



Perearstiabi patsientide visiitide arvu ja haiguskestuse prognoosid Eesti Haigekassa ning tervise infosüsteemi andmete põhjal



Tervise Arengu Instituut
National Institute for Health Development

Tervise Arengu Instituut

Perearstiabi patsientide visiitide arvu ja haiguskestuse prognoosid Eesti Haigekassa ning tervise infosüsteemi andmete põhjal

Kirill Smirnov, Diana Sokurova

Tallinn 2019

Tervise Arengu Instituudi **missioon** on luua ja jagada teadmisi, et tõenduspõhiselt mõjutada tervist toetavaid hoiakuid, käitumist, poliitikat ja keskkonda eesmärgiga suurendada inimeste heaolu Eestis.

Väljaande andmete kasutamisel viidata allikale.

Soovitav viide käesolevale väljaandele: Smirnov K, Sokurova D. Perearstiabi patsientide visiitide arvu ja haiguskestuse prognoosid Eesti Haigekassa ning tervise infosüsteemi andmete põhjal. Tallinn: Tervise Arengu Instituut; 2019.

Sisukord

Lühendid.....	4
Mõisted.....	4
Kokkuvõte.....	5
Sissejuhatus.....	7
1 Metoodika	8
1.1 Loendusandmete mudelid	8
1.2 Logistilised mudelid	8
1.3 Elukestusmudelid	9
1.4 Mudelite diagnostika.....	9
2 Mudelite tegemine Eesti Haigekassa andmestiku põhjal.....	10
2.1 Eesti Haigekassa andmestiku kirjeldus.....	10
2.2 Saadud loendusandmete mudeli kirjeldus	11
2.3 Saadud logistilise mudeli kirjeldus.....	16
3 Mudelite testimine tervise infosüsteemi andmestiku abil	19
3.1 Tervise infosüsteemi andmestiku kirjeldus.....	19
3.2 Loendusandmete mudeli testimine	20
3.3 Logistilise mudeli testimine.....	22
4 Haiguskestuse prognoosimine tervise infosüsteemi andmestikus.....	23
Kasutatud kirjandus	28
LISA.....	29

Lühendid

EHK	Eesti Haigekassa
RHK-10	rahvusvaheline haiguste ja tervisega seotud probleemide statistiline klassifikatsioon, 10. versioon
TIS	tervise infosüsteem

Mõisted

Epikriis	kokkuvõte patsiendi haigusjuhtumist, milles kajastatakse antud haigusjuhtumi dünaamika, lähtudes arsti käsutuses olevast sellekohasest teabest. Ambulatoorne epikriis koostatakse iga ambulatoorse haigusjuhtumi kohta selle lõppemisel
Raviarve	ravikulude arve, mille tervishoiuteenuse osutaja esitab patsiendile osutatud teenuste eest Haigekassale. Ühel raviarvel võib olla info mitme visiidi kohta
Tõenäosus	arv, mis iseloomustab juhusliku sündmuse võimalikkust. Tõenäosus võrdub ühega, kui sündmus kindlasti toimub ja võrdub 0-ga, kui sündmuse toimumine on võimatu
Visiit	perearsti või -õe vastuvõtud (esmane, korduv, profülaktiline) ning telefoni ja e-posti teel toimunud nõustamised

Kokkuvõte

Analüüsis prognoositi statistiliste mudelite abil patsientide visiitide arvu perearstiasse ja tõenäosust, et patsient pöördub ühe ja sama diagnoosiga aasta jooksul vähemalt kaks korda ning samuti, kui pikk on patsiendi haiguskestus.

Saadud tulemused selgitasid, millest sõltub visiitide sagedus ja milliste patsientidega tuleb perearstiasse kõige enam tegeleda. Antud informatsioon aitab hinnata ja paremini planeerida perearstiabi tööjaotust teades patsienti, tema diagnoosi ja visiidi toimumist. Töös kasutati Eesti Haigekassa (EHK) ja tervise infosüsteemi (TIS) andmeid.

Nii EHK kui ka TIS andmestikus leiti vaatlused puuduvate andmetega (puudub patsiendi sugu, vanus või diagnoos). Tähele tuleb panna, et TIS andmestikus on puuduvate andmetega epikriiside osatähtsus 6 korda suurem kui EHK andmestikus. 2017. aastal oli TIS andmestikus kirjeldatud ainult 720 362 unikaalset patsienti, EHK andmestikus oli patsientide arv 1 011 794. Seega annab EHK andmestik meile rohkem informatsiooni ja selle kvaliteet on parem.

Loendusandmete mudelit kasutatakse juhul, kui on vaja prognoosida sündmuse toimumise arvu teatud ajavahemikus. Antud analüüsis on see patsiendi pöördumiskordade ehk visiitide arv perearstiabi kasutamisel 2017. aastal. EHK andmete põhjal saadud mudelist selgus, et aasta jooksul tehtud visiitide arv sõltub patsiendi vanuserühmast, diagnoosirühmast, diagnoosi tüübist (esialgne, esmane, korduv), esimese visiidi tüübist (perearsti/-õe vastuvõtt, nõustamine telefoni või e-posti teel) ning esimese visiidi kuupäevast, kuid ei sõltu patsiendi soost.

Mudeli järgi pöördub perearstiasse kõige sagedamini patsient vanuses 0–4, kelle esmane diagnoos on rühmast S00–T98 (vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed) ja tema esimeseks visiidiks on pereõe nõustamine telefoni teel, mis toimub 1. jaanuaril. Keskmine visiitide arv aastas antud tingimustel on viis.

Kõige harvem pöördub patsient perearstiasse vanuses 15–24, kelle esialgne diagnoos on rühmast H00–H59 (silma- ja silmamanuste haigused) ja tema esimeseks visiidiks on perearsti vastuvõtt aasta viimasel päeval (31. detsember). Keskmine visiitide arv aastas on 0,34.

Prognoositud visiitide arv erineb reaalsest visiitide arvust 63% juhtudest ühe visiidi võrra EHK testandmestikus ja 73,5% juhtudest TIS andmestikus. Ülejäänud juhtudel on prognoosid suurema veaga, vastavalt 37% EHK ja 26,5% TIS andmestikust.

Mudeli prognoos on täpne juhul, kui patsiendi reaalne visiitide arv aasta jooksul on üks, kaks või kolm. Näiteks, kui patsient tegi aasta jooksul kaks visiiti, siis mudeli prognoos on keskmiselt 1,9 visiiti. Küll aga ei saa mudeliga hinnata neid patsiente, kes on teinud palju visiite. Näiteks, kui patsient tegi aastas realselt kuus visiiti on tema keskmine prognoositud visiitide arv 2,3.

Logistilisi mudeleid kasutatakse juhul, kui prognoositud tunnus on binaarne ehk tunnusel on ainult kaks võimalikku väärtust (0 või 1). Antud töö raames prognoositi logistilise mudeli abil, kas patsient pöördus perearstiasse aastas rohkem kui ühe korra või mitte. EHK andmete põhjal saadud mudelist selgus, et tõenäosus, kus patsient teeb aasta jooksul vähemalt kaks visiiti, sõltub soost, vanuserühmast, diagnoosirühmast, diagnoosi tüübist (esialgne, esmane, korduv), esimese visiidi tüübist (perearsti/-õe vastuvõtt, nõustamine telefoni või e-posti teel) ja esimese visiidi kuupäevast.

Saadud mudeli järgi on suurem šanss pöörduda perearstiasse rohkem kui üks kord aastas poisil vanuses 0–4, kelle esmane diagnoos on rühmast S00–T98 (vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed) ja tema esimene visiit on pereõe nõustamine telefoni teel 1. jaanuaril. Tõenäosus, et ta tuleb ka teist korda, on 0,93.

Väikseim šanss pöörduda perearstiasse rohkem kui üks kord aastas on naisel vanuses 15–24, kellel on korduv diagnoos rühmast H00–H59 (silma- ja silmamanuste haigused) ja tema esimene

visiit on perearsti vastuvõtt 31. detsembril. Tõenäosus, et ta pöördub perearstiabisse vähemalt kaks korda, on sel juhul 0,13.

Saadud mudel prognoosib õigesti, kas patsient tuleb perearsti juurde ainult üks kord või teeb vähemalt kaks visiiti aasta jooksul 64% juhtudest EHK andmestiku ja 52% juhtudest TIS andmestiku põhjal.

TIS andmestiku põhjal tehtud mudeli järgi sõltub haiguskestus soost, vanuserühmast, diagnoosirühmast, diagnoosi tüübist (esialgne, esmane, korduv) ning esimese visiidi tüübist (perearsti/-õe vastuvõtt, nõustamine telefoni teel).

Saadud mudeli järgi on haiguskestus kõige lühem mehel vanuses 45–64, kellel on korduv diagnoos rühmast Z00–Z99 (tervise seisundit mõjustavad tegurid ja kontaktid tervise teenistusega) ja tema esimene kontakt oli perearsti nõustamine telefoni teel. Haiguse prognoositud pikkus on 1,84 päeva.

Kõige pikkema haiguskestusega on naine vanuses 65–84, kellel on esmane diagnoos rühmast S00–T98 ja tema esimene visiit on pereõe nõustamine telefoni teel. Antud juhul kestab haigus 24,96 päeva.

TIS andmestiku põhjal on haiguskestuse mudeli prognoosi vea varieeruvus +/- 5 päeva 54% juhtudest. Ülejäänud 46% juhtudest on prognoosi vea varieeruvus veel suurem.

Sissejuhatus

Käesoleva töö eesmärk on mudelite koostamine prognoosimaks, mitu korda aastas pöördub patsient ühe ja sama diagnoosiga perearstiabisse ning milline on tõenäosus, et patsient pöördub sama diagnoosiga aasta jooksul korduvalt. Prognoosimudeli koostamiseks kasutatakse kaht mudeliklassi: loendusandmete ja logistilist mudelit. Mudelite koostamiseks kasutati 2017. aasta Eesti Haigekassa (EHK) ravikindlustuse andmekogu perearstiabi raviarvete andmeid.

Mudelite sobivust kontrolliti testandmestikuga, milleks plaaniti kasutada tervise infosüsteemi (TIS) edastatud ambulatoorsete epikriiside andmeid sama perioodi kohta. Sobivuse tulemusel oleks võimalik teada, kui hästi kirjeldab EHK andmetest tehtud mudel TIS andmeid. Selgus, et antud mudeli viisil ei saa võrrelda EHK ja TIS andmeid, kuna mudel kirjeldab liiga hästi TIS andmestikku. Selleks võib olla kaks põhjust:

1. mudel on hea ja sobib nii EHK kui TIS andmetele,
2. nii TIS kui ka EHK andmed kirjeldavad samu inimesi.

Seega töös kasutati sobivuse kontrollimiseks EHK testandmestiku, kuid lisaks vaadeldi, kui hästi kirjeldab mudel ka TIS andmeid. Kui õigete prognooside osatähtsus tuleb väike, siis see näitab, et mudel on halb või andmed ei ole võrreldavad.

Tervishoiuteenuse osutajad edastavad EHK-le raviarveid. TIS kogub tervishoiuteenuse osutajatelt andmeid lõpetatud epikriiside kohta. Kuna käesoleva töö rõhk on suunatud aasta jooksul perearstiabisse tehtud visiitide arvule ühe ja sama diagnoosiga, mitte aga visiitide arvule ühe epikriisi või raviarve raames, siis seepärast otsustati grupeerida andmeid patsiendi ja diagnoosi järgi. Grupeeritud andmestikke kasutati prognoosimudeli koostamiseks.

Lisaks koostati elukestusmudel haiguskestuse prognoosimiseks. Elukestusmudel tehti TIS andmete põhjal, kuna EHK andmetes puudub informatsioon haiguse lõpust. Antud andmete puhul loeti haigestumiseks ka retseptiravimite väljakirjutamine ja arstlik läbivaatus.

Käesolev töö on jätk Tervise Arengu Instituudi 2018. aasta analüüsile „Perearstiabisse pöördumise põhjused, 2017“ (<https://www.tai.ee/et/terviseandmed/uuringud/download/468>).

1 Metoodika

1.1 Loendusandmete mudelid

Loendusandmete mudelit kasutatakse juhul, kui on vaja prognoosida sündmuse toimumise arvu teatud ajavahemikus. Käesoleva töö raames on see patsiendi visiitide arv perearstiabisse 2017. aastal. Olemas on mitu erinevat loendusandmete mudelit. Kõige enam kasutatakse neist Poissoni ning Negative Binomial (NB-mudel) mudelit.

Kuna Poissoni jaotuse korral peab kehtima vastavus $D(Y) = E(Y)$, on Poissoni mudelit mõistlik kasutada tunnuse Y prognoosimiseks ainult siis, kui andmestikus puudub üle- või alahajuvus ($D(Y) > E(Y)$ ülehajuvuse korral ning $D(Y) < E(Y)$ alahajuvuse korral). Õeldakse, et alahajuvus ei tekita probleeme mudeli parameetrite hindamisel ja seda on vähem uuritud, aga suur ülehajuvus võib tekitada vigu. (1) Kui ülehajuvus on suur ($D(Y) - E(Y) > 5$), siis Poissoni mudeli asemel on NB-mudel eelistatum. (3)

Peale tavaliste Poissoni ning NB-mudelite eksisteerivad ka ZTP (*zero truncated Poisson*) ning ZTNB (*zero truncated NB*). Neid mudeleid kasutatakse juhul, kui andmestikus puuduvad nullid ehk vaadeldud ajavahemikus prognoositud sündmus toimub vähemalt üks kord.

Poissoni ning NB-mudeli korral arvutatakse prognoositud tunnus y järgmise valemi järgi:

$$y = \exp(x^T B),$$

kus x – seletavate tunnuste väärtuste vektor, B – mudeli koefitsientide vektor.

1.2 Logistilised mudelid

Logistilisi mudeleid kasutatakse juhul, kui prognoositud tunnus on binaarne ehk tunnusel on ainult kaks võimalikku väärtust (0 või 1). Näiteks, kas patsient pöördus perearstiabisse rohkem kui ühe korra või mitte, on binaarne.

Logistilise mudeli prognoosimiseks on võimalik kasutada *Logit*, *Probit* ning *Clog-Log* seosefunktsioonid.

Olgu π tõenäosus, et prognoositud tunnus on võrdne ühega, siis

$$\text{Logit mudeli korral } \pi = \frac{\exp(x^T B)}{1 + \exp(x^T B)},$$

$$\text{Probit mudeli korral } \pi = \Phi(x^T B);$$

$$\text{Clog-log mudeli korral } \pi = 1 - \exp(-\exp(x^T B)),$$

kus x – seletavate tunnuste väärtuste vektor, B – mudeli koefitsientide vektor ning Φ – standardse normaaljaotuse tihedusfunktsioon.

Logistilise mudeli interpreteerimiseks on tihti kasutusel šanss. Šansiks nimetatakse sündmuse esinemise tõenäosuse ja sündmuse mitteesinemise tõenäosusesuhet $\Pi = \frac{\pi}{1-\pi}$. Logit mudeli korral $\Pi = \exp(x^T B)$.

Mudeli sobivuse hindamiseks on võimalik kasutada *Hosmer-Lemeshow*i testi. Kui saadud testi p -väärtus on $< 0,05$, siis saadud mudel ei sobi hästi andmete kirjeldamiseks.

1.3 Elukestusmodelid

Elukestusmodelid kasutatakse juhul, kui on vaja prognoosida sündmuse toimumise aega. Näiteks, käesolevas töös prognoositakse aega päevades haiguse algusest kuni hetkeni, mil patsient saab terveks.

Oluline mõiste elukestusmodelite kasutamisel on *hazard* funktsioon. *Hazard* funktsiooni saab interpreteerida kui tõenäosust, et patsient saab terveks perioodil $(t, t + \Delta t)$, kui on teda, et ta on haige ajal t ning $\Delta t \rightarrow 0$. (2)

Eksisteerib mitu erinevat elukestusmodelit. Kõige populaarsemad nendest on Weibull, Log-normal ning Log-logistic mudel. Nende mudelite kasutamiseks eeldatakse, et uuritud tunnus on pärit vastavast jaotusest.

Alternatiivne variant on mittepameetiline Cox mudel. Selle mudeli kasutamine ei eelda mingit konkreetset jaotust, aga ta vajab erinevate vaatluste jaoks *hazard* funktsioonide proportsionaalsust. (2) Selleks, et seda kontrollida, kasutatakse *Grambsch-Therneau* testi. Kui testi p -väärtus on $< 0,05$, siis saadud Cox mudel ei sobi andmete kirjeldamiseks.

1.4 Mudelite diagnostika

Mudelite võrdlemiseks kasutatakse *Akaike Information Coefficient* (AIC), mis arvutatakse järgmise valemi järgi:

$$AIC = 2k - 2\ln(L),$$

kus k - parameetrite arv mudelis, L - tõepärafunktsiooni maksimaalne väärtus. Tõepärafunktsiooniks nimetatakse funktsiooni, mis esitab valimi saamise tõenäosuse sõltuvalt jaotusparameetrite väärtusest.

Väiksema AIC-i väärtusega mudel kirjeldab andmeid paremini (3). Seega kui võrrelda omavahel kaht mudelit AIC väärtustega vastavalt AIC_1 ning AIC_2 nii, et $AIC_1 > AIC_2$ siis teine mudel (väiksema AIC-i väärtusega) on eelistatum kasutamiseks ja prognoosimiseks.

Prognoosi täpsuse uurimiseks kasutatakse töös RMSE (*Root mean square error*), MAPE (*Mean Absolute Percentage Error*) ning MAD (*Mean Absolute Difference*) kriteeriumeid.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

$$MAD = \frac{1}{n} \left| \sum_{i=1}^n \hat{y}_i - y_i \right|$$

kus $\hat{y}_i$ - prognoositud väärtus, y_i - reaalne väärtus.

2 Mudelite tegemine Eesti Haigekassa andmestiku põhjal

2.1 Eesti Haigekassa andmestiku kirjeldus

Eesti Haigekassa ravikindlustuse andmekogus on kõik ravikindlustatute raviarved ja ravikindlustamata patsientide üldarstiabi arved. Käesolev töö hõlmab neid raviarveid, mis algasid ning lõppesid 2017. aastal. Kokku on andmestikus 3 940 681 sobivat arvet. Igast raviarvest on teada:

- **Raviarve number**
Unikaalne raviarve kood.
- **Patsiendi ID**
Kaheksanumbiline kood, mis eristab patsienti. Andmestikus on olemas 1 011 794 unikaalset patsienti.
- **Sugu**
40% raviarvetest on kirjutatud meespatsientidele ja 60% naispatsientidele. 10-nel raviarvel oli patsiendi sugu teadmata.
- **Vanus**
Patsiendi vanus arve kirjutamise ajal. Isikute vanus varieerub andmestikus 0-st kuni 107-aastani. Neljal raviarvel oli vanus teadmata.
- **Vanuserühm**
Patsiendi vanuserühm raviarve kirjutamise ajal. Käesoleva töö raames kasutatakse seitset vanuserühma (0–4; 5–14; ..., 85+). Kõige rohkem raviarveid on välja antud vanuserühma 65–84 patsientidele (29,01%). Väiksemad rühmad on 15–24 (5,54%) ning 85+ (5,52%).
- **Põhidiagnoos**
Patsiendi põhihaiguse diagnoos. Kokku on andmestikus 7962 unikaalset diagnoosi. Need on kodeeritud rahvusvahelise haiguste klassifikatsiooni RHK-10 järgi. Kõige sagedasem on Z76.0 ehk korduva retsepti väljakirjutamine.
- **Diagnoosirühm**
Patsiendi põhihaiguse diagnoosirühm. Kõik diagnoosid on jagatud 20-ks rühmaks (LISA). Andmestikus puudus V00–Y98 rühm. Kõige suurem rühm on tervises seisundit mõjustavad tegurid ja kontaktid tervise teenistusega koodidega Z00–Z99 (22,21%). Väikseim diagnoosirühm on sünniperioodis tekkivad teatavad seisundid (P00–P96), mille osatähtsus on alla 0,02%.
- **Põhidiagnoosi tüüp**
Patsiendi põhihaiguse diagnoosi tüüp. Kõik diagnoosid on jagatud kolmeks rühmaks. „Esialgne“ tüüp tähendab, et patsiendi diagnoos ei ole veel kliiniliselt kinnitatud. „Esmane“ tüüp tähendab, et patsiendil on antud diagnoos kinnitatud esmakordselt. „Korduv“ tüüp tähendab, et patsiendil on diagnoos kinnitatud varasemaltki. Esialgseid diagnoosid on väga harvad ja esinevad ainult 2,65%-l raviarvel. Kõige sagedasemad on korduvad diagnoosid, nende osatähtsus on 69,60%.
- **Visiitide summa**
Patsiendi visiitide arv ühe raviarve raames.
- **Visiitide tüüp**
Vastavalt EHK tervishoiuteenuste loetelule kasutati koode:
 - perearsti esmane vastuvõtt (9001);
 - perearsti korduv vastuvõtt (9002);
 - perearsti profülaktiline vastuvõtt (9003);

- perearsti kindlustamata isikute plaaniline vastuvõtt (9016);
- perearsti telefoni teel toimunud nõustamine (9018);
- perearsti elektronposti teel toimunud nõustamine (9019);
- pereõe vastuvõtt (9061);
- pereõe telefoni teel toimunud nõustamine (9064);
- pereõe elektronposti teel toimunud nõustamine (9065)

Analüüsis grupeeriti kõik perearsti vastuvõtukoodid (9001, 9002, 9003 ja 9016) kokku.

- **Alguskuupäev**
Raviarve alguse kuupäev.
- **Lõppkuupäev**
Raviarve lõpu kuupäev.

Analüüsi ei kaasatud 33 raviarvet, kuna neil puudus patsiendi sugu, vanus või diagnoosi tüüp. Kokku on vaatluse all **3 940 648** raviarvet.

Ühel patsiendil võib aasta jooksul olla mitu raviarvet sama diagnoosiga. Kuna meid huvitab patsiendi summaarne visiitide arv aastas ühe ja sama diagnoosiga, mitte raviarve jooksul, tuli raviarved grupeerida patsiendi ning põhidiagnoosi järgi. Grupeeritud uues andmestikus on 2 724 733 rida. Igas reas on järgmised tunnused:

- patsiendi ID;
- patsiendi sugu;
- patsiendi maksimaalne vanus sel aastal;
- vanuserühm;
- patsiendi põhidiagnoos;
- diagnoosirühm;
- päevade arv esimesest visiidist aasta lõpuni sama diagnoosiga;
- põhidiagnoosi tüüp esimese visiidi ajal sel aastal;
- esimese visiidi tüüp selle diagnoosiga sel aastal;
- patsiendi visiitide arv selle diagnoosiga aasta jooksul.

2.2 Saadud loendusandmete mudeli kirjeldus

Järgnevalt kirjeldatakse mudelit prognoosimaks patsiendi visiitide arvu perearstiabi kasutamisel ühe ja sama diagnoosiga.

Tabel 1. Patsientide visiitide arvu jaotus

Visiitide arv	1	2	3	4	5	6	7+
Patsientide arv	1 203 746	721 523	335 517	192 976	95 295	62 247	113 429
Osatähtsus	44,19%	26,48%	12,31%	7,08%	3,50%	2,28%	4,16%

Andmestiku uurimisel leiti mitu patsienti, kelle summaarne visiitide arv ei ole loogiline. Näiteks üks patsient tegi 906 visiiti 87 päeva jooksul. Suure visiitide arvuga patsientide osatähtsus on üsna väike (tabel 1). Seega, nimetatud patsiendi visiitide arv ja teised patsiendid, kellel oli rohkem kui seitse visiiti, lisati „7+“ rühma. See parandab andmete kvaliteeti ja lihtsustab mudeli koostamist.

Visiitide arvu prognoosimiseks kasutati järgmiseid tunnuseid:

- patsiendi sugu;

- patsiendi maksimaalne vanus sel aastal või vanuserühm;
- patsiendi põhidiagnoos;
- diagnoosirühm;
- päevade arv esimesest visiidist aasta lõpuni sama diagnoosiga;
- põhidiagnoosi tüüp esimese visiidi ajal sel aastal;
- esimese visiidi tüüp selle diagnoosiga sel aastal.

Mudeli tegemiseks on EHK andmestik jagatud kaheks osaks. Esimene osa on treeningandmed ehk need, mida on kasutatud mudeli õppimiseks ja teine osa on testandmed, mis on vajalikud prognoosimiseks ning prognoosi täpsuse leidmiseks. Treeningvaatluste valimiseks on andmestik jagatud seitsmeks kihiks visiitide arvu järgi. Igas kihist valitakse juhuslikult 70% treeningandmete hulka, ülejäänud 30% on testandmed.

Analüüsi protsessis on vaatluse all erinevad Poissoni mudelid, sealhulgas ZTP mudelid. NB-mudelit pole vaadeldud, kuna andmestikus puudub ülehajuvus. Soo tunnus ei ole oluline iga mudeli korral. Seega edasi on vaadatud ainult järgmisi tunnuseid: vanus või vanuserühm, diagnoosirühm, päevade arv aasta lõpuni, diagnoosi tüüp ning esimese visiidi tüüp.

Tabel 2. Mudelite võrdlus

	AIC
Poisson vanuserühmaga	7 136 603
Poisson vanusega	7 158 217
ZTP vanuserühmaga	6 407 448
ZTP vanusega	6 441 551

Väikseima AIC-i väärtusega mudel kirjeldab andmeid paremini (3). Tabelis 2 on näha, et AIC-i väärtus on väiksem, kui mudelis kasutatakse tunnuse vanuse asemel vanuserühm. Kõige väiksem AIC on ZTP mudelil parameetritega vanuserühm, diagnoosirühm, diagnoosi tüüp, visiidi tüüp ning päevade arv esimesest visiidist aasta lõpuni. ZTP mudel on eelistatum ka nullide puudumise tõttu (andmestikus on patsiente, kes on vähemalt korra pöördunud perearstiabisse). Saadud mudel on esitatud allpool.

$$\begin{aligned}
\text{Visiitide arv} = & \exp(-0,07 - 0,43 \times I(\text{vanus} = 15 - 24) - 0,26 \times (vanus = 25 - 44) \\
& - 0,14 \times I(vanus = 45 - 64) - 0,35 \times I(vanus = 5 - 14) \\
& - 0,14 \times I(vanus = 65 - 84) - 0,15 \times I(vanus = 85 +) \\
& + 0,05 \times I(diagngr = C00 - D48) + 0,06 \times I(diagngr = D50 - D89) \\
& - 0,06 \times I(diagngr = E00 - E90) + 0,04 \times I(diagngr = F00 - F99) \\
& + 0,02 \times I(diagngr = G00 - G99) - 0,56 \times I(diagngr = H00 - H59) \\
& - 0,27 \times I(diagngr = H60 - H95) + 0,14 \times I(diagngr = I00 - I99) \\
& + 0,03 \times I(diagngr = J00 - J99) - 0,18 \times I(diagngr = K00 - K93) \\
& - 0,38 \times I(diagngr = L00 - L99) - 0,02 \times I(diagngr = M00 - M99) \\
& - 0,1 \times I(diagngr = N00 - N99) + 0,007 \times I(diagngr = O00 - O99) \\
& - 0,52 \times I(diagngr = P00 - P96) - 0,46 \times I(diagngr = Q00 - Q99) \\
& - 0,25 \times I(diagngr = R00 - R99) + 0,23 \times I(diagngr = S00 - T98) \\
& - 0,39 \times I(diagngr = Z00 - Z99) + 0,27 \times (tüüp = esmane) \\
& + 0,17 \times (tüüp = korduv) + 0,0025 \times \text{päevade arv} \\
& + 0,02 \times (vis tüüp = 9018) + 0,11 \times (vis tüüp = 9019) \\
& + 0,2 \times (vis tüüp = 9061) + 0,29 \times (vis tüüp = 9064) \\
& + 0,36 \times (vis tüüp = 9065))
\end{aligned}$$

Vaatamata sellele, et lõppmudelil on parameetreid palju, on tulemuste interpreteerimine üsna lihtne. Parameetrid positiivse koefitsiendiga suurendavad prognoositud visiitide arvu ja negatiivsed vähendavad seda. Mudelist on näha, et koefitsiendid on kõikide vanuserühmade ees negatiivsed. See tähendab, et patsiendid vanuses 0–4 pöörduvad perearstiasse sagedamini kui teevad seda vanemad patsiendid. Näiteks, kui on kaks isikut erineva diagnoosiga, kuid samast diagnoosirühmast, sama diagnoosi ja visiidi tüübiga ning pöördumiskuupäev on sama, kuid üks neist on vanuserühmas 0–4 ja teine 5–14, siis vanem patsient pöörduv perearstiasse aasta jooksul keskmiselt $\frac{1}{\exp(-0,35)} = 1,43$ korda harvem, kui laps vanuses 0–4.

Erinevaid teisi diagnoosirühmi võrreldi diagnoosirühmaga A00–B99 (teatavad nakkus- ja parasiithaigused). See tähendab, et kui on kaks isikut samast vanuserühmast, sama diagnoosi ja visiidi tüübiga ning pöördumiskuupäev on sama, kuid ühel neist on diagnoos diagnoosirühmast A00–B99 ja teisel H00–H59 (silma- ja silmamanuste haigused), siis teine patsient pöörduv samal aastal perearstiasse keskmiselt $\frac{1}{\exp(-0,56)} = 1,75$ korda harvem, kui esimene. Aga, kui teisel patsiendil on diagnoos diagnoosirühmast I00–I99 (vereringeelundite haigused), siis pöörduv ta perearstiasse $\exp(0,14) = 1,15$ korda sagedamini kui esimene patsient.

Võrreldes kaht samade parameetritega patsienti, kellest ühel on esimesel visiidil esialgne diagnoos ehk kui diagnoos ei ole veel kinnitatud, siis ta pöörduv aasta jooksul perearstiasse keskmiselt $\exp(0,27) = 1,31$ korda vähem kui patsient esmase diagnoosiga ja $\exp(0,17) = 1,19$ korda vähem kui korduva diagnoosiga patsient.

Visiitide arv sõltub kuupäevast, millal patsient pöördus. Kui kaks sarnast patsienti pöörduvad perearstiasse, siis patsient, kes pöördus päeva võrra varem, pöörduv sel aastal keskmiselt $\exp(0,0025) = 1,0025$ korda sagedamini kui hilisem patsient.

Samuti sõltub visiitide arv esimese visiidi tüübist. Näiteks patsient, kelle esimene visiit oli pereõe vastuvõtt (koodiga 9061), pöörduv $\exp(0,2) = 1,22$ korda sagedamini kui sarnane patsient, kelle esimene visiit oli perearsti vastuvõtt (9001).

Saadud mudeli järgi pöörduv kõige enam perearstiasse patsient vanuses 0–4, kelle esmane diagnoos on diagnoosirühmast S00–T98 (vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed) ja tema esimene visiit on olnud pereõe nõustamine telefoni teel ning see toimus aasta esimesel päeval. Nende patsientide keskmine visiitide arv aastas on 5,9.

Kõige harvem pöörduv perearstiasse patsient vanuses 15–24, kellel on esialgne diagnoos diagnoosirühmast H00–H59 (silma- ja silmamanuste haigused) ja tema esimene visiit oli perearsti vastuvõtt, mis toimus aasta viimasel päeval. Sel juhul on nende patsientide keskmine visiitide arv aastas 0,34.

Tabel 3. Mudelite võrdlus

	RMSE	MAPE	MAD
Poisson vanuserühmaga	1,52	63,76%	1,14
Poisson vanusega	1,52	64,24%	1,15
ZTP vanuserühmaga	1,56	51,11%	1,08
ZTP vanusega	1,57	51,27%	1,09

Saadud mudeli (ZTP vanuserühmaga) RMSE on 1,56; MAPE on 51,11%; MAD on 1,08 (tabel 3). Seega ei saa öelda, et mudel on väga täpne, kuid see on parim mudel vaadeldud mudelite hulgas MAPE ja MAD järgi (RMSE põhjal on Poisson vanuserühmaga parem, aga erinevus pole suur).

Tabel 4. Prognoosi täpsus testandmetes

Prognoosi erinevus reaalsest väärtusest (visiidid)	<0,5	0,5–1	1–2	2–3	>3
Osatähtsus	36,57%	25,85%	23,00%	6,71%	7,87%

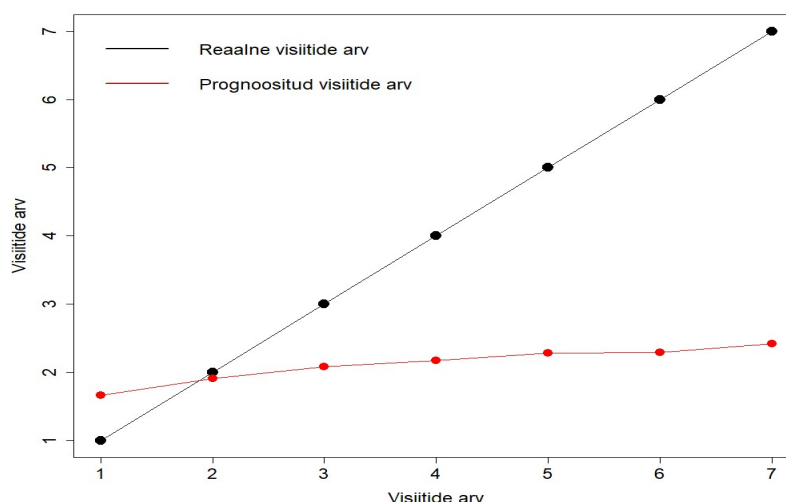
Tabelis 4 on näha, et 36,57% andmetest on prognoositud täpsusega 0,5 ning 62,42% juhtudest prognoosi viga on ühest väiksem.

Vaatleme eraldi prognoosi täpsust patsientidele erinevate visiitide arvuga.

Tabel 5. Prognoosi täpsus testandmetes reaalseste visiitide arvu järgi

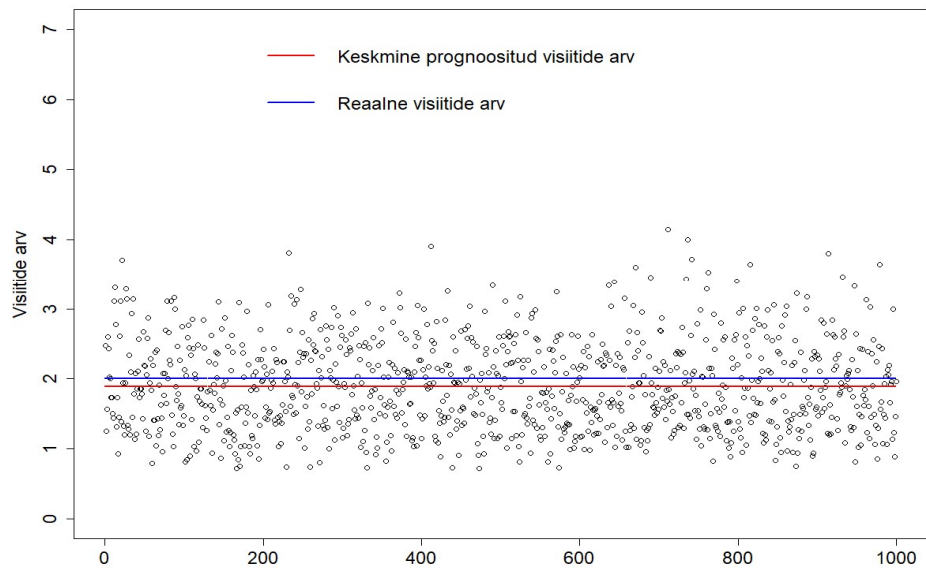
Prognoosi erinevus reaalsest väärtusest	<0,5	0,5–1	1–2	2–3	>3
Ühe visiidiga patsientide hulk	46,46%	26,51%	24,37%	2,57%	0,09%
Kahe visiidiga patsientide hulk	49,55%	39,47%	10,76%	0,22%	0,00%
Kolme visiidiga patsientide hulk	25,70%	26,44%	45,39%	2,46%	0,00%
Nelja visiidiga patsientide hulk	2,83%	8,56%	46,60%	40,11%	1,90%
Viie visiidiga patsientide hulk	0,02%	0,78%	13,40%	50,91%	34,90%
Kuue visiidiga patsientide hulk	0,00%	0,00%	0,59%	13,68%	85,73%
Vähemalt seitsme visiidiga patsientide hulk	0,00%	0,00%	0,00%	1,30%	98,70%

Tabelis 5 on näha, et nii ühe, kahe kui kolme reaalse visiidi korral töötab mudel üsna täpselt ja 96–99% juhtudest on prognoosi viga väiksem kui kaks. Kuid suurema reaalseste visiitide arvu korral prognoosib mudel ebatäpselt. Näiteks kuue visiidi puhul on prognoosi viga kahest suurem 99,41% juhtudel.



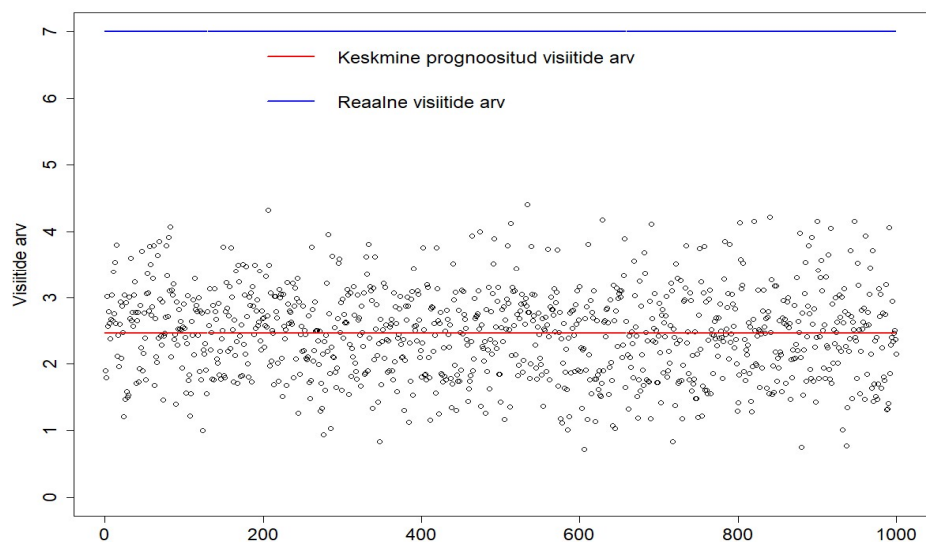
Joonis 1. Keskmine prognoositud ning reaalne visiitide arv EHK andmete põhjal

Sama on kujutatud ka joonisel 1, kus punane joon näitab, kuidas muutub keskmine prognoositud visiitide arv reaalsete visiitide (must sirge) suurenemisel. Samuti on näha, et keskmine prognoositud visiitide arv kasvab väga aeglaselt, mistõttu suure visiitide arvu hulgas on prognoos ebatäpne.



Joonis 2. Prognoositud visiitide arv, kui reaalne visiitide arv on võrdne kahega

See ebatäpsus on ülevaatliselt visualiseeritud joonisel 2 ning joonisel 3. Kui reaalne visiitide arv testandmestikus on võrdne kahega (joonis 2), siis keskmine prognoositud visiitide arv on umbes 1,9 ja suur osa prognoosidest (mustad punktid) on reaalse väärtuse lähedal.



Joonis 3. Prognoositud visiitide arv, kui reaalne visiitide arv on võrdne seitsmega

Kui reaalne visiitide arv on seitse või rohkem (joonis 3), siis keskmine prognoositud visiitide arv on ainult 2,6 ning minimaalne prognoosi viga on kahest suurem.

2.3 Saadud logistilise mudeli kirjeldus

Antud peatükis kirjeldatakse mudelit tõenäosuse prognoosimiseks, et patsient pöördub perearstiabisse sama diagnoosiga vähemalt kaks korda aastas. Selleks loodi uus binaarne tunnus, mis on võrdne nulliga, kui visiitide arv on üks, ja ühega, kui visiitide arv kaks või rohkem.

Tabel 6. Visiitide arv treeninguandmestikus

Visiitide arv	Üks	Kaks või rohkem
Patsientide arv	946 167	1 182 335
Osatähtsus	44,45%	55,55%

Mudeli õppimiseks ning prognoosi täpsuse kontrollimiseks kasutatakse sama treening- ja testandmestikku nagu eelmises punktis. Treeningandmestikus on 44,45% patsientidest, kes pöördusid perearstiabisse ühe ja sama diagnoosiga ainult korra aastas ning 55,55% pöördusid vähemalt kaks korda (tabel 6).

Tabel 7. Seosefunktsioonide võrdlemine

Seosefunktsioon	AIC	Hosmer-Lemeshow'i testi <i>p</i> -väärtus
Logit	2711414	$< 2 \times 10^{-16}$
Probit	2711513	$< 2 \times 10^{-16}$
Clog-log	2715391	$< 2 \times 10^{-16}$

Binaarse tunnuse prognoosimiseks on võimalik kasutada *Logit*, *Probit* ning *Clog-Log* seosefunktsiooni. Need kolm mudelit testiti parameetritega sugu, vanuserühm, diagnoosirühm, diagnoosi tüüp, visiidi tüüp ning päevade arv aasta lõpuni. *p*-väärtused näitavad, et kõik parameetrid on olulised kõigi mudelite korral. AIC väärtuse põhjal (tabel 7) parimaks mudeliks on mudel *Logit* seosefunktsiooniga (esitatud allpool). Kuid *Hosmer-Lemeshow*'i testi põhjal võib öelda, et mitte ükski mudel ei kirjelda andmeid hästi (*p*-väärtus $< 0,05$).

$$\begin{aligned}
\Pi = & \exp(-0,42 - 0,03 \times I(\text{sugu} = \text{naine}) - 0,52 \times I(\text{vanus} = 15 - 24) \\
& - 0,3 \times I(\text{vanus} = 25 - 44) - 0,13 \times I(\text{vanus} = 45 - 64) \\
& - 0,48 \times I(\text{vanus} = 5 - 14) - 0,16 \times I(\text{vanus} = 65 - 84) \\
& - 0,24 \times I(\text{vanus} = 85 +) - 0,23 \times I(\text{diagngr} = C00 - D48) \\
& + 0,25 \times I(\text{diagngr} = D50 - D89) - 0,04 \times I(\text{diagngr} = E00 - E90) \\
& - 0,2 \times I(\text{diagngr} = F00 - F99) - 0,18 \times I(\text{diagngr} = G00 - G99) \\
& - 0,96 \times I(\text{diagngr} = H00 - H59) - 0,51 \times I(\text{diagngr} = H60 - H95) \\
& + 0,22 \times I(\text{diagngr} = I00 - I99) + 0,07 \times I(\text{diagngr} = J00 - J99) \\
& - 0,31 \times I(\text{diagngr} = K00 - K93) - 0,77 \times I(\text{diagngr} = L00 - L99) \\
& - 0,14 \times I(\text{diagngr} = M00 - M99) - 0,2 \times I(\text{diagngr} = N00 - N99) \\
& + 0,007 \times I(\text{diagngr} = O00 - O99) - 0,75 \times I(\text{diagngr} = P00 - P96) \\
& - 0,78 \times I(\text{diagngr} = Q00 - Q99) - 0,37 \times I(\text{diagngr} = R00 - R99) \\
& + 0,26 \times I(\text{diagngr} = S00 - T98) - 0,92 \times I(\text{diagngr} = Z00 - Z99) \\
& + 0,47 \times (\text{tüüp} = \text{esmane}) - 0,11 \times (\text{tüüp} = \text{korduv}) \\
& + 0,0037 \times \text{päevade arv} + 0,05 \times (\text{vis tüüp} = 9018) \\
& + 0,26 \times (\text{vistüüp} = 9019) + 0,47 \times (\text{vistüüp} = 9061) \\
& + 0,79 \times (\text{vis tüüp} = 9064) + 0,97 \times (\text{vis tüüp} = 9065))
\end{aligned}$$

Positiivsed koefitsiendid mudelis suurendavad šanssi, et patsient pöördub perearstiasse vähemalt kaks korda aastas ühe diagnoosiga ja negatiivsed koefitsiendid vähendavad seda.

Näiteks, kui perearstiasse pöördus samal päeval mees ja naine samast vanuserühmast, diagnoosidega samast rühmast ning sama diagnoosi ja visiidi tüübiga, siis šanss, et mees pöördub rohkem kui ühe korra on $\frac{1}{\exp(-0,03)} = 1,03$ korda suurem kui naise puhul.

Patsientidel vanuserühmas 0–4 on suurem šanss perearstiasse pöörduda enam kui ühe korra, kui teistes vanuserühmades. Kui perearstiasse pöörduvad kaks sarnast patsienti, üks neist on vanuserühmas 0–4 ja teine 15–24, siis šanss, et lapsega pöördutakse veel üks kord on $\frac{1}{\exp(-0,52)} = 1,68$ suurem, kui patsiendil vanuses 15–24.

Patsiendid, kelle diagnoosi tüüp on esmane, tulevad perearsti juurde $\exp(0,47) = 1,6$ korda suurema šansiga, kui sama patsient, kelle diagnoosi tüüp on esialgne.

Teise visiidi toimumise šanss sõltub ka kuupäevast, mil patsient pöördub perearstiasse. Kui kaks võrdset patsienti pöörduvad perearstiasse, siis patsient, kes pöördus päev varem, teeb sel aastal veel korra visiidi $\exp(0,0037) = 1,0037$ korda suurema šansiga, kui hilisem patsient.

Samuti sõltub šanss diagnoosirühmast ja esimese visiidi tüübist. Näiteks, kui ühel patsiendil on diagnoos rühmast H00–H59 (silma- ja silmamanuste haigused) ja teisel on A00–B99 (teatavad nakkus- ja parasiithaigused) ning ülejäänud näitajad on sarnased, siis teisel patsiendil on šanss veel korra pöörduda perearstiasse $\frac{1}{\exp(-0,96)} = 2,61$ korda suurem kui esimesel.

Saadud mudeli järgi on kõige suurem šanss, et patsient tuleb perearsti juurde rohkem kui üks kord aastas, poistel vanuses 0–4, kellel esmane diagnoos on diagnoosirühmast S00–T98 (vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed) ja esimeseks visiidiks on pereõe nõustamine telefoni teel, mis toimus aasta esimesel päeval. Väikseim šanss, et patsient pöördub perearstiasse rohkem kui üks kord aasta jooksul, on naistel vanuses 15–24, kellel on korduv diagnoos diagnoosirühmast H00–H59 ja tema esimene visiit on perearsti vastuvõtt, mis toimus aasta viimasel päeval.

Teades šanssi, on võimalik leida tõenäosust, et patsient pöördub aastas perearstiasse sama diagnoosiga ka teist korda.

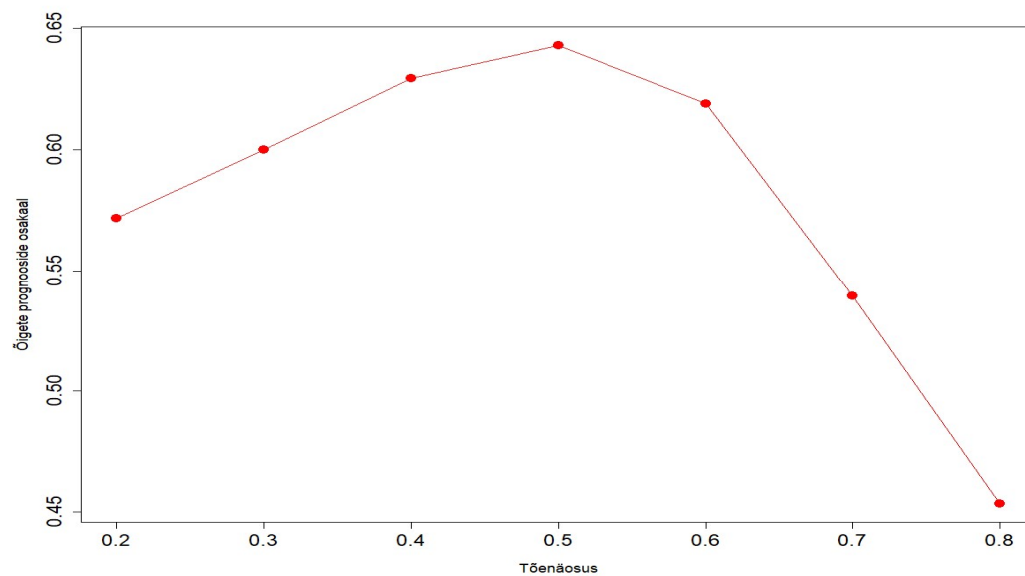
Kasutades testandmestikku, kontrollime, kui hästi töötab valitud mudel. Eeldame, et teise visiidi tõenäosus on suurem või võrdne 0,5-st, siis prognoosime, et patsient pöördub rohkem kui üks kord

perearstiabisse selle diagnoosiga aastas. Vastasel juhul on prognoosiks ainult üks visiit selle diagnoosiga sel aastal.

Tabel 8. Prognoosi täpsus

Prognoos	Reaalne visiitide arv	
	Üks	Vähemalt kaks
Üks	20,37%	12,96%
Vähemalt kaks	22,74%	43,93%

Tabelist 8 on näha, et mudel prognoosib õiget tulemust 64,3% juhtudest ehk kaks kolmandikku patsientidest on õigesti klassifitseeritud.



Joonis 4. Mudeli täpsus erinevate valitud tõenäosuste korral

Kui proovida teisi valitud tõenäosusi, muutub mudel ebatäpsemaks. Seda on näha joonisel 4. Näiteks, kui valida tõenäosuseks 0,8, klassifitseerib mudel patsiente õigesti ainult 45% juhtudest.

3 Mudelite testimine tervise infosüsteemi andmestiku abil

Eelmises peatükis kirjeldatud mudelid olid tehtud Eesti Haigekassa andmestiku põhjal. Järgnevas peatükis vaatleme, kui hästi töötavad saadud mudelid (ZTP ning Logit-mudel) tervise infosüsteemi andmete peal.

3.1 Tervise infosüsteemi andmestiku kirjeldus

Tervise infosüsteemi (TIS) andmestik on oma struktuuri põhjal sarnane Eesti Haigekassa andmestikuga. TIS-i koondatakse sarnane informatsioon perearsti visiitidest üle Eesti. Siinkohal on TIS andmestikus huvipakkuvad 2017. aasta 1 669 820 epikriisi.

Igast epikriisist on teada

- **Dokumendi number**
Unikaalne kood epikriisile.
- **Patsiendi ID**
Numbriline kood, mis identifitseerib patsiendi. Kodeerimine erineb EHK andmestikust. TIS andmestikus on olemas 720 362 unikaalset patsienti, mida on 1,5 korda vähem, kui EHK andmestikus.
- **Sugu**
Patsiendi sugu. 42,0% epikriisidest on välja antud meestele ning 58,0% naistele. Samuti leidub 12 epikriisi, millel sugu on teadmata.
- **Vanus**
Patsiendi vanus haiguse alguse ajal. Isikute vanus varieerub andmestikus 0-st kuni 107-aastani nagu EHK andmestikus. Leidub 10 epikriisi teadmata vanusega.
- **Vanuserühm**
Patsiendi vanuserühm epikriisi kirjutamise ajal. Kuna vanuserühmi kasutatakse saadud mudelites prognoosimiseks sarnaselt, moodustati vanuserühmad nagu oli EHK andmestikus (0–4; 5–14; ..., 85+).
- **Eriala kood**
Kodeeritud eriala. Vaadeldi ainult peremeditiini (koodiga E300) ning pereõendust (N320).
- **Põhidiagnoos**
Patsiendi põhidiagnoos. Kokku on andmestikus 7467 unikaalset diagnoosi, mis on samuti kodeeritud rahvusvahelise haiguste klassifikatsiooni RHK-10 järgi nagu EHK andmestikus. 68-l epikriisil on diagnoos teadmata.
- **Põhidiagnoosi tüüp**
Patsiendi põhidiagnoosi tüüp. Kõik diagnoosid on jagatud kolmeks rühmaks nagu EHK andmestikus (esialgne, esmane, korduv).
- **Visiitide summa**
Patsiendi visiitide arv ühe epikriisi raames.
- **Visiitide tüüp**
Esimese visiidi tüüp epikriisi raames. 1 – vastuvõtt, 2 – koduviit, 3 – nõustamine telefoni teel, 4 – nõustamine elektronposti teel. Koduviite antud analüüsis ei vaadelda. Kasutades tunnust „eriala kood“ on põhimõtteliselt võimalik visiidi tüüpi kodeerida nagu EHK andmestikus.
- **Algus- ja lõpukuupäev**
Epikriisi alguse ja lõpu kuupäev.

- TIS andmestikus puudub tunnus „diagnoosirühm“. Kuna see on vajalik saadud mudelite kasutamiseks, moodustati tunnus iseseisvalt.

Analüüsi ei kaastatud 90 epikriisi, kuna nendes puudus patsiendi sugu, vanus või diagnoos. Seega on vaatluse all **1 669 730** epikriisi. Tähele tuleb panna, et TIS andmestikus on vigaste epikriiside osatähtsus 6,42 korda suurem kui EHK andmestikus.

Eelmises peatükis kirjeldatud mudelite kasutamiseks moodustakse uus andmestik grupeeritud epikriisidega patsiendi ning diagnoosi järgi, nagu tehti EHK andmestikuga. Uues andmestikus on 1 384 424 rida. Igas reas on järgmised tunnused:

- patsiendi ID;
- patsiendi sugu;
- patsiendi maksimaalne vanus sel aastal;
- vanuserühm;
- patsiendi põhidiagnoos;
- diagnoosirühm;
- päevade arv esimesest visiidist selle diagnoosiga aasta lõpuni;
- põhidiagnoosi tüüp esimese visiidi ajal sel aastal;
- esimese visiidi tüüp selle diagnoosiga sel aastal;
- patsiendi visiitide arv selle diagnoosiga aasta jooksul.

3.2 Loendusandmete mudeli testimine

Tervise infosüsteemi andmestikus erineb patsientide visiitide arvu jaotus Haigekassa andmete põhjal saadud jaotusest. Siin on ühekordsete visiitide osatähtsus kaks korda suurem (EHK – 44,19%, TIS – 84,78%). Seetõttu suurema visiitide arvuga patsientide osatähtsus on palju väiksem kui EHK andmetes (tabel 1 ja tabel 9).

Tabel 9. Patsientide visiitide arvu jaotus

Visiitide arv	1	2	3	4	5	6	7+
Patsientide arv	1 153 340	137 459	37 333	14 847	7 287	3 928	6 242
Osatähtsus	84,78%	10,10%	2,74%	1,09%	0,54%	0,29%	0,46%

Eelmises peatükis kirjeldatud ZTP mudeli RMSE on 0,97 ning MAD on 0,71. See on 1,5 korda väiksem, kui EHK testandmestikus. Seega võib öelda, et TIS andmestikus annab mudel täpsema prognoosi.

Tabel 10. Prognoosi täpsus TIS andmetes

Prognoosi erinevus reaalsest väärtusest	<0,5	0,5–1	1–2	2–3	>3
Osatähtsus	47,55%	25,93%	22,75%	2,82%	0,94%

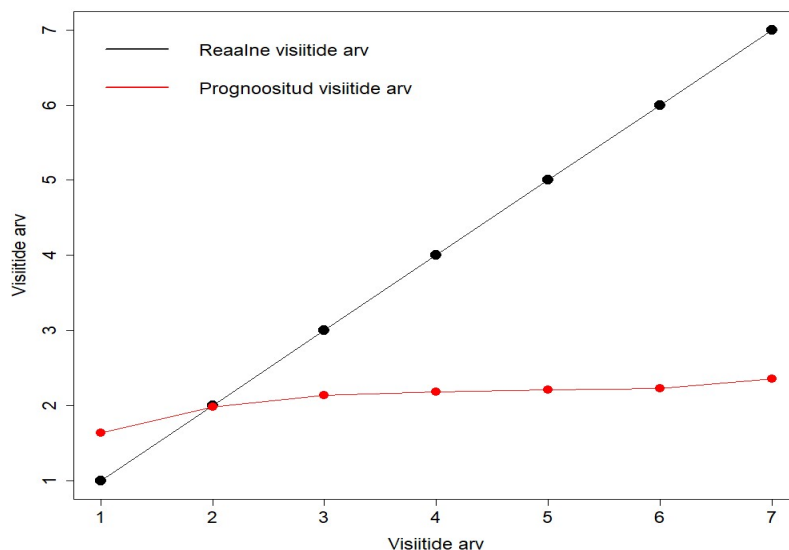
Tabelis 10 on näha, et 47,55% andmetest on prognoositud täpsusega 0,5 ning 73,48% juhtudest on prognoosi viga ühest väiksem. See on 10% võrra täpsem, kui oli EHK testandmestikus. Nagu eelmises peatükis öeldud, on saadud ZTP mudeli prognoosid väga ebatäpsed, kui patsiendi visiitide

arv on suurem neljast. Kuna TIS andmestikus on vähem neid patsiente, kelle visiitide arv ületab nelja, näitab mudel paremat tulemust.

Tabel 11. Prognoosi täpsus TIS andmetes reaalseste visiitide arvu järgi

Prognoosi erinevust reaalsest väärtusest	<0,5	0,5–1	1–2	2–3	>3
Ühe visiidiga patsientide hulk	48,75%	25,10%	23,72%	2,42%	0,01%
Kahe visiidiga patsientide hulk	53,64%	37,27%	9,04%	0,05%	0,00%
Kolme visiidiga patsientide hulk	28,58%	28,59%	41,37%	1,46%	0,00%
Nelja visiidiga patsientide hulk	1,25%	9,52%	49,01%	39,29%	0,94%
Viie visiidiga patsientide hulk	0,05%	0,38%	11,35%	48,40%	39,81%
Kuue visiidiga patsientide hulk	0,00%	0,00%	0,46%	12,81%	86,74%
Vähemalt seitsme visiidiga patsientide hulk	0,00%	0,00%	0,00%	0,70%	99,30%

Vaatleme eraldi prognoosi täpsust patsientide erinevate visiitide arvuga. Kui võrrelda omavahel tabeli 5 ja tabeli 11 andmeid, siis on näha, et prognoosi täpsuses ei ole suurt erinevust. Mudel töötab hästi, kui reaalseste visiitide arv on üks, kaks või kolm. Sel juhul on prognoosi viga väiksem kui kaks visiiti umbes 98% juhtudest. Kuid suure visiitide arvu korral on prognoos halb, nagu oli ka EHK testandmestikus. Näiteks, kuue visiidi puhul on prognoosi viga suurem kahest visiidist 99,55% juhtudest.



Joonis 5. Keskmine prognoositud ning reaalne visiitide arv TIS andmete põhjal

Joonisel 5 on näha, kuidas muutub keskmine prognoositud visiitide arv reaalseste visiitide arvu suurenemisel. Samuti nähtub, et keskmine prognoositud visiitide arv kasvab väga aeglaselt. Seetõttu on visiitide suure arvu korral prognoos ebatäpne. Olukord on täpselt sama, mis oli EHK testandmestikus (joonis 1).

3.3 Logistilise mudeli testimine

Selles peatükis analüüsime, kui hästi töötab saadud *logit*-mudel, mis TIS andmestiku põhjal prognoosib tõenäosust, et patsient pöördub perearstiabisse sama diagnoosiga vähemalt kaks korda aastas. Selleks on loodud TIS andmestikus uus binaarne tunnus, mis on võrdne nulliga, kui visiitide arv on üks, ning ühega, kui visiitide arv on kaks või rohkem.

Tabel 12. Visiitide arv TIS andmestikus

Visiitide arv	Üks	Kaks või rohkem
Patsientide arv	1 153 340	207 096
Osatähtsus	84,78%	15,22%

TIS andmestiku järgi pöördus 84,78% patsientidest perearstiabisse ühe ja sama diagnoosiga ainult üks kord ja 15,22% pöördusid vähemalt kaks korda aastas (tabel 12).

Eeldame, et kui prognoositud tõenäosus on suurem või võrdne 0,5-st (nagu eelmises peatükis), siis prognoosi tulemuseks on vähemalt kaks visiiti aasta jooksul sama diagnoosiga. Vastasel juhul on prognoosi tulemus üks visiit sama diagnoosiga aastas.

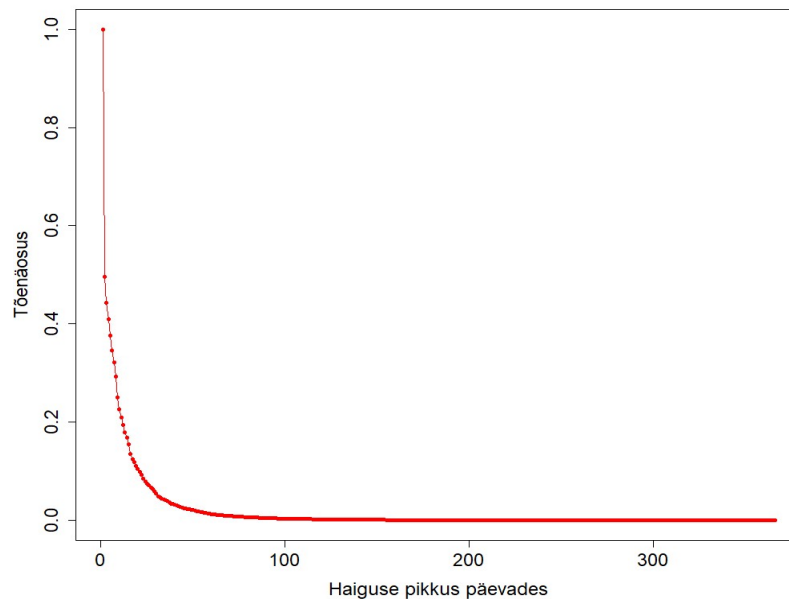
Tabel 13. Prognoosi täpsus

Reaalne visiitide arv	Üks	Vähemalt kaks
Prognoos		
Üks	40,80%	3,91%
Vähemalt kaks	43,97%	11,31%

Tabelis 13 on näha, et mudel prognoosib õigesti patsiendi visiite 40,8% juhtudest, kui visiit toimub korra aastas ja vaid 11,31% juhtudest õigesti, kui patsient pöördus perearstiabisse vähemalt kaks korda. Ehk ainult pooled patsiendid on õigesti klassifitseeritud. Saadud tulemus on 12% võrra madalam, kui oli prognoos EHK testandmestikus. Mudeli järgi pöördub patsient perearstiabisse rohkem kui korra aastas, kuid tegelikult pöördus patsient vaid ühe korra. Selline viga esineb 43,97% juhtudest. Teist tüüpi viga, kus patsient pöördub mudeli järgi perearstiabisse ühe korra aastas, kuid tegelikult pöördub mitu korda, esineb palju harvem (3,91%).

4 Haiguskestuse prognoosimine tervise infosüsteemi andmestikus

Patsiendi haiguskestusest sõltub perearstiabi töömaht. Mida kauem kestab patsiendi haigus, seda rohkem tööd on perearsti meeskonnal. Seega on huvipakkuvaks tunnuseks patsiendi ühe haiguse pikkus 2017. aasta jooksul. Selle tunnuse modelleerimiseks valitakse parim elukestusmodel *Weibull*, *log-logistic*, *log-normal* ning *Cox* mudelite hulgast. Andmed on võetud TIS andmestikust. Lihtsa juhusliku valiku järgi on andmestik jagatud kaheks osaks – treeningandmestik (70,0%) ning testandmestik (30,0%).



Joonis 6. Haiguskestuse funktsioon

Joonisel 6 on esitatud haiguskestuse funktsioon, mis näitab tõenäosust p , et patsient on haige kauem kui x päeva. Haiguse alguses (0. päev) on kõik patsiendid haiged ehk tõenäosus, et patsient on haige kauem kui null päeva on võrdne ühega. Treeningandmestiku põhjal saavad pooled patsiendid terveks ühe päeva jooksul. Seega tõenäosus, et patsient on haige kauem kui ühe päeva, on 0,496. Haiguse pikkus enam kui seitse päeva on tõenäosusega 0,29.

Tabel 14. Parameetriliste mudelite saadud AIC

Mudel	AIC
Weibull	6 577 055
Log-logistic	6 250 598
Log-normal	6 214 736

Esmalt vaatleme kolme parameetrilist mudelit (*Weibull*, *Log-logistic*, *Log-normal*) tunnustega sugu, vanuserühm, diagnoosirühm, esimese visiidi ning diagnoosi tüüp. Iga mudeli korral on kõik parameetrid olulised (p -väärtus $< 0,05$). Võrreldes mudeleid omavahel AIC'i põhjal on parimaks *Log-normal* elukestusmodel (tabel 14).

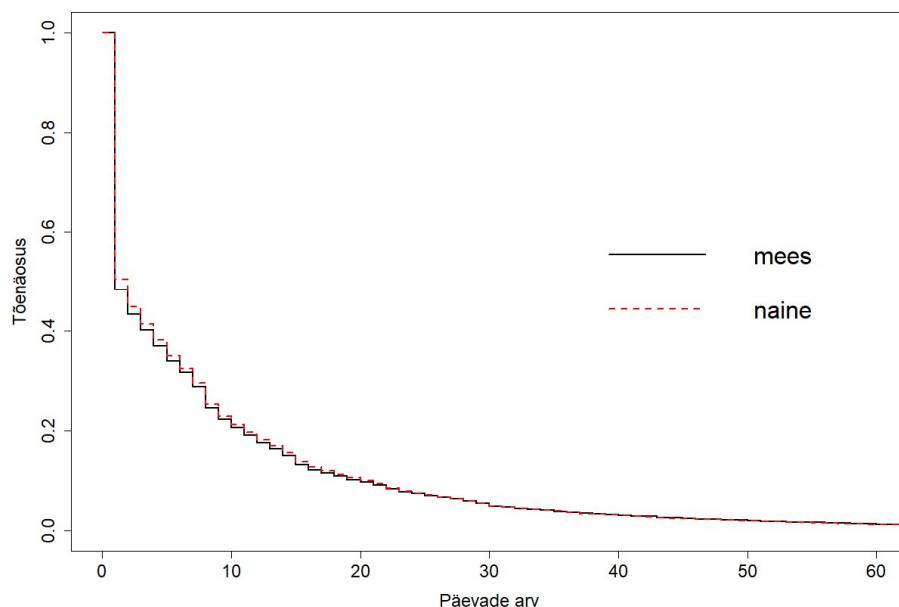
Mitteparameetrilise Cox mudeli korral on kõik tunnused olulised. Mudeli sobivuse kontrollimiseks kasutame *Grambsch-Therneau* testi, mis kontrollib, kas Cox mudeli eeldused on täidetud (2). Testi põhjal ei sobi saadud mudel andmete kirjeldamiseks (p -väärtus $<0,05$). Seega edasi prognoosimiseks kasutatakse *Log-normal* mudelit (esitatud allpool).

$$\begin{aligned}
 E(\ln(T)) = & \exp(1,23 + 0,01 \times I(\text{sugu} = \text{naine}) - 0,002 \times I(\text{vanus} = 15 - 24) \\
 & + 0,11 \times I(\text{vanus} = 25 - 44) - 0,2 \times I(\text{vanus} = 45 - 64) \\
 & - 0,14 \times I(\text{vanus} = 5 - 14) + 0,13 \times I(\text{vanus} = 65 - 84) \\
 & - 0,02 \times I(\text{vanus} = 85 +) - 0,03 \times I(\text{diagngr} = C00 - D48) \\
 & + 0,06 \times I(\text{diagngr} = D50 - D89) + 0,02 \times I(\text{diagngr} = E00 - E90) \\
 & - 0,31 \times I(\text{diagngr} = F00 - F99) - 0,34 \times I(\text{diagngr} = G00 - G99) \\
 & - 0,55 \times I(\text{diagngr} = H00 - H59) - 0,28 \times I(\text{diagngr} = H60 - H95) \\
 & - 0,03 \times I(\text{diagngr} = I00 - I99) + 0,03 \times I(\text{diagngr} = J00 - J99) \\
 & + 0,015 \times I(\text{diagngr} = K00 - K93) - 0,46 \times I(\text{diagngr} = L00 - L99) \\
 & - 0,007 \times I(\text{diagngr} = M00 - M99) + 0,06 \times I(\text{diagngr} = N00 - N99) \\
 & + 0,16 \times I(\text{diagngr} = O00 - O99) + 0,16 \times I(\text{diagngr} = P00 - P96) \\
 & - 0,37 \times I(\text{diagngr} = Q00 - Q99) - 0,08 \times I(\text{diagngr} = R00 - R99) \\
 & + 0,54 \times I(\text{diagngr} = S00 - T98) - 0,8 \times I(\text{diagngr} = Z00 - Z99) \\
 & + 0,22 \times (\text{tüüp} = \text{esmane}) - 0,12 \times (\text{tüüp} = \text{korduv}) \\
 & - 0,24 \times (\text{vis tüüp} = 9018) - 0,03 \times (\text{vis tüüp} = 9061) \\
 & + 0,35 \times (\text{vis tüüp} = 9064))
 \end{aligned}$$

Saadud mudel näitab keskmist logaritmitud haiguskestust. Oodatud haiguse pikkuse saamiseks on vaja kasutada saadud skaala parameetrit σ . Antud juhul $\sigma=1,2$. Seega oodatud haiguse pikkuse saab järgmisest valemist:

$$\text{Haiguskestus} = E(\ln(T)) * \exp\left(\frac{\sigma^2}{2}\right) = E(\ln(T)) * 2,09,$$

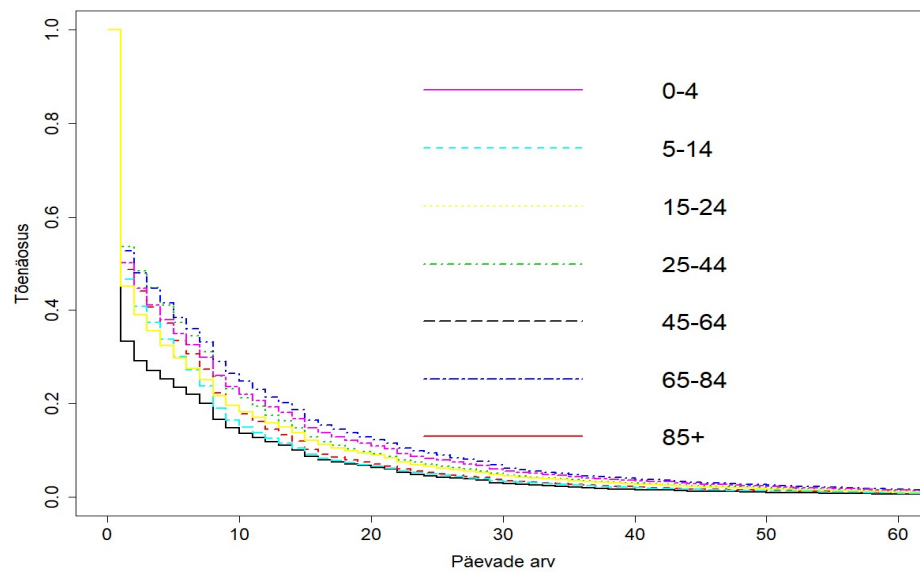
kus $E(\ln(T))$ on eelnevalt kirjeldatud mudel.



Joonis 7. Haiguskestuse funktsioon meeste ja naiste jaoks

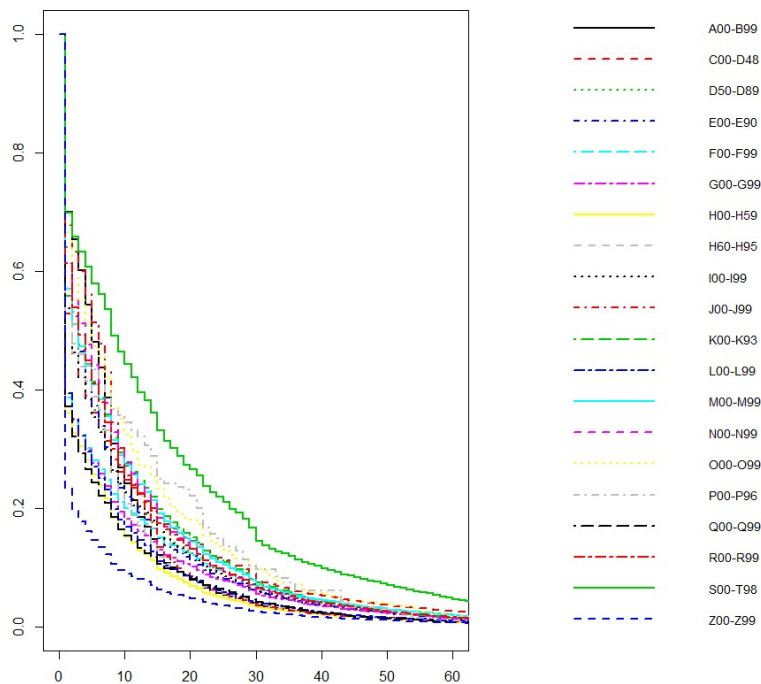
Positiivsed koefitsiendid suurendavad haiguskestuse aega ja negatiivsed vähendavad seda. Kui vaadeldakse kaht patsienti ühest vanuserühmast ja sama diagnoosirühmaga, sama diagnoosi ja

esimese visiidi tüübiga, kuid üks neist on mees ja teine naine, siis naise haiguskestus on keskmiselt $\exp(0,01) = 1,01$ korda pikem. Joonisel 7 on näha, et naiste ja meeste haiguskestuse funktsioonid on väga sarnased ning naiste puhul kattub joon natuke meeste omaga.



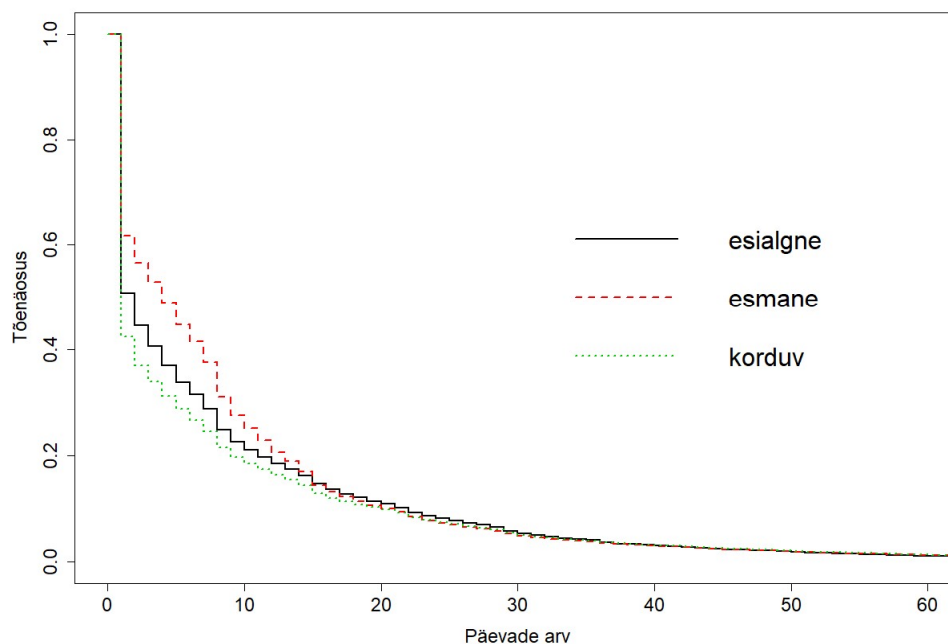
Joonis 8. Haiguskestuse funktsioon erinevate vanuserühmade jaoks

Haiguskestust mõjutab patsiendi vanuserühm (joonis 8). Kui üks patsient on vanuserühmast 0–4 ja teine 45–64 ning kõik teised tunnused on neil võrdsed, siis vanem patsient terveneb $\frac{1}{\exp(-0,2)} = 1,22$ korda kiiremini.



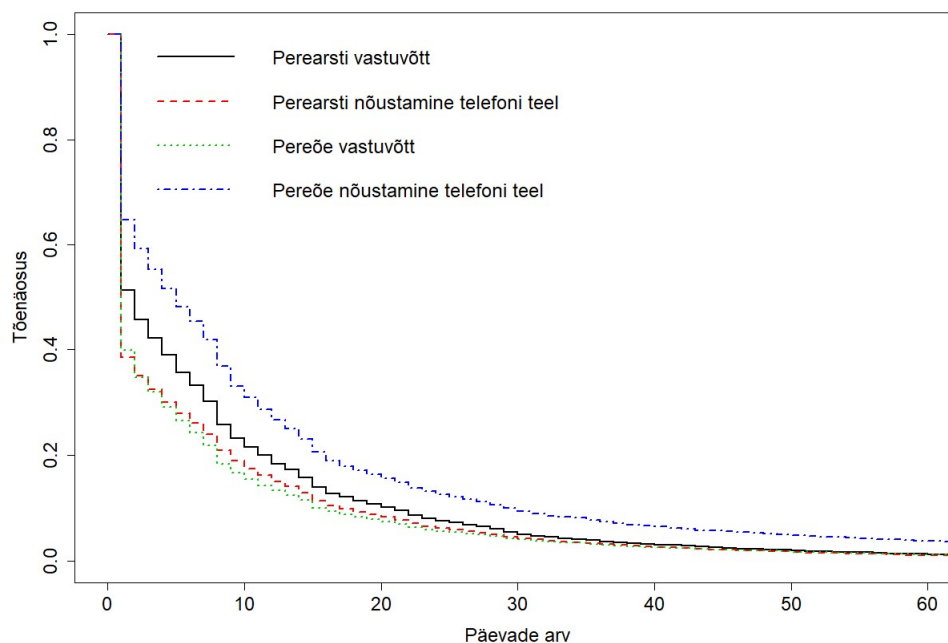
Joonis 9. Haiguskestuse funktsioon erinevate diagnoosirühmade jaoks

Haiguskestus erineb diagnoosirühmades (joonis 9). Kõige pikem haiguskestus on patsientidel, kelle diagnoos kuulub diagnoosirühma S00–T98 (vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed), kuid kõik teised tunnused on võrdsed. Kui võrrelda kaht sarnast patsienti, kellest ühel on diagnoos S00–T98 ja teisel A00–B99 (teatavad nakkus- ja parasiithaigused) rühmast, siis esimese patsiendi haiguskestus on keskmiselt $\exp(0,54) = 1,72$ korda pikem kui teisel. Kui aga võrrelda sarnaseid patsiente diagnoosidega Z00–Z99 (tervise seisundit mõjustavad tegurid ja kontaktid terviseteenistusega) ja A00–B99 rühmast, siis esimene patsiendi haiguskestus on $\frac{1}{\exp(-0,8)} = 2,23$ korda lühem kui teisel.



Joonis 10. Haiguskestuse funktsioon erinevate diagnoosi tüüpide jaoks

Diagnoosi tüüp on veel üks oluline tunnus, mis mõjutab haiguskestust (joonis 10). Kui vaatluse all on kaks patsienti ühest vanuserühmast, diagnoosidega ühest rühmast, sama soo ning esimese visiidi tüübiga, kuid ühel neist on esialgne diagnoos ja teisel esmane, siis esialgse diagnoosiga patsiendi haiguskestus on keskmiselt $\exp(0,22) = 1,25$ korda lühem, kui patsiendil esmase diagnoosi tüübiga. Kui aga teise patsiendi diagnoosi tüüp on korduv, siis on tema haiguskestus $\frac{1}{\exp(-0,12)} = 1,13$ korda lühem, võrreldes esialgse diagnoosiga patsiendist.



Joonis 11. Haiguskestuse funktsioon esimese visiidi erinevate tüüpide jaoks

Viimane mõjutav faktor on esimese visiidi tüüp (joonis 11). Kui võrrelda omavahel sarnast patsienti, kelle esimene visiit on perearsti vastuvõtt (koodiga 9001) selle patsiendiga, kelle esimene visiit on pereõe nõustamine telefoni teel (9064), siis peale pereõe telefonikonsultatsiooni on patsiendi haiguskestus keskmiselt $\exp(0,35) = 1,42$ korda pikem, kui patsiendil, kes pöördus esmalt perearsti vastuvõtule.

Saadud mudeli järgi on haiguskestus kõige lühem meespatsiendil vanuses 45–64, kellel on korduv diagnoos diagnoosirühmast Z00–Z99 (terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega) ja tema esimeseks visiidiks oli perearsti nõustamine telefoni teel. Haiguse prognoositud pikkus on 1,84 päeva. Kõige pikema haiguskestusega on naine vanuses 65–84, kellel on esmane diagnoos rühmast S00–T98 (vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed) ja tema esimene visiit on pereõe nõustamine telefoni teel. Sel juhul kestab haigus 24,96 päeva.

Kasutades treeningandmeid uurime, kui täpne on saadud mudel. Mudeli RMSE on 14,27 ning MAD on võrdne 7,33-ga.

Tabel 15. Prognoosi vea jaotus

Vea suurus (päevades)	<1	1–5	5–9	9–20	20–29	29–40	40–100	>100
Osatähtsus	5,78%	47,78%	33,01%	7,22%	2,72%	1,24%	1,95%	0,30%

Tabelis 15 on näha, et ainult 5,78% juhtudest prognoosib mudel haiguskestuse perioodi täpsusega üks päev. Prognoosi viga 1–5 päeva esineb kõige rohkem – 47,78%-l patsientidest. Prognoosi viga kuni kümme päevani esineb 86,57% juhtudest. See on hea tulemus, kuid reaalsuses on suur erinevus, kas patsient paraneb haigusest ühe või kümne päeva möödudes.

Kasutatud kirjandus

1. Käärrik E. Andmeanalüüs II. Loengukonspekt; 2013.
2. Käärrik M. Elukestusmodelid. Loengukonspekt; 2018.
3. Käärrik M. Üldistatud lineaarsed modelid. Loengukonspekt; 2019.

LISA

Rahvusvahelise haiguste ja tervisega seotud probleemide statistilise klassifikatsiooni 10. versiooni peatükid

Kood	Nimetus
A00–B99	Teatavad nakkus- ja parasiithaigused
C00–D48	Kasvajad
D50–D89	Vere- ja vereloomeelundite haigused ja teatavad immuunmehhanismiga seotud haigusseisundid
E00–E90	Sisesekretsiooni, toitumis- ja ainevahetusehaigused
F00–F99	Psüühika- ja käitumishäired
G00–G99	Närvisüsteemihaigused
H00–H59	Silma- ja silmamanuste haigused
H60–H95	Kõrva- ja nibujätkehaigused
I00–I99	Vereringeelundite haigused
J00–J99	Hingamiselundite haigused
K00–K93	Seedeelundite haigused
L00–L99	Naha- ja nahaaluskoe haigused
M00–M99	Lihaskonna ja sidekoe haigused
N00–N99	Kuse- ja suguelundite haigused
O00–O99	Raseduse ja sünnitusega seotud tüsistused
P00–P96	Sünniperioodis tekkivad teatavad seisundid
Q00–Q99	Kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad
R00–R99	Mujal klassifitseerimata sümptomid, tunnused, kliiniliste ja laboratoorsete leidude hálbed
S00–T98	Vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed
V01–Y98*	Haigestumise ja surma välispõhjused
Z00–Z99	Tervise seisundit mõjustavad tegurid ja kontaktid terviseteenistusega

* Andmestikud puudus V01–Y98 rühm.

Tervisestatistilised andmed ja informatsioon:

- **Tervisestatistika ja terviseuuringute andmebaas**
<http://www.tai.ee/tstua>
- **Tervise Arengu Instituudi tervisestatistika osakonna veebileht**
<http://www.tai.ee/tegevused/tervisestatistika>
- **Andmepäring Tervise Arengu Instituudile**
tai@tai.ee
- **Statistikaameti andmebaas**
<http://www.stat.ee/>
- **Euroopa Liidu statistika**
<http://ec.europa.eu/eurostat>
- **Maaailma Terviseorganisatsiooni (WHO) Euroopa esinduse tervise andmebaas**
<http://data.euro.who.int/hfadbf/>
- **Majandusliku Koostöö ja Arengu Organisatsiooni (OECD) andmebaas**
http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STAT

