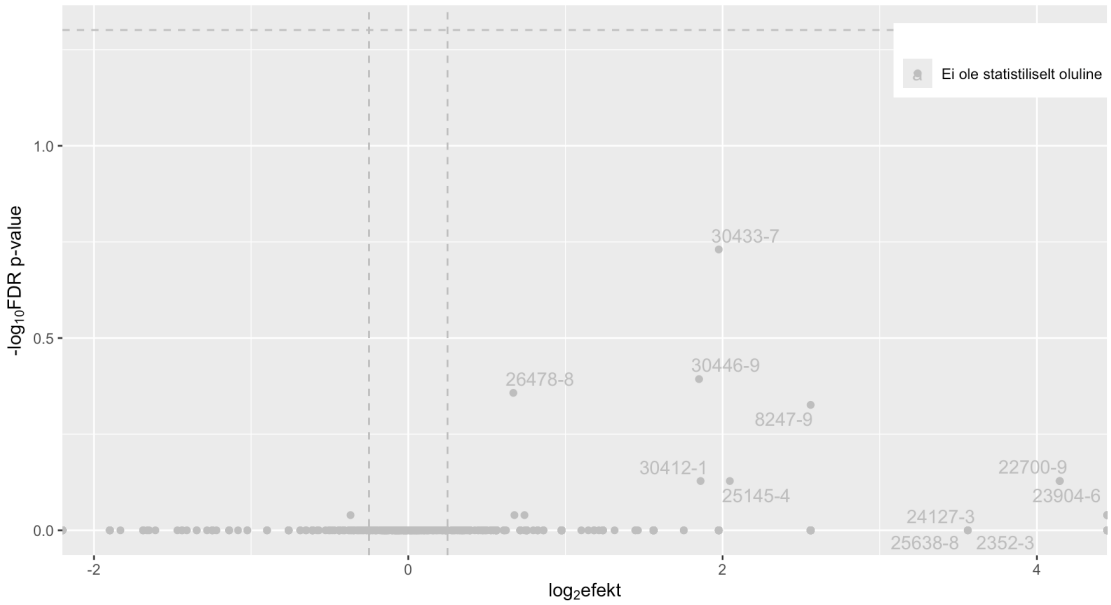
Joonis 17. Analüüsi toimumise ja haigestumise seos.

Tabelis 5 on suurima efektiga analüütide nimetused, mille seos haigestumisega on ühtlasi statistiliselt oluline.

Tabel 5. Analüüsi toimumise seos haigestumisega

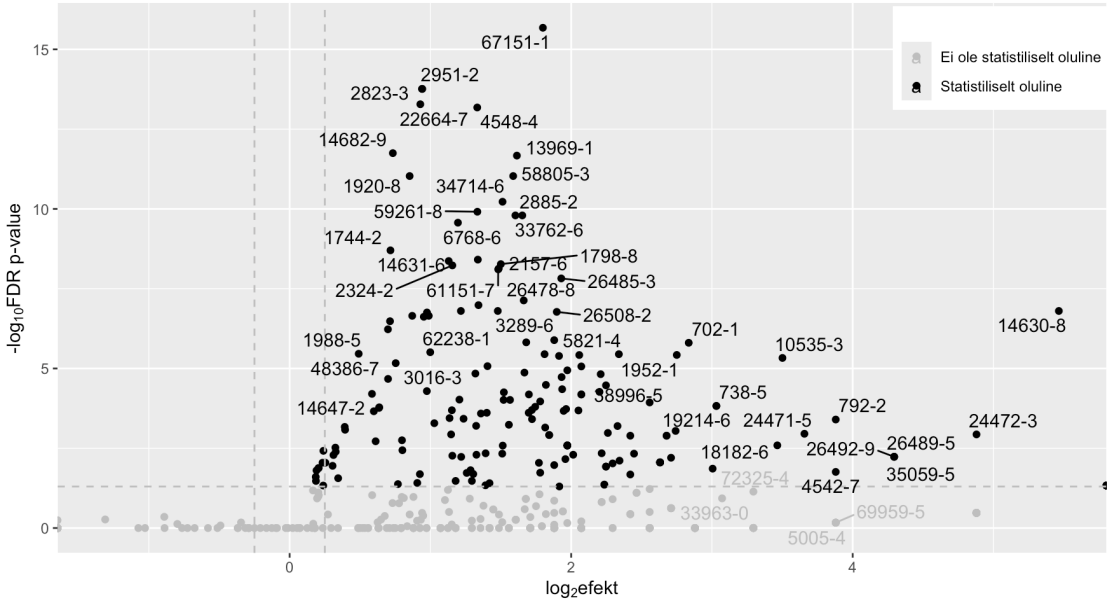
LOINC	**Analüüdi nimi**	**p-väärtus**	**Efekt**
30412-1	Atüüpilised lümfotsüüdid	0,024	4,59
13047-6	Plasmarakud %	0,019	3,22
48421-2	C-reaktiivne valk kapillaarses veres	0,002	2,60
2472-9	Immuunglobuliin M	0,013	1,96
2458-8	Immuunglobuliin A	0,031	1,79

Joonisel 18 on seosed analüüside toimumise ja antibiootikumiravi toimumise vahel. 354 analüüdi efekt 936-st on suurem kui üks, ent mitte ühegi analüüdi seos ei ole statistiliselt oluline. Põhjus võib olla selles, et analüüsitulemused ei ennusta põletikulisi protsesse tulevikus. Antibiootikumiravi vajaduse kindlakstegemiseks kasutatakse küll laboratoorseid teste, ent need tehakse vahetult enne ravikuuri ning seos mõõtmise ja kaugemas tulevikus tehtava antibiootikumiravi vahel puudub.



Joonis 18. Analüüsi toimumise ja antibiootikumiravi seos.

Analüüside toimumise ja pikemal kui nädalapikkusel haiglaravil viibimise vaheline seos on kujutatud joonisel 19. 936-st analüüdist 178 on haiglaraviga statistiliselt olulises seoses ning 385 analüüdi efekt on suurem kui üks. Seejuures on kõigi statistiliselt olulises seoses olevate analüütide efekt suurem kui üks.



Joonis 19. Analüüsi toimumise ja haiglaravile sattumise seos.

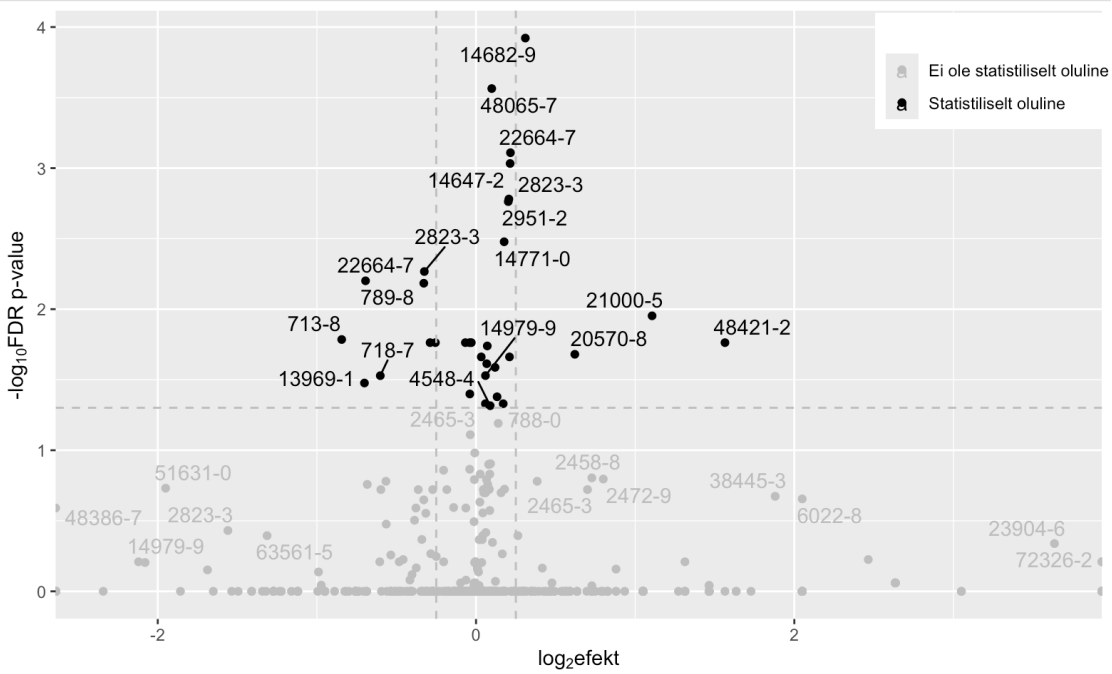
Tabelis 6 on suurima efektiga analüütide nimetused, mille seos haiglaravile sattumisega on ühtlasi statistiliselt oluline.

Tabel 6. Analüüsi toimumise seos haiglaravile sattumisega

LOINC	**Analüüdi nimi**	**p-väärtus**	**Efekt**
14630-8	Bilirubiin	$< 0{,}001$	44,15
24472-3	Trombotsüütide funktsiooni uuring kollageeni ja adenosiindifosfaadiga	0,001	29,44
26489-5	Mononukleaarsete leukotsüütide arv liikvoris	0,006	19,62
4542-7	Haptoglobiin	0,017	14,72
10535-3	Digoksiin	$< 0{,}001$	11,32

5.2 Analüüsitulemuse diskreetse väärtuse seos tervisesündmustega

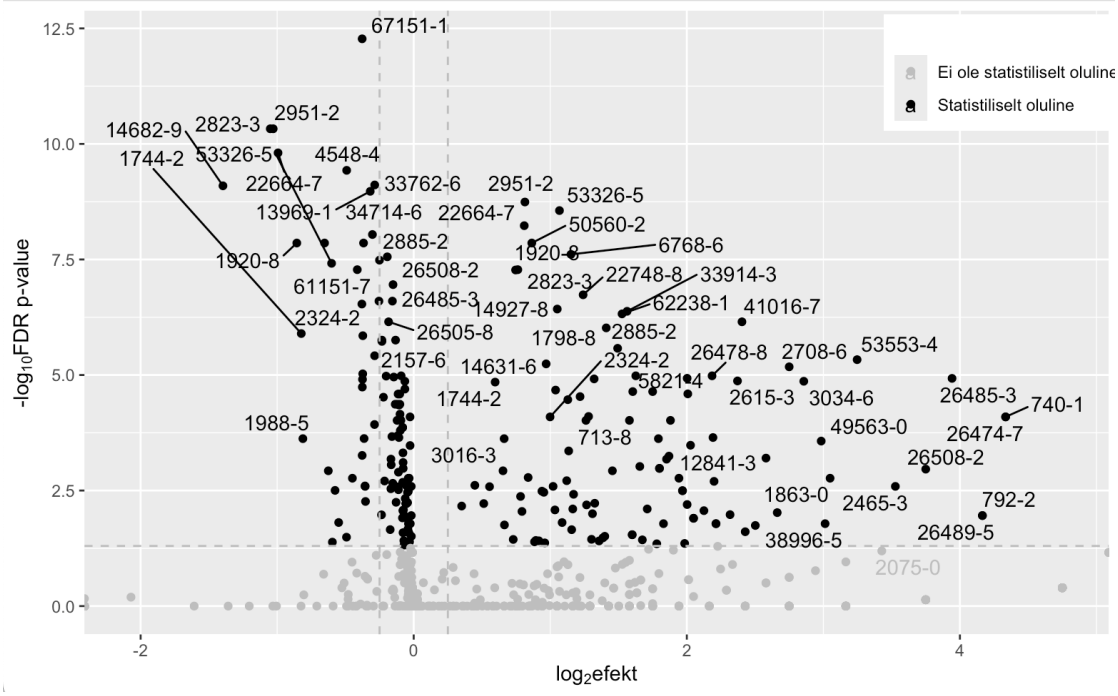
Joonistel 20 ja 21 on referentsvahemike järgi diskretiseeritud analüüsitulemuste ning tervisesündmuste vaheliste seoste efekt-olulisusgraafikud. Vaatame siinkohal vaid seoseid haiguse esinemise ning haiglaravile sattumisega, sest seosed antibiootikumiraviga on kõigi analüüsitulemuste puhul ebaolulised.



Joonis 20. Analüüsi diskretiseeritud väärtuste ja haiguse seos.

Tabel 7. Analüüsi diskretiseeritud väärtuste ja haiguse seos.

LOINC	**Analüüdi nimi**	**p-väärtus**	**Efekt**
48421-2	C-reaktiivne valk kapillaarses veres	0,017	2,96
21000-5	Erütrotsüütide suurusjaotuvus	0,011	2,15
20570-8	Hematokrit (erijuhtivus)	0,020	1,54



Joonis 21. Analüüsi diskretiseeritud väärtuste ja haiglaravile sattumise seos.

Tabel 8. Analüüsi diskretiseeritud väärtuste ja haiglaravile sattumise seos.

LOINC	**Analüüdi nimi**	**p-väärtus**	**Efekt**
26474-7	Lümfotsüütide arv (madal)	$< 0{,}001$	$20{,}18$
740-1	Metamüelotsüütide suhtarv	$< 0{,}001$	$2{,}18$
26489-5	Mononukleaarsete leukotsüütide arv liikvoris	$0{,}011$	$17{,}94$
26485-3	Monotsüütide suhtarv	$< 0{,}001$	$15{,}37$
26508-2	Kepptuumsete neutrofiilide suhtarv	$0{,}001$	$13{,}45$

5.3 Prognoosimine

Prognoosime binaarset tunnust: kas tervisesündmus toimub või ei toimu. Argumenttunnusteks on 742 analüüdi mõõteandmed. Ennustamiseks treeniti 8434 patsiendi andmete peal otsustusmetsa ning logistilise regressiooni mudelid. Testandmestikus olid lisaks 8422 patsiendi andmed ning mudelite headust hinnati AUC abil. Otsustusmetsa treenimiseks kasutati R paketti *randomForest* ning logistilise regressiooni jaoks paketti *stats*.
Hinnati kolm gruppi mudeleid, mille puhul kasutati erinevaid argumenttunnuseid. Esimese grupi puhul on argumenttunusteks iga analüüdi kohta, kas seda on mõõdetud või ei ole. Teise grupi argumenttunnused on keskväärtuse ja standardhälbe abil diskretiseeritud mõõtetulemused ning kolmanda grupi argumenttunnused on etteantud referentsvahemike abil diskretiseeritud mõõtetulemused. Tabelites 9-11 on mudelite AUCd.

Tabel 9. Mõõtmise toimumise AUC

	Random Forest (train/test)	**Logistiline regressioon (train/test)**
Haigus	0,552 / 0,563	0,624 / 0,528
Antibiootikumid	0,524 / 0,532	0,594 / 0,515
Haiglaravi	0,675 / 0,704	0,650 / 0,530

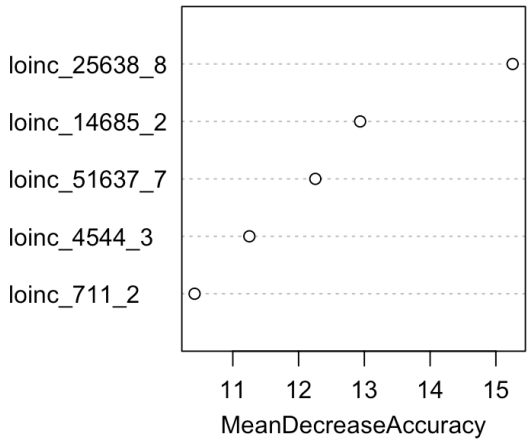
Tabel 10. Keskmise ja standardhälbe AUC

	Random Forest (train/test)	**Logistiline regressioon (train/test)**
Haigus	0,555 / 0,559	0,636 / 0,504
Antibiootikumid	0,531 / 0,508	0,502 / 0,507
Haiglaravi	0,683 / 0,691	0,546 / 0,500

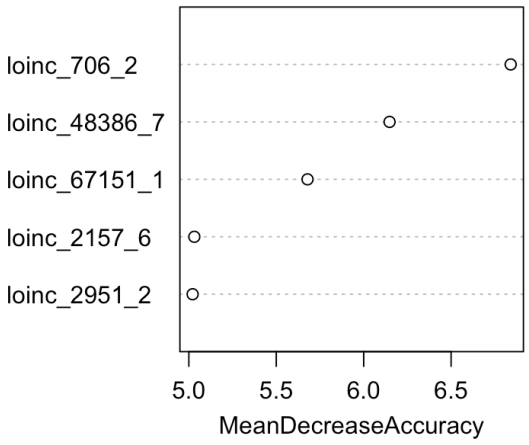
Tabel 11. Referentsvahemike AUC

	Random Forest (train/test)	**Logistiline regressioon (train/test)**
Haigus	0,564 / 0,566	0,620 / 0,503
Antibiootikumid	0,519 / 0,509	0,656 / 0,502
Haiglaravi	0,678 / 0,637	0,761 / 0,500

Koostatud mudelite prognoosivõime on madal, sest testandmetel hinnatud AUCd jäävad $0{,}5$ lähedale, mis tähendab, et mudelid ei ole palju paremad juhuslikust mudelist. Kolmest tervisesündmusest kõige paremini prognoositav on pikemale kui nädalapikkusele haiglaravile sattumine. Sellele viitas juba fakt, et suure osa analüütide mõõtmise toimumine oli haiglaraviga statistiliselt olulises seoses.
Üllatav on asjaolu, et keskmise ja standardhälbe järgi diskretiseeritud ning referentsvahemike järgi diskretiseeritud mudelitel ei ole olulist vahet. Võiks arvata, et referentsvahemike järgi diskretiseerimine annab paremad tulemused, ent kohati oli keskmise ja standardhälbe mudel isegi parem. Selle tulemuse selgitamiseks on joonistel 22 ning 23 vastavalt haigust ja haiglaravi prognoosivate otsustusmetsa mudelite olulisima ennustusmõjuga tunnused ehk analüüdid. Selgub, et haigust prognoosiva mudeli kõige olulisemad analüüdid on eosinofiilsete granulotsüütide katioonne proteiin, vitamiin B12, trombokrit, hematokrit ning eosinofiilide arv. Haiglaravi prognoosiva mudeli olulisimad analüüdid on basofiilide suhtarv, trombotsüütide suurte vormide suhtarv, troponiin T, kreatiini kinaas ja naatrium. Suuremal osal neist analüütidest on küll vanusest ja soost sõltuvad referentsvahemikud, aga seda ainult lapsi ning täiskasvanuid eristavalt ehk kõigi täiskasvanute referentsvahemikud on samad. Seejuures on andmestikes esinevatest mõõtmistest 89% tehtud täiskasvanutel ehk enamiku mõõtmiste puhul on oluliste analüütide referentsvahemikud võrdsed. Seega annab referentsvahemike abil diskretiseerimine suuresti samad tulemused, mis keskväärtuse ja standardhälbe abil diskretiseerimine ning see seletab, miks mudelite prognoosivõimel ei ole olulist vahet.

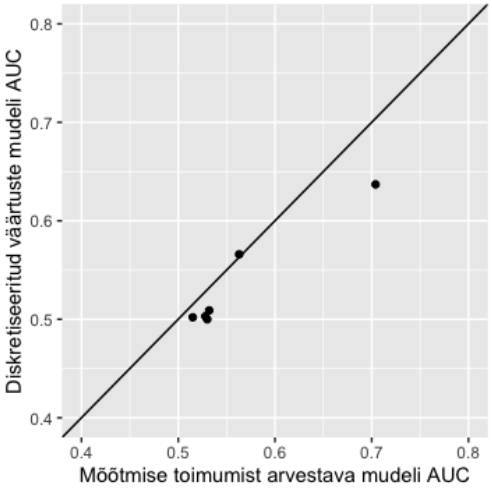


Joonis 22. Olulisima ennustusvõimega tunnused haiguse prognoosimisel.



Joonis 23. Olulisima ennustusvõimega tunnused haiglaravi prognoosimisel.

Parima mudeli AUC testandmetel on ligikaudu $0{,}7$. Selleks, et saaksime väita, et referentsvahemike abil diskretiseeritud andmetel põhinev mudel on parem mõõtmisfaktidel põhinevast mudelist, võiks diskretiseeritud andmete mudeli AUC olla $0{,}7$ lähedal ning mõõtmisfaktide mudeli AUC juhusliku mudeli ehk $0{,}5$ lähedal. Jooniselt 24 selgub, et see ei vasta tõele. Mudelite testandmetel hinnatud AUCd on väga sarnased ning enamasti on hoopis mõõtmisfaktidel põhineva mudeli AUC suurem. Seega võib järeldada, et analüüsi diskreetsest väärtusest olulisem näitaja tervisesündmuse prognoosimisel on see, et mõõtmine üldse läbi viidi.



Joonis 24. AUC võrdlus

6 Kokkuvõte

Magistritöö üks eesmärk oli tuvastada vigu laborianalüüside tulemuste andmestikus ehk analüüsitulemusi, mis on seotud kas vale LOINC koodiga või vale mõõtühikuga ning kontrollida referentsvahemike õigsust. Vigade tuvastamiseks analüüsitulemuste hulgas kasutati klasterdamist Gaussi segumudeli abil. Normaaljaotusega tulemuste puhul oli klasterdamine edukas - mudel jagas analüüsitulemused mitmesse klastrisse, kui kasutatud oli näiteks erinevaid mõõtühikuid. Seega sai tekkinud klastrite arvu järgi tuvastada probleemsed analüüdid. Selgus aga, et suur osa analüüsitulemusi on normaaljaotuse asemel eksponentjaotusega ning nende puhul Gaussi segumudel kasulikke klasterdustulemusi ei andnud. Tulevikus peaks selliste analüütide tulemuste klasterdamiseks seega kasutama segumudelit, mille komponendid on eksponentjaotusega.
Referentsvahemike õigsuse kontrolliks uuriti erinevate vahemike arvu ja varieeruvust analüüdi kohta. Referentsvahemike keskpunktide võrdlemisel erinevate soo- ja vanusegruppide vahel selgus, et suurema osa analüütide puhul esinevad olulised erinevused. Seega on üks referentsvahemike rohkuse ja varieeruvuse põhjus soost ning vanusest sõltumine. Lisaks selgus, et vahemike keskpunktide võrdlemine on kasulik viis referentsvahemike varieeruvuse uurimiseks, sest see vähendab laborite erinevusest tingitud müra. Vigade esinemise kontrolliks tuleks edaspidi vaadelda referentsvahemike varieeruvust soo- ja vanusegruppide kaupa. Kui varieeruvus on ka soo ja vanuse mõju eemaldamisel suur, siis on tegemist ilmselt vigade esinemisega andmestikus.
Töö teine eesmärk oli uurida, kas disketiseeritud analüüsitulemustest on kasu tervisesündmuste ennustamisel ning kas erinevad diskretiseerimise viisid viivad erinevate tulemusteni. Selleks võrreldi mudeleid, mis kasutasid erinevate tervisesündmuste esinemise ennustamiseks argumenttunnustena mõõtmisfakte, keskmise ja standardhälbe abil diskretiseeritud mõõtmistulemusi ning referentsväärtuste abil diskretiseeritud tulemusi. Selgus, et keskmise ja standardhälbe abil diskretiseeritud tulemuste ja referentsväärtuste abil diskretiseeritud tulemuste mudelite ennustusvõimetel ei ole olulist vahet. Lisaks selgus, et mudelid, mis kasutasid argumenttunnustena mõõtmisfakte, on sama hea ennustusvõimega nagu mudelid, mis kasutasid diskretiseeritud mõõtetulemusi ehk analüüsi diskreetsest väärtusest olulisem näitaja tervisesündmuse prognoosimisel on see, et mõõtmine läbi viidi.

Viidatud kirjandus

[1] R. W. Forsman. "Why is the laboratory an afterthought for managed care organizations?" *Clinical Chemistry* 42.5 (mai 1996), lk. 813–816. ISSN: 1530-8561. DOI: 10.1093/clinchem/42.5.813.

[2] F. Ceriotti ja J. Henny. ""Are my laboratory results normal?" Considerations to be made concerning reference intervals and decision limits". *Ejifcc* 19.2 (2008), lk. 106.

[3] Eesti Laborimeditsiini Ühing. *LOINC töörühm*. https://www.elmy.ee/tooruhmad/loinc/. (13.05.2024).

[4] *About LOINC*. https://loinc.org/about/. (13.05.2024).

[5] *LOINC Term Basics*. https://loinc.org/get-started/loinc-term-basics/. (13.05.2024).

[6] N. Slhessarenko ja A. Andriolo. "The importance of determining reference intervals for Laboratory Medicine". *Jornal Brasileiro de Patologia e Medicina Laboratorial* (2016). ISSN: 1676-2444. DOI: 10.5935/1676-2444.20160019.

[7] Tervise Arengu Instituut. *RHK ehk rahvusvaheline haiguste klassifikatsioon*. https://www.tai.ee/et/instituudist/meditsiiniterminoloogia-kompetentsikeskus/who-klassifikaatorid/rhk-ehk-rahvusvaheline. (13.05.2024).

[8] World Health Organization. *Anatomical Therapeutic Chemical (ATC) Classification*. https://www.who.int/tools/atc-ddd-toolkit/atc-classification. (13.05.2024).

[9] M. Oja *et al.* "Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned". *JAMIA Open* 6.4 (2023). DOI: 10.1093/jamiaopen/ooad100.

[10] L. Scrucca *et al.* "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models". *The R Journal* 8.1 (2016), lk. 289. ISSN: 2073-4859. DOI: 10.32614/rj-2016-021.

[11] A. Ugrjumova. "Mudelipõhise klasteranalüüsi ja K-medoidide meetodi võrdlemine kvalitatiivsete tunnustega andmete klasterdamisel". Magistritöö. Tartu Ülikool, 2020.

[12] M. Schonlau ja R. Y. Zou. "The random forest algorithm for statistical learning". *The Stata Journal: Promoting communications on statistics and Stata* 20.1 (märts 2020), lk. 3–29. ISSN: 1536-8734. DOI: 10.1177/1536867x20909688.

[13] S. Sarkar ja H. Midi. “Importance of Assessing the Model Adequacy of Binary Logistic Regression”. *Journal of Applied Sciences* 10.6 (märts 2010), lk. 479–486. ISSN: 1812-5654. DOI: 10.3923/jas.2010.479.486.

[14] Ş. Corbacioglu ja G. Aksel. “Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value”. *Turkish Journal of Emergency Medicine* 23.4 (2023), lk. 195. ISSN: 2452-2473. DOI: 10.4103/tjem.tjem_182_23.

[15] TensorFlow. *MixtureSameFamily*. https://www.tensorflow.org/probability/api_docs/python/tfp/distributions/MixtureSameFamily. (13.05.2024).

[16] SYNLAB. *Referentsväärtuste tabel*. https://synlab.ee/arstile/laboriteatmik/referentsvaartused/. (13.05.2024).

[17] A. Adamson. “Assessment of the suitability of the Estonian Health Record data for the prediction of ischemic stroke”. Magistritöö. Tartu Ülikool, 2021.

Lisad

I. GitLabi repositoorium

`https://gitlab.cs.ut.ee/health-informatics/student-theses/annika_talvet_lab_analyses` (privaatne repositoorium)

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Annika Talvet**,
*(*autori nimi)

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

   Laborianalüüside diskretiseerimine ja analüüs ,
   *(*lõputöö pealkiri)

   mille juhendaja on Sven Laur,
   *(*juhendaja nimi)

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

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

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

Annika Talvet
15.05.2024