TARTU ÜLIKOOL

Loodus- ja täppisteaduste valdkond

Arvutiteaduse instituut

Andmeteaduse õppekava

Sille Habakukk

COVID-19 ennustavate riskimudelite loomine Eesti terviseandmete põhjal

Magistritöö (15 EAP)

Juhendaja: Raivo Kolde, PhD

Tartu 2022

COVID-19 ennustavate riskimudelite loomine Eesti terviseandmete põhjal

Lühikokkuvõte:

COVID-19 pandeemia on pannud tervishoiusüsteemid üle maailma töötama kõrgendatud koormusel. Jätkusuutliku tervishoiuteenuse osutamiseks on tihti vaja teha otsuseid, milliste patsientidega tegeleda eelisjärjekorras. Taoliste meditsiiniliste otsuste tegemisel kasutatakse tihti ennustavaid riskimudeleid, mida COVID-19 jaoks viiruse uudsuse tõttu pandeemia alguses ei eksisteerinud. Esimese haigestumislaine ajal loodud riskimudelid andsid pealtnäha häid tulemusi, kuid olles treenitud vähestel haigusega seotud andmetel pigem ei saavutanud rahuldavaid tulemusi välisel valideerimisel teistel andmestikel. Seetõttu ei saanud neid mudeleid kasutusele võtta ka Eestis.

Selle töö eesmärk oli kasutada elektroonilisi terviseandmeid, et luua riskimudelid, mis ennustaksid hästi COVID-19 haiguskulgu ka Eesti rahvastikul. Riskimudelid treeniti ennustama patsientide haigla- või intensiivravi vajadust või surma 30 päeva jooksul peale COVID-19 nakatumist. Tulemuseks saadud riskimudelid kasutasid juhumetsa algoritmi, mis ei ole riskimudelite loomisel standardpraktika, kuid oli stabiilselt hea COVID-19 raske põdemise eristamisel ja vältis paremini andmetele ülesobitamist. Samas leiti, et mudelite ennustatud tõenäosuste absoluutväärtused vajavad kalibreerimist kui haiguspilt muutub ajas.

Loodud riskimudelid kasutasid arvuliselt paljusid ennustavaid tunnuseid, mistõttu sobiksid peale välise sõltumatu valideerimise rakendamiseks rahvatervise valdkonnas otsuste vastu võtmiseks. Lisaks näidati, et ainult patsiendi eelneva haigusloo põhjal on võimalik ennustada COVID-19 haiguskulgu, ilma patsiendi sümptomeid arvesse võtmata.

Võtmesõnad:

COVID-19, ennustusmudelid, meditsiiniinformaatika

CERCS:B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Creating Prediction Models Based on Estonian COVID-19 Data

Abstract:

The COVID-19 pandemic has strained healthcare systems all around the world. In order to keep up providing quality health care services, it is often necessary to determine which patients to prioritise. Predictive models can inform such medical decisions but did not exist for the novel COVID-19 at the pandemic's beginning. Predictive models created during the first wave of the pandemic reported good predictions but, having been trained on small datasets, did not produce satisfactory results upon external validations on different datasets. Therefore these models also could not be put to use in Estonia.

This thesis aimed to use electronic health records to create prediction models that would predict COVID-19 outcomes well in the Estonian population. Prediction models were trained to predict whether or not a patient would need hospitalisation, be admitted to intensive care or die within 30 days of contracting the virus. The resulting models used the Random Forest algorithm, which is not standard for prediction models, but had stable performance when predicting adverse outcomes of COVID-19 and avoided over-fitting to the data. However the models' absolute value predictions need to be calibrated to account for the disease course changing over time.

The created prediction models used a significant number of predictors, making the models more suitable for use in the public health policy creation process. The models would need to be independently externally validated before use. The prediction models show that good predictive performance is achievable using only registry data, without factoring in any symptoms.

Keywords:

COVID-19, prediction models, medical informatics

CERCS:B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

1 Sissejuhatus

COVID-19 pandeemia on kurnanud haiglasüsteeme paljudes maailma riikides. Haigestunutest võrdlemisi suur osa vajab arstide sekkumist ja haiglaravi. COVID-19 patsientide ravivajaduste tõttu on nii Eestis kui ka mujal peatatud plaanilist ravi ning täheldatud meditsiiniprofessionaalide kurnatust ja läbipõlemist, mis koos vähendavad rahvastikule pakutava tervishoiu teenuse kvaliteeti. Ülekoormuse vältimiseks ja paremaks ressursiplaneerimiseks on vaja teada, milliseid patsiente on kõige kiiremini vaja ravida või eelisjärjekorras vaktsineerida. Selles otsustusprotsessis saab kasutada riskimudeleid [2]. Riskimudelid on ennustavad mudelid, millega on võimalik igale patsiendile määrata haigestumise puhul komplikatsioonide tekkimise tõenäosused.

Riskimudelite loomisega tegelevad näiteks terviseinformaatikud, kellest osad on moodustanud *Observational Health Data Sciences and Informatics* ehk OHDSI (hääldatakse *Odyssey*) kogukonna. OHDSI kogukond arendas pandeemia alguses välja riskimudelid [23], mille headust Eesti COVID-19 haigete andmestikul valideeris oma bakalaureusetöös Marc David [5]. Davidi töö leidis, et need kiiresti loodud riskimudelid on osutunud ebatäpseteks. Põhjuseid selleks võib spekuleerida mitu:

- esialgse andmete vähesuse kompenseerimiseks kasutati COVID-19 lähendina gripiandmeid,
- muteerunud viiruse haiguspilt erineb esialgsest,
- vaktsineeritute haigusteekond võib olla teistsugune,
- populatsioonid, millel treeniti mudelid, erinevad populatsioonist, millel mudel valideeriti,
- kasutatud matemaatilised mudelid polnud parimad.

Pandeemia algusest on möödas juba kaks aastat ja haigestunute kohta Eestis on nüüd rohkem andmeid saadaval. Tartu Ülikooli COVID-19 teemalise teadusprojekti Coriva

projekti raames tehtud uuring [9] leidis seoseid mitmete eelnevalt esinenud diagnooside ja COVID-19 haiguse raskusastme vahel. Sellele järgneb loomulik küsimus, kas eelnevalt esinenud diagnoosid või muud terviseandmed sisaldavad piisavalt infot ka ennustava mudeli jaoks. Seetõttu sai Coriva üheks eesmärgiks treenida Eesti terviseandmete põhjal riskimudeleid, mis ennustaks COVID-19 haigete haigla- ja intensiivravi vajaduse ning surma tõenäosuseid [17]. Selle eesmärgi täitmise protsessist ja tulemustest on kokku pandud käesolev magistritöö.

Käesoleva töö peatükis „Taust“ kirjeldatakse täpsemalt riskimudelite tausta ja komponente, seletatakse lahti riskimudelite headuse hindamise võimalusi, kirjeldatakse terviseandmete olemust ning antakse lühiülevaade varasematest COVID-19 riskimudelitest. Peatükis „Uuring“ tuuakse samm-samm haaval välja uuringu läbimiseks tehtud otsused ja lõpliku riskimudeli tulemused. Jaotises „Lisad“ on välja toodud ennustavate tunnuste valiku sammu seadistused ja tulemused ning lihtlitsents lõputöö reprodutseerimiseks.

2 Taust

Siin peatükis antakse ülevaade riskimudelite teoreetilisest poolest, COVID-19 terviseandmetest ja matemaatiliste mudelite headuse hindamisest.

2.1 Riskimudelid

Ennustused on tervishoiuvaldkonnas olulisel kohal. Ennustusi kasutatakse ravi määramisel ja kõrvalmõjude vältimiseks, riskirühmade väljaselgitamiseks ennetusmeetodite rakendamiseks, kliiniliste uuringute planeerimisel [16]. Objektiivsete ennustuste tegemiseks saab kasutada ennustusmudeleid ehk riskimudeleid, mis on matemaatilised mudelid, mis võtavad arvesse patsiendi sotsiaal-demograafilisi andmeid ja haigus- ning raviajalugu.

Nagu teistegi kliiniliste mudelitega, on riskimudeli koostamiseks vaja defineerida populatsioon, kellel mudelit plaanitakse rakendada, valim ehk milliste inimeste andmetel mudel treenitakse, tulem ehk mis sündmust ennustatakse ning mis ajaperioodil inimesi jälgitakse.

Riskimudeli erinevus teistest kliinilistest mudelitest seisneb eesmärgis. Kui üldiselt kliinilised uuringud otsivad põhjuslike seoseid, siis riskimudelid peavad andma võimalikult täpseid ennustusi. Seetõttu võivad riskimudeli ennustavad tunnused olla vaid korreleeruvad, mitte põhjuslikud [10].

Riskimudelid annavad hinnangu riskile ehk tõenäosuse tulemi tekkimisele protsentuaalselt vahemikus 0-100%. Binaarse tulemi (haigestub/ei haigestu, vajab ravi/ei vaja ravi) korral, määrates kindlaks riskilävendi (*risk threshold*), saab teha üldistuse, et inimestel, kelle risk on suurem kui määratud lävend, tekib tulem (positiivne tulem) ja inimestel, kelle risk on alla määratud lävendi, tulemit ei teki (negatiivne tulem). Kuna ideaalseid mudeleid leidub harva, siis tuleb riskilävendi valimisel kaaluda potentsiaalset kasu ja kulu, mida valepositiivsed või valenegatiivsed hinnangud võivad kaasa tuua [7]. Kulu võib

siinkohal tähendada materiaalset kulu rahas või ressurssides kui ka mittemateriaalseid tagajärgi, nagu kõrvaltoimete esinemist, elukvaliteedi langemist ravi viibimisest tekkinud tüsistuste tõttu jms.

2.1.1 Riskimudelite näited

Mitmed riskimudelid on meditsiinis juba pikalt kasutuses.

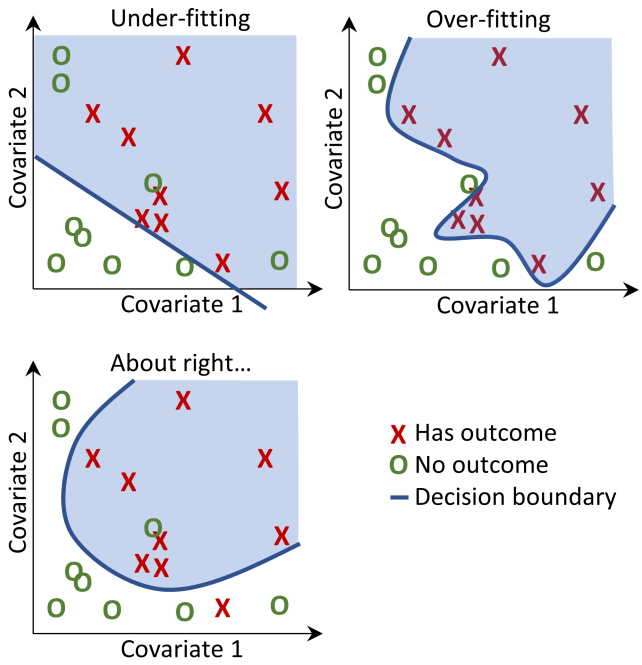
Neist võib-olla tunduim on Apgari skoor [8], mida arvutatakse vastsündinutel minuti ja viie minuti möödumisel sünnist. Apgari skoori kovariaatideks on hinnangud skaalal 0,1,2 hingamisele, südametööle, lihastoonusele, refleksidele ja nahavärvile. Tulem, mida ennustatakse, on vajadus beebile hingamisabi või südamemassaaži andmiseks. Apgari skoor alla 7 näitab, et beebi vajab meditsiinilist sekkumist.

Veel on tuntud Glasgow kooma skaala [18], mis hindab inimese koomaseisundi kestvust ja tugevust, kasutades tunnustena inimese kõnevõimet, silmade liigutamist ja motoorseid liigutusi. Charlsoni komorbiidsusindeks ennustab 1-aasta surmariski, võttes arvesse 19 erineva haiguse esinemist ning kaasumist [3].

2.1.2 Matemaatilised mudelid

Riskimudeli loomisel soovitakse leida seoseid ennustavate tunnuste ja tulemi vahel. Riskimudeli üks osa on matemaatiline mudel, mille abil leitakse otsustuspiir, mille põhjal mudel jaotab patsiendid juhtumiteks ja mitte-juhtumiteks. Kui eeldada, et ennustavaid (arvulisi, pidevaid) tunnuseid on kaks, saab patsiendid asetada kahemõõtmelisele joonisele 1, kus x-teljel on ühe, ning y-teljel teise tunnuse väärtus. Matemaatilisest mudelist sõltub kui keeruline otsustuspiir saab olla. Liiga lihtsa otsustuspiiri korral jaotatakse liiga palju patsiente valesti, mida kutsutakse alasobitamiseks. Väga keeruka otsustuspiiri korral võib mudel liiga täpselt järgida andmestiku eripärasid, mida kutsutakse ülesobitamiseks. Matemaatilise mudeli valikul tuleks leida kesktee ala- ja ülesobitamise vahel.

Matemaatilisi mudeleid on palju erinevaid, selles töös vaadati lähemalt lasso logistilist

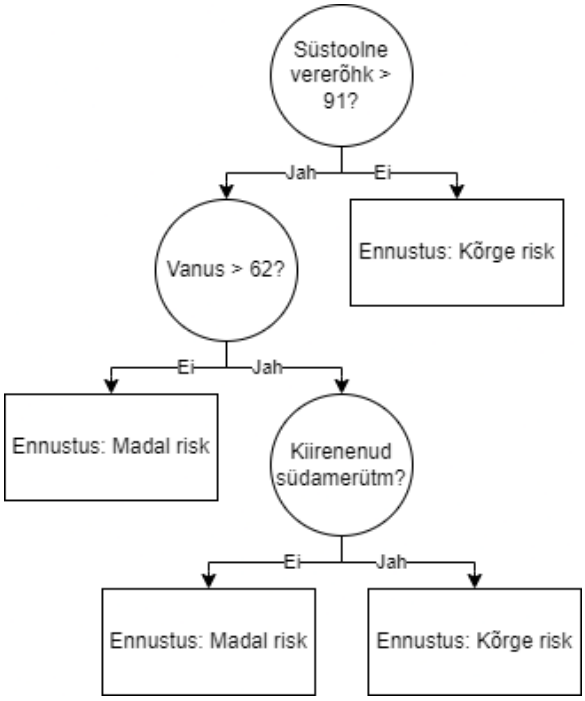


Joonis 1. Joonistel [10] on toodud alasobitatud (*under-fitting*), ülesobitatud (*over-fitting*) ja hästi sobitatud riskimudelitele vastavad otsustuspiirid.

regressiooni (*LassoLogisticRegression*), juhumetsa (*RandomForest*), gradientvõimendust (*GradientBoostingMachine*) ja kohastuvvõimendust (*AdaBoost*).

Eelnevas alapeatükis väljatoodud riskimudelid ja ka paljud teised, mida praktikas kasutatakse on üpris lihtsad, kuna arendati välja ajastul, mil arvutid polnud nii levinud ja arvutusvõime oli madalam. Näiteks toodud riskiskoorid kasutavad regressioonmudeleid. Kõige laialdasemalt kasutataksegi riskimudelites logistilisi regressioonmudeleid, kuna need võimaldavad kaasata nii kategoorilisi kui arvulisi tunnuseid, mitte-lineaarseid teisendusi ja ka tunnuste koosmõjusid [16]. Logistilise regressiooni tulemusi on ka suhteliselt lihtne tõlgendada, mille tõttu kasutatakse neid ka juhtudel, kui keerukamad mudelid oleks võib-olla sobivamad [7]. Logistilist regressiooni kombineeritakse tihti lasso (*least absolute shrinkage and selection operator*) reguleerimisega, mis võimaldab automaatselt mudelist välja jätta ebaolulisi ennustavaid tunnuseid [10].

Samas personaalmeditsiini populaarsuse kasvu tõttu on kasvanud ka nõudlus mude-



Joonis 2. Lihtne otsustuspuu, mis ennustab patsiendi surmariski. [1]

lite järgi, mis võtaksid arvesse rohkem individuaalsust [7], mille jaoks tuleks kasutada rohkem andmeid ja võib-olla ka keerukamaid matemaatilisi mudeleid. Keerukamad matemaatilised mudelid on tihti kombineeritud lihtsamatest mudelitest, näiteks otsustuspuudest [10]. Otsustuspuud näevad välja nagu voodiagrammid (joonis 2). Otsustuspuid iseloomustab nende sügavus (*depth*), mis näitab mitu korda puu hargneb, joonisel 2 on puu sügavuseks kolm.

Juhumetsad, gradientvõimendus ja kohastuvvõimendus kõik kombineerivad otsustuspuid, kuid veidi erinevalt. Juhumetsas on palju otsustuspuid, kuid piiratakse palju ennustavaid tunnuseid igas otsustuspuus võib kasutada ning erinevates puudes kasutatakse erinevaid tunnuseid [10]. Gradientvõimenduse korral lisatakse otsustuspuid iteratiivselt ja iga iteratsiooni korral pannakse enam rõhku eelmistes iteratsioonides valesti klassifitseeritud patsientidele [10]. Kohastuvvõimendus on sarnane gradientvõimendusele, kuid kasutab otsustuspuid, mille sügavus on tavaliselt 1 [10, 11]. Kui palju otsustus-

puid on igas eeltoodud matemaatilises mudelis on määratav treenimise hetkel mudelite hüperparameetrite kaudu.

Piirangud otsustuspuude sügavustele ja arvule ning ennustatavate tunnuste arvule (nii otsustuspuudes kui ka logistilises regressioonis) on olulised ülesobitamise ennetamiseks [10].

2.1.3 Riskimudelite rakendamine

Riskimudelitel on rakendusi nii rahvatervise, kliinilise praktika kui ka teadusliku uuringute valdkondades. Paljud riskimudelid, mida tervishoiusüsteemis kasutatakse, nõuavad arstidelt nö käsitsi arvutamist, tänapäeval tihiti ka Exceli tabelite kasutamist [16]. Samas rahvatervise otsuste juures kasutatakse terviseandmeid agregeeritult, mis on tihti vähemalt osaliselt automatiseeritud protsess.

Riskimudeli tulevasest kasutusalast sõltub seega ka kui keeruline riskimudel olla võib. Selgelt pole kirurgil võimalik sisestada sadu või tuhandeid patsiendi terviseandmeid rakendusse, et aegkriitilisel hetkel riskimudeli ennustust saada. Seega riskimudelid, mida kasutavad arstid eelkõige patsiendivisiidi ajal peavad olema suhteliselt lihtsad - mõistlikult piiratud arvu ennustavate tunnustega ja kergesti arvutatav.

Seevastu, kui tehakse otsuseid rahvatervise poliitika üle, näiteks sõeluuringute korraldamisel või vaktsiini eelisjärjekorras saajate valimisel, on kasutusel registriandmed rahvastiku kohta. Seega kui riskimudelid on arendatud eesmärgiga informeerida ametnike rahvatervise otsuste tegemiseks, siis peaks need mudelid arvestama ainult andmeid, mis on registrites kergesti kättesaadaval. Kui riskimudel rakendada registripäringute osana, siis on võimalik mudeli ennustustes kasutada ka kordades suuremat arvu tunnuseid, kuna kogu arvutusprotsessi teeb läbi masin, mitte inimene.

2.2 Mudelite headuse hindamine

Mudeli headuse hindamisel vaadatakse kõigepealt, kui täpselt on mudel saanud tegelikkuse ennustamisega hakkama. Seda iseloomustab segadusmaatriks [12], mille kuju on välja toodud tabelis 1. Mudeli ennustuste jagunemine tõesteks positiivseteks/negatiivseteks ning valepositiivseteks ja -negatiivseteks sõltub valitud riskilävendist [16].

		Mudeli ennustus	
		Tulem esineb	Ei esine
Tegelikkus	Tulem esineb	Tõene positiivne (TP)	Valenegatiivne (VN)
	Ei esine	Valepositiivne (VP)	Tõene negatiivne (TN)

Tabel 1. Segadusmaatriks toob välja binaarse tulemi korral, millised mudeli ennustused olid täpsed ja millised mitte.

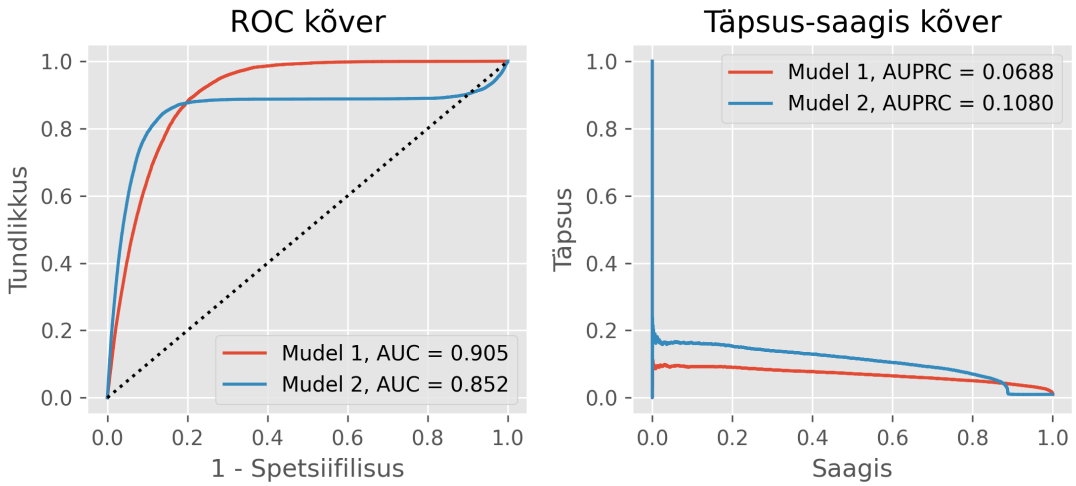
2.2.1 Diskrimineerimine

Riskimudeli täpsust tulemi esinemise ennustamisel nimetatakse diskrimineerimisvõimeks [7]. Binaarse tulemi korral on üheks tuntuimaks mudeli diskrimineerimisvõime näitajaks ROC (*receiver operating characteristic*) kõvera alune pindala ehk AUC (*area under the curve*). ROC kõver on tundlikkuse (*sensitivity*, valem (1)) ja spetsiifilisuse (*specificity*, valem (2)) vahelise suhte joonis. AUC kirjeldab tõenäosust, et tulemiga patsiendile ennustatakse suurem risk, kui juhuslikule patsiendile, kellel tulemit ei esine [16]. Täiesti juhusliku mudeli AUC väärtus on järelikult 0.5 ja ideaalse mudeli AUC on 1.

$$tundlikkus = \frac{TP}{TP + VN} \qquad (1)$$

$$spetsiifilisus = \frac{TN}{TN + VP} \qquad (2)$$

Kuigi AUC ja ROC-kõvera joonised on klassifitseerivate mudelite kirjeldamisel laialdaselt kasutusel, siis on välja toodud, et kallutatud andmetel võib AUC anda väga häid tulemusi, omamata tegelikkuses head ennustusvõimet [6]. Kallutatud andmed on

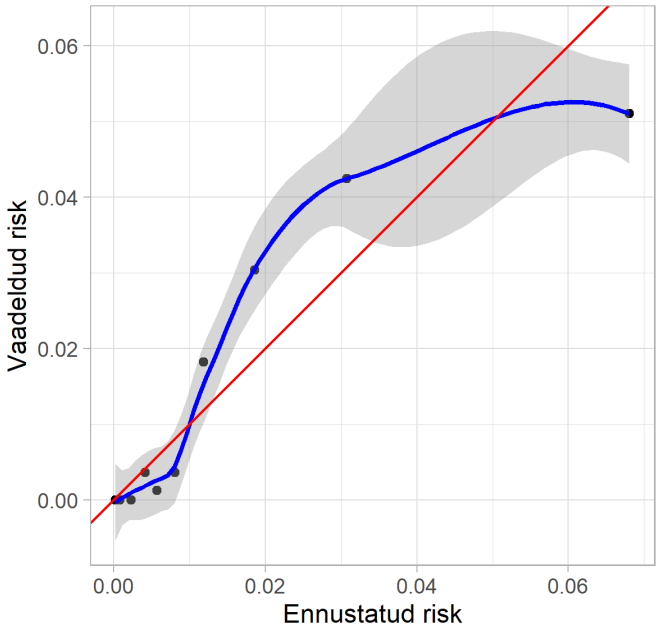


Joonis 3. Simuleeritud kallutatud andmetel näide ROC ja täpsus-saagis kõverast. [12]

sellised, kus ennustatavat tulemit esineb oluliselt vähem kui pooltel juhtudel andmestikus. Siis soovitatakse peale AUC vaadata täpsuse (*precision*, valem (3)) ja saagise (*recall*, valem (4)) vahelise suhte joonise (PR kõver ehk *precision-recall curve*) alust pindala ehk AUPRC (*area under precision-recall curve*) [6]. AUPRC on vähem mõjutatud kallutatud andmetest, kuna selle arvutamisel ei kasutata tõeselt negatiivseid tulemusi. Täiesti juhusliku mudeli AUPRC väärtus on tulemi esinemise suhtarv andmestikus ja ideaalse mudeli AUPRC on 1. Seetõttu on AUPRC tõlgendamise jaoks oluline analüüsis välja tuua AUPRC baasväärtus, kuna see sõltub konkreetsest andmestikust.

$$täpsus = \frac{TP}{TP + VP} \qquad (3) \qquad\qquad saagis = \frac{TP}{TP + VN} \qquad (4)$$

Saagist AUPRCs ja tundlikkust AUCs arvutatakse sama valemiga, seda suurust nimetatakse ka veel tõeselt positiivsete suhteks (TPR, *True Positive Rate*). ROC ja täpsus-saagis kõveratel on esimesel juhul TPR x-teljel ning teisel juhul y-teljel, näitekõverad on toodud joonisel 3.



Joonis 4. Sinine joon on Loessi algoritmiga silutud [16] kalibratsioonikõver, punane joon näitab ideaalset kalibratsiooni.

2.2.2 Kalibratsioon

Mudeli sobivust andmestikule on võimalik kirjeldada kalibratsiooniga. Kalibratsioon näitab kui hästi ennustatud riskid on vastavuses vaadeldud riskidele [10]. Kui mudel ennustab 70% patsiendile haigust, siis kalibreeritud mudeli korral peaks tegelikkuses 70 patsiendil 100st haigus esinema.

Parim moodus kalibreerimist analüüsida on läbi kalibratsioonikõvera ehk joonise, mille x-teljele kantakse kvantiilide kaupa grupeeritult andmepunktide keskmine ennustatud risk ning y-teljele kantakse samade grupeeritud punktide vaadeldud tulemi keskmine esinemistõenäosus [7]. Ideaalselt kalibreeritud mudeli kalibratsioonikõvera punktid asetsevad $x = y$ joonel.

Kalibratsioonikõvera näide koos usaldusvahemikega on toodud joonisel 4. Punktid, mis asetsevad $x = y$ joonest kõrgemal, näitavad, et nendele punktidele ennustatakse

madalamat riski, kui tegelikkuses vaadeldi. Samamoodi punktid, mis asetsevad $x = y$ joone all, näitavad, et ennustatud risk oli kõrgem, kui tegelikkuses.

2.2.3 Valideerimine

Riskimudelid annavad parimaid tulemusi oma treeningandmestikul. Riskimudelist on aga kasu ainult siis kui see annab häid ennustusi uutel patsientidel ehk kui mudel on hästi üldistuv [16]. Mudeli üldistuvuse hindamist nimetatakse valideerimiseks. Valideerimine näitab mudeli töökindlust ja rakendatavust praktilistes olukordades. Valideerimist on laias laastus kahte tüüpi - sisemine ja väline valideerimine.

Sisemist valideerimist viiakse läbi samas populatsioonis, millel mudelit treenitakse. Üheks enam-levinud meetodiks sisemiseks valideerimiseks on eraldatud valimi kasutamine. Sellisel juhul eraldatakse enne mudeli treenimist juhuslikult osa andmestikust, mida mudelile treenimise käigus ei näidata. Hiljem saab nende nägemata andmete pealt anda hinnang mudeli üldistatavusele sama populatsiooni jaoks [7]. Eraldatud valimi kasutamisel on probleemiks vähelevinud tunnused, mis juhuslikkuse tõttu võivad jaotuda treening- ja valideerimishulkadesse ebaproportsionaalselt. Samuti on eraldatud valimi kasutamine keeruline kui kogu treeningandmestik on väike, kuna siis jääb puudu andmeid hea mudeli treenimiseks kui ka õiglase valideerimistulemuse saamiseks [10]. Nende probleemidele pakub lahendust teine sisemise valideerimise meetod - k-korda ristvalideerimine. Ristvalideerimise jaoks jagatakse andmestik juhuslikult k võrdseks osaks. Üks osavalim jäetakse kõrvale valideerimiseks, ja ülejäänud $k - 1$ osavalimil treenitakse mudel. Seda protsessi korratakse iga osavalimiga, kokku valideeritakse mudelit k korda [16]. Mudeli valideerimismõõdikud arvutatakse ristvalideerimise kõigi valideerimiste tulemuste keskmisena.

Välist valideerimist viiakse läbi mingis mõttes erinevas, kuid treeningpopulatsiooniga sarnases populatsioonis [7]. Väline valideerimine on eriti oluline kui tahta arendatud riskimudelit rakendada meditsiinis [10]. Väline valideerimine saab olla näiteks ajali-

ne või geograafiline. Ajalise valideerimise korral on valideerimisandmestik kogutud erineval ajaperioodil võrreldes treeningandmestikuga, tavaliselt koosneb hiljutisemalt ravitud patsientidest [16]. Ajaline valideerimine sarnaneb eraldatud valimiga sisemise valideerimisega, kuid erineb valimi eraldamise põhimõte poolest. Kuna ajaga võivad muutuda nii ravimeetodid kui andmete kogumise põhimõtted ka samas andmestikus, siis peetakse ajalist valideerimist ikkagi väliseks valideerimiseks [16]. Geograafiline valideerimine tähendab mudeli valideerimist andmestikul, mis on kogutud teisel geograafilisel alal, näiteks teise haigla, riigi või hoopis maailmajao patsientide seast [7]. Geograafilise valideerimise läbiviimiseks on vaja koostööd mitmete uuringugruppide vahel, kuna tihti jääb ühele uuringugrupile saadaolevatest andmetest väheks usaldusväärsete tulemuste saamiseks [16]. Kõige suuremat usaldust mudeli töökindluse vastu loob erapooletu valideerimine, mille viivad läbi mudeli arendajatest täiesti eraldiseisvad inimesed andmestikul, millele mudeli arendajatel polnud ligipääsu [16].

2.3 Tarkvara riskimudelite loomiseks

Terve ülevaade selles alapeatükis on refereeritud *The Book of OHDSI* raamatust [10].

OHDSI kogukond tegutseb selle nimel, et parandada maailma inimeste tervist läbi rahvusvahelise koostöö. Oma eesmärgi täitmiseks OHDSI liikmed kaardistavad hea tava terviseandmete uurimisel, arendavad vabavaralist analüütilist tarkvara, mis järgiks neid häid tavasid, ning rakendavad arendatud tarkvara, et leida vastuseid kliiniliselt olulistele küsimustele.

Terviseandmeid kogutakse üha suuremas kogustes üle maailma, saavutades suurandmete mahu. Terviseandmeid kogutakse rahvastikuregistrites, ravisüsteemides, kindlustustes ning tihti täiesti unikaalse vormistusega. Põhjalike uuringute jaoks on neid andmestike vaja kombineerida, mis tulenevalt nende vormistamisele nõuab iga uuringu puhul suurt tööd andmete puhastamisel. Ühtne terviseandmete ülesmärkimise standard aitaks muuta terviseuuringuid lihtsamaks ja soodustaks ka koostööd uurimisgruppide vahel. OHDSI

loodud ühtne andmemudel ehk CDM (*common data model*) proovib lahendada seda probleemi. CDM on kõigi OHDSI arendatud tarkvaralahenduste alustala, kuna nende lahenduste kasutamiseks peab iga uurija oma andmestiku viima CDM kujule.

OHDSI loodud tööriist ATLAS on veebipõhine graafiline kasutajaliides standardiseeritud uuringute disainimiseks. ATLAS lihtsustab uuringute jooksutamist, võimaldades kasutajal programmikoodi kirjutamata genereerida andmestiku kirjelduse ja koosseisu ning defineerida kõik vajalikud uuringu osad alates valimitest, lõpetades uuringu peenhäälestusest. Riskimudelite loomiseks sisaldab ATLAS patsiendipõhiste ennustuste ehk PLP (*patient level prediction*) paketti. PLP pakett genereerib kasutaja valikute põhjal vajalikud andmebaasipäringud, R ja Python koodid terve riskimudeli treenimise ja valideerimise protseduur läbi viimiseks.

2.4 COVID-19 ja andmed

COVID-19 haigust põhjustab SARS-CoV-2 viirus, mis levib enamikel juhtudel piisknakkusena. COVID-19 sümptomid sarnanevad gripi sümptomitega, avaldudes erineva raskusastmetega. Samas on koroonaviirus nakkavam ja kõrgema suremusega kui gripp [19]. Nagu teisedki viirushaigused, on COVID-19 aastate jooksul muteerunud, mis väljendub vahel muutustes nakkuvuses ja/või haigestumise raskuses [13]. Sageli räägitakse ka haigestumise lainetest ehk ajaperioodidest kui haigestunute arv päevas on kõrge. Kuna haigestumise lainete tekkimine sõltub kehtivatest piirangutest, viirusetüvedest, rahvakäitumisest, möödudes seega riigiti või piirkonniti erinevatel aegadel, siis ei ole ühtselt kindlaks määratud haigestumislainete piirkuupäevi. Selle töö raames käsitletakse esimest haigestumislainet kui vahemikku veebruar kuni august 2020 kaasa arvatud. Teine laine arvestati september 2020 kuni juuni 2021 kaasa arvatud. Kolmas laine arvestati juuli kuni november 2021.

COVID-19 haiguse täpsemaks uurimiseks Eesti populatsioonis väljastas Haigekassa Coriva projekti raames vahemikus veebruar 2020 kuni veebruar 2021 kõigi positiivse

SARS-CoV-2 testi või COVID-19 diagnoosi saanud isikute terviseandmed. Lisaks leiti iga sellise isiku kohta 4 kontrollisikut. Kontrollisikuid ei sobitatud vanuse, soo ega muude tunnuste poolest. Kokku sai andmestikku üle 380 tuhande isiku terviseandmed, mis on hinnanguliselt 29% kogu elanikkonnast [15].

Lõplikus andmestikus on COVID-19 diagnoosiga isikuid rohkem kui esialgselt planeeritud 1:4 suhtes, sest vahemikus veebruar 2021 kuni november 2021 haigestus ka kontrollisikuid. Täpsemalt on lõplikus andmestikus 128 200 COVID-19 diagnoosiga patsienti. Terviseandmetest väljastati patsientide raviarved ja väljakirjutatud retseptid vahemikust veebruar 2017 kuni november 2021. Surmaregistrist päriti juurde andmed valimisse sattunud isikute surmade kohta.

Raviarvetel on patsiendi üldandmed nagu sugu ja sünniaasta ning ka andmed patsiendile osutatud terviseteenuste ja -protseduuride kohta. Veel on raviarvetele märgitud diagnoosid, arstivisiitide kuupäevad ning visiiditüübid.

Retseptide andmetes on välja toodud patsiendi diagnoos, kuni kolm ravimi toimeainet, müügiks ettenähtud kogus ja ka ravimi tarvitamise juhised.

Eestis pole terviseandmete hoiustamine standardiseeritud CDM kujul, millel oleks võimalik peatükis „Tarkvara riskimudelite loomiseks“ väljatoodud tööriistu kohe rakendada. Lisaks erinevale andmestruktuurile kasutavad Eesti andmestikud CDMist erinevaid kodeeringuid diagnoosi-, ravimi-, teenus- ja protseduurikoodide jaoks. Tarkvara Eesti terviseandmete teisendamiseks CDM vormingusse on aastaid arendanud personaalmeditsiini osakond firmas STACC koostöös Tartu Ülikooli terviseinformaatika töörühmaga. Selle tarkvara edasiarendusega tuli autoril ka käesoleva töö raames tegeleda, kuna Coriva projekti käigus saadud andmekoosseis ei võimaldanud tarkvara üks-ühele rakendamist.

2.5 Varasemad COVID-19 riskimudelid

Kohe COVID-19 pandeemia algusest hakkasid teadlased üle maailma tööle, et luua riskimudeleid, mis ennustaks haiguse levimust ja prognoose. Juba märtsis 2020 avaldatus

metauuringus [24] analüüsiti 232 avaldatud COVID-19 seotud riskimudelit. Uuritud riskimudelite kasutatavust hinnati madalaks, eelkõige ebaselge või kallutatud andmete kogumise, kehvasti dokumenteeritud arendusprotsessi ja/või riskimudelite ülesobitamise tõttu. Ainult üks haiguskulgu ennustav mudel osutus Wyants jt [24] analüüsis paljulubavaks, kuid vajas nende kinnitusel edasist valideerimist.

Mainitud uuringu viisid läbi Clift jt [4], saades tulemuseks kaks riskimudelit. Üks riskimudel ennustas kui kaua peale COVID-19 nakatumist patsient komplikatsioonidesse sureb, teine ennustas kui kaua peale nakatumist patsient vajab haiglaravi. See uuring viidi läbi ainult esimese laine haigestunute andmetega Suurbritannias. Ka antud töö autorid soovitasid pandeemia olukorra muutuste korral mudelit üle treenida, viidates viiruse mutatsioonidest ja riiklikest piirangutest tulenevale vajadusele mudeli kalibratsiooni parandada. Kahjuks Clift jt [4] loodud mudeli rakendamine Eestis pole võimalik, kuna nende mudel kasutab konkreetselt Suurbritannia tervishoiusüsteemile omaseid tunnuseid, sealhulgas ka näitaks kodust postiindeksit.

Ka OHDSI kogukonna teadlased, eesotsas Williams jt, koostasid COVID-19 ennustavad riskimudelid (COVER) [23]. Nende riskimudelite ennustatavad tulemid olid COVID-19 haige sattumine kopsupõletikuga haigla- ja intensiivravile ning COVID-19 nakatumisest põhjustatud surm. Williams jt lahendasid COVID-19 andmestike vähesuse probleemi, treenides oma mudelid gripihaigete andmetel, jättes COVID-19 haigete andmed valideerimiseks. COVER mudelite töökindlust Eesti kontekstis uuris David [5] oma bakalaureuse töös ning leidis mitmeid murekohti nende praktilises rakendatavuses. Üheks puuduseks toodi välja puuduvate andmete probleem, kui mudeli treenimisel kasutatud andmestikus esines andmeid, mida Eesti andmestikus ei eksisteeri, siis saadud mudelid Eesti andmestikul ei saa hästi toimida.

Mõlemad välja toodud uuringud kasutasid ainult regressioonmudeleid. Williams jt kasutasid Lasso logistilist regressiooni [23], iga mudel sisaldas 7 kuni 521 ennustavat tunnust. Clift jt ei täpsustatud, mis algoritmi nad mudeli treenimisel kasutasid, peale

selle, et nad arvutasid regressioonikordajaid [4] välja toodud 41 ennustavale tunnusele.

3 Uuring

Tulenevalt COVID-19 suurest koormusest tervishoiusüsteemidele, eriti haiglatele ja nende võimekusele jätkata plaanilise raviga [21], said uuringusse valitud konkreetsed küsimused, mis mõjutavad enim haiglakohtade hõivatust:

Haiglaravi Millised patsiendid satuvad suurema tõenäosusega COVID-19 diagnoosi järgselt haiglaravile?

Intensiivravi Millised patsiendid vajavad suurema tõenäosusega COVID-19 diagnoosi järgselt intensiivravi?

Surm Millised patsiendid surevad suurema tõenäosusega COVID-19 diagnoosi järgselt?

3.1 Uuringu ülesehitus

Riskimudeli treenimiseks tuleb teha mitmeid valikuid. Mõned otsused saab ja tuleb teha enne uuringu alustamist - defineerida uurimisküsimused, valim, tulem. Osad valikud samas nõuavad empiirilist lähenemist, näiteks milline masinõppe algoritm sobib andmetega kõige rohkem, millised ennustavad tunnused ehk kovariaadid on piisavad ja tarvilikud, et saavutada parimad tulemused.

Teekond lõpliku mudelini nägi välja järgnev:

1. Defineeriti treeningvalim ning tulemid vastavalt uuringu eesmärkidele.
2. Valiti välja masinõppe algoritm, mis andis parimad tulemused valideerimisel.
3. Parimale algoritmile leiti kõige sobilikum ennustavate tunnuste komplekt.
4. Treeniti lõplik mudel.

3.1.1 Valim

Püstitatud uurimisküsimuse järgi sai uuringu valimiks valitud Eesti haigekassa süsteemis osalevad patsiendid, kellel diagnoositi nakatumine COVID-19 viirusesse. COVID-19 nakatumist tuvastati kas raviarvel oleva diagnoosi või positiivse PCR testi esinemise põhjal.

Juba teisel nädalal pärast esimest vaktsiinidoosi leiduvad antikehad kaks korda enamal patsiendil kui enne vaktsineerimist [22]. Mudeli loomisel taheti simuleerida pandeemia alguse olukorda, kus COVID-19 haigus oli veel uus ja riskimudelid kõige vajalikumad. Seetõttu valiti treeningvalim võimalikult sarnane vaktsiini leiutamise eelse populatsiooniga, arvates vaktsineerimiskuuri alustanud patsiendid treeningvalimist välja enne seda kui nende vaktsineerimiskuur oli lõpuni viidud. Täpsemalt, selle uuringu jaoks loeti patsient vaktsineerituks, kui vähemalt üks vaktsiinidoos oli tehtud rohkem kui 14 päeva enne haigestumist.

3.1.2 Tulemid

Vastavalt kolmele uurimisküsimusele defineeriti ka kolm tulemit: haiglaravivajadus, intensiivravivajadus ja surm. Kuna ennustusmudelid kasutavad kõiki korreleeritud, mitte ainult põhjuslike seoseid, siis on oluline tulemite defineerimisel pöörata tähelepanu ajavahemikele sündmuste vahel, ehk kui suur on vahe tulemi realiseerumise ja COVID-19 haigestumise vahel. Kui haigestumise ja tulemi vaheks lubada liiga pikk periood, siis saadud mudeli põhjal järelduste tegemisel ei oleks enam selge, kas tulemit mõjutas haigestumine või mõni muu vahepeal toimunud sündmus. Siin uuringuks valisime COVID-19 haigestumise ja tulemi ajavahemikuks maksimaalselt 30 päeva.

Haiglaravi Patsiendil loeti olevat tulem „Haiglaravi“ kui tal esines 30 päeva jooksul pärast valimisse sattumist raviarve, mille tüüp oli „Haiglaravi“.

Intensiivravi Patsiendil loeti olevat tulem „Intensiivravi“ kui tal esines 30 päeva jooksul

pärast valimisse sattumist raviarve, mille tüüp oli „Intensiivravi“.

Surm Patsiendil loeti olevat tulem „Surm“ kui tema kohta oli 30 päeva jooksul pärast valimisse sattumist registreeritud surmateatis.

Rasedaid testiti COVID-19 suhtes enne sünnitust [20]. Seetõttu esineb valimis märkimisväärne osa patsiente, kes sattusid COVID-19 diagnoosiga haiglasse, kuid mitte COVID-19 ravivajaduse tõttu. Haiglasse sattumine sünnituse tõttu ei olnud selle uuringu jaoks huvipakkuv tulem. Ka intensiivravi vajadus ning suremus on rasedate ja sünnitajate seas isegi mitte-pandeemia olukorras muust populatsioonist kõrgem. Seetõttu arvati kõigist tulemitest välja patsiendid, kellel 90 päeva ehk trimestri jooksul enne valimisse sattumist esines raviarvetel mõni raseduse või sünnitusega seotud diagnoos või protseduur.

3.1.3 Valideerimine

Mudelite iteratiivsel treenimisel kasutati sisemiseks valideerimiseks ristvalideerimist.

Terviseandmete tundlikkuse tõttu ei leidu autorile teadaolevalt avalikke andmebaase COVID-19 haigete kohta, mis sisaldaks piisava detailset terviseinfot, et õpitud mudeleid väliselt valideerida sõltumatul andmestikul. Samas oli Coriva andmestik oma populatsiooni suure hõlmatuse ja COVID-19 ajas muutuva olemuse tõttu võimalik jaotada ajaliseks valideerimiseks, mis iseloomustab kui hästi treenitud mudelid üldistuvad ajas. Teadaolevalt muteeruvad viirused kiiresti, mistõttu ka haigestumise ning erinevate tulemite esinemise sagedus muutub.

Esimese haigestumise laine ajal oli testimine korraldatud väga piiravalt. Seetõttu ei ole teadaolevaid esimese laine haigestunuid sisaldav valim hea üldistus kõigile esimese laine ajal haigestunute populatsioonile, kuna andmestikust on välja jäänud ebaproportsionaalselt palju kinnitamata haigestunuid.

Teise haigestumise laine ajal oli testimine avatud kõigile, mistõttu võib sellel ajal

			Patsientide arv			
Laine	**Ajaperiood**	**Staatus**	**Kokku**	**Haigla-ravi**	**Intensiiv-ravi**	**Surm**
1.	veebruar - august 2020	Vaktsineerimata	2115	147	81	37
2.	september 2020 - juuni 2021	Vaktsineerimata	90928	2255	960	810
		Vaktsineeritud	2526	25	7	<5
3.	juuli - november 2021	Vaktsineerimata	37758	336	117	78
		Vaktsineeritud	8388	36	10	<5

Tabel 2. Andmestikus sisalduvate patsientide jaotus kõigi valideerimishulkade ja tulemite vahel.

haigestunutest koosnevat valimit lugeda parimaks üldistuseks kõigile teise laine ajal COVID-19 nakkuse saanud inimestele. Teise laine alguses hakati piiratud osa populatsioonist juba vaktsineerima, ja laine lõpuharjal lubati vaktsineerima ka üldpopulatsioon.

Kolmanda, ja andmestikus viimasena esindatud, haigestumise laine ajal oli testimine ja ka vaktsineerimine olnud pikemalt avatud kõigile.

Treening- ning valideerimisvalimid peaksid olema võimalikult sarnaselt koostatud. Võttes arvesse nii testimise avatust populatsioonile kui ka vaktsineerimise levimust valiti teise laine vaktsineerimata COVID-19 haigestunud treeningvalimiks ja ajaline valideerimine teostati kolmanda laine vaktsineerimata haigestunute valimil. Coriva andmestikus oli teise laine vaktsineerimata haigestunuid 90928 ja kolmanda laine vaktsineerimata haigestunuid 37758. Patsientide jaotumine valideerimishulkade ja tulemite vahel on toodud välja tabelis 2

3.2 Masinõppe algoritmi valik

PLP tarkvarapaketis on sisse ehitatud kaheksa algoritmi, mille vahel valida riskimudeli ehitamisel. Edaspidi on neist väljatoodud neli, mis andsid kõige paremaid tulemusi - lasso logistiline regressioon, juhumets, gradientvõimendus ning kohanduvvõimendus.

Iga algoritmi kohta treeniti kolm mudelit - üks iga tulemi kohta - kokku kaksteist

riskimudelit. Et isoleerida ja võrrelda ainult algoritmi mõju riskimudeli headusele jäid ülejäänud mudeli treenimise parameetrid vaikeväärtusteks ning olid kõigil treenitud mudelitel samad. Algoritmide hüperparameetreid siin sammul ei muudetud. Ennustavaid tunnuseid uuritakse täpsemalt järgmises alapeatükis, kuid kasutatud PLP vaikeseadistus on tabelis 6 lisas I välja toodud kui seadistus 13, koos täpselt mudelisse kaasatud tunnustega. Riskimudelid treeniti ajalise valideerimise võimaldamise tarbeks teise laine vaktsineerimata COVID-19 haigete andmetel. PLP pakett seadistati mudeleid treenima kasutades 5-kordset ristvalideerimist. Ajaline valideerimine viidi läbi kõigil treenitud mudelitel kolmanda laine vaktsineerimata patsientide andmetel.

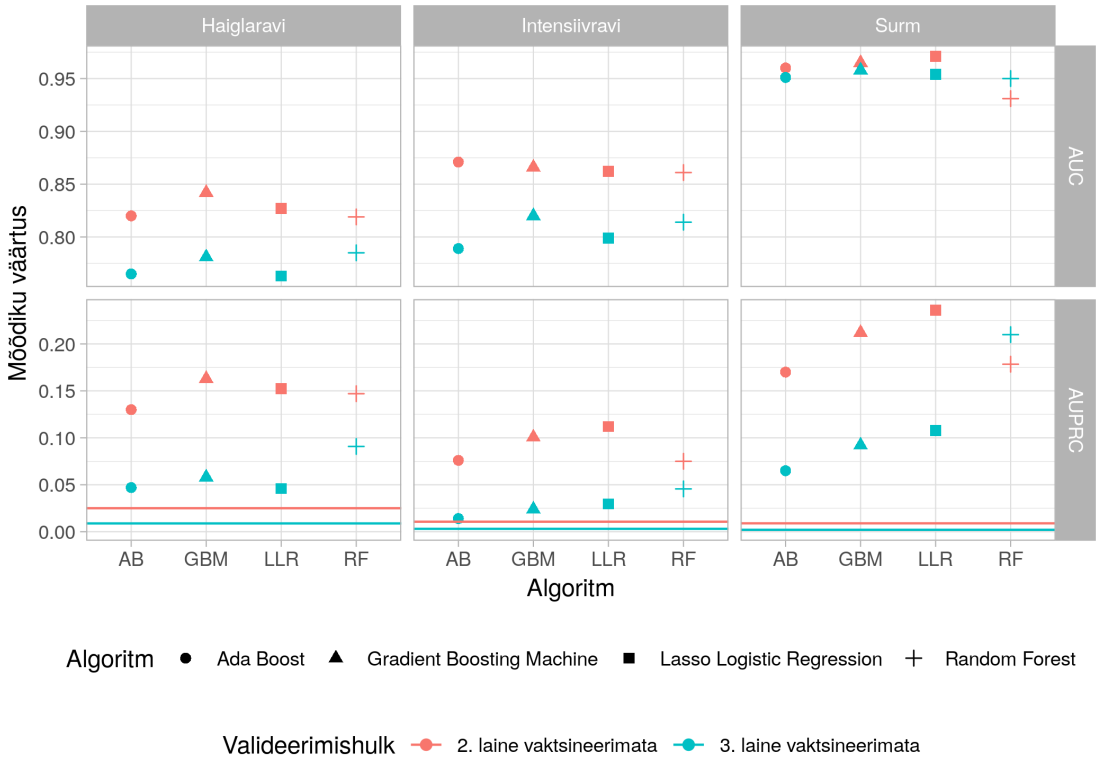
3.2.1 Tulemused

Valideerimise tulemusel saadud AUC ja AUPRC väärtused on kujutatud joonisel 5.

AUC väärtused eri algoritmide korral erinevad vähe, sama valideerimishulga sees kõige rohkem 0.03 võrra. Mudelite AUC näitajad püsivad head ka ajalisel valideerimisel. Surma ennustusmudelid saavutasid AUC tulemused üle 0.925 nii ristvalideerimise kui ka ajalise valideerimise korral, kõigist mudelitest näitasid parimat diskriminiseerimisvõimet just surma mudelid. AUC põhjal teistest selgelt paremat algoritmi ei leidu.

AUPRC joonisel toodud horisontaaljooned näitavad tulemite osakaalu kogu valimist. Tegu on selgelt kallutatud andmetega, tulemitest enim esines 2.5% teise laine valimil haiglaravivajadus, teiste tulemite esinemine jääb alla 1% valimitest. Seega oligi oodata, et kallutatud andmete korral ebatäpse AUC põhjal järeldusi teha ei saa ja peaks parima algoritmi valimiseks kasutama AUPRC mõõdikuid.

Lisaks AUPRC arvulistele väärtustele jooniselt 5 on hea võrrelda PR-kõveraid joonisel 6. Ristvalideerimisel saadud AUPRC väärtused on mudelite lõikes üpris sarnased. Suur erinevus tuleb välja ajalise valideerimise tulemustest, millest on näha, et juhumetsade mudelid kaotavad diskrimineerimisvõimest vähem kui teised mudelid, andes kolmanda laine patsientidel parimad tulemused.

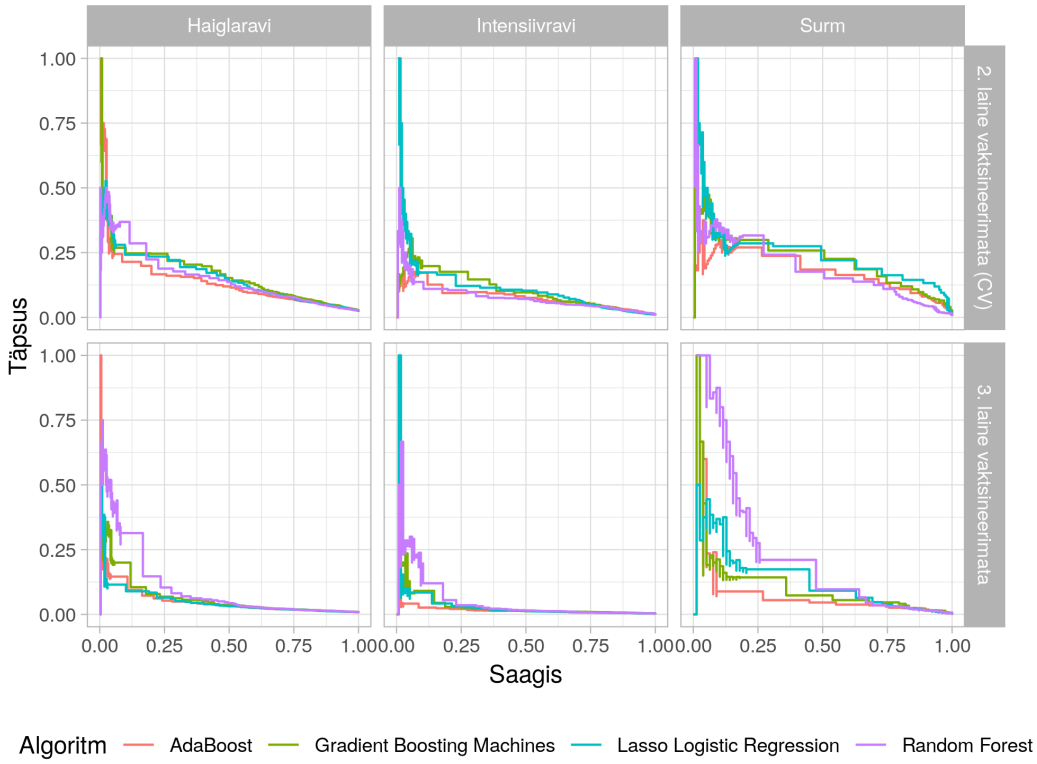


Joonis 5. Erinevate masinõppealgoritmide saavutatud AUC ja AUPRC mõõdikud kõigi tulemite korral. AUPRC baasväärtused mõlema valideerimishulga jaoks on toodud vastavat värvi horisontaaljoonega.

Juhumetsa algortimi kasutanud mudelite üldistusvõime oli parim kõigi tulemite jaoks, mistõttu edaspidi kasutatakse uuringus ainult juhumetsa algoritmi.

3.3 Kovariaadid

Peale mudeli algoritmi on teine oluline valik riskimudeli treenimisel ennustavate tunnuste ehk kovariaatide valik. PLP pakett võimaldab mudeli treenimisel arvesse võtta 26 eri tüüpi kovariaati [10]. Mõnele nendele vastavat infot pole Eesti terviseandmetest võimalik kätte saada, näiteks rassi või rahvust. Viieteistkümnele kovariaaditüübile saab seadistada 3-8 ajaperioodi, milles tunnust vaadeldakse. Lisaks on ühel tunnusel neli erinevat alavalikut. Kokku on kovariaadiseadistuse loomisel võimalik teha 103 valikud. Järgnevalt tuuakse



Joonis 6. PR-kõverad eri algoritmide ja kõigi tulemite korral.

välja, milliseid neist selles uuringus kasutati.

Sugu ja vanusegrupp Vanust grupeeritakse viieaastaste vahemikena. Tunnusel „Sugu“ on kaks väärtust, tunnusel „Vanusegrupp“ on 22 väärtust.

Indekskuu Indekskuu *(IndexMonth)* võimaldab mudelis arvesse võtta kalendrikuud, mil patsient arvati uuringuvalimisse. Selle uuringu raames tähistab indekskuu kalendrikuud, mil patsient haigestus COVID-19. Tunnusel „Indekskuu“ on 12 võimalikku väärtust.

Riskiskoorid Mudelisse on võimalik lisada nelja erinevat riskiskoori - CHADS2, DCSI, CHA2DS2VASc, Charlson.

Kõigi järgnevate tunnuste esinemist saab vaadelda lühiajaliselt (30-7 päeva jook-

sul enne indekskuupäeva) või pikaajaliselt (365-7 päeva jooksul enne indekskuupäeva). COVID-19 kinnitava testitulemuse saab patsient mõne päeva jooksul pärast arstile pöördumist, kuid juba esimese pöördumise ajal võib arst määrata mõne diagnoosi või ravimi raskemini põdevale haigele. Sellised andmed lekitaksid infot patsiendi haiguskäigu kohta enne „ametlikku“ haigestumist, mistõttu haigestumisele eelneva nädala terviseandmed on jäetud välja, et mudel jääks rakendatavaks tervele populatsioonile, mitte ainult haigestunud patsientidele.

Haigusseisund Haigusseisundi all mõeldakse enamasti diagnoose, aga ka täpsustusi nagu näiteks vähistaadiumid. Haigusseisundit saab arvesse võtta kolme eri tunnusega: (1) Haigusseisundi esinemine (*ConditionOccurence*), (2) Haigusseisundi kestvus (*ConditionEra*) ja (3) üldistatud haigusseisundi kestvus (*ConditionGroupEra*). Haigusseisundi esinemine vaatleb üksikut diagnoosi kirjet. Haigusseisundi kestvus on arvutatud diagnoosikirjete põhjal ja kirjeldab perioodi, mil patsiendil haigusseisund esines. Patsiendil loetakse olevat üldistatud haigusseisund kui tal on haigusseisund või mõni haigusseisundiga seotud alamdiagnoos [10]. Näiteks Kui patsiendil on diagnoositud I50.1 Vasaku vatsakese puudulikkus [14], siis loetakse tal olevat ka I50 Südamepuudulikkus. Coriva andmestikus on erinevaid haigusseisundeid 6563.

Ravim Ravimeid on sarnaselt haigusseisunditega võimalik arvestada kolme eri moodi: (1) Ravimi tarvitamine (*DrugExposure*), (2) ravimi kasutusaeg (*DrugEra*) ja (3) ravimiklassi kasutusaeg (*DrugGroupEra*). Ravimi tarvitamine vaatab, kas patsiendile on ravimit välja kirjutatud. Ravimi kasutusaeg on ravimiretseptide põhjal arvutatud, kui pikal perioodil patsient neid ravimeid võtab. Ravimiklass grupeerib kokku sarnase kasutusalaga ravimid. Näiteks kui patsient on tarbinud antibiootikum penitsilliini, siis tal märgitakse olevat nii penitsilliini kasutamine, kui ka üldiselt antibiootikumide tarbimine. Coriva andmestikus on erinevaid ravimeid 915.

Protseduur Protseduurid, mida patsiendile on tehtud. Coriva andmestikus on erinevaid protseduure 410.

Mõõtmine Tunnust (*Measurement*) võetakse arvesse ainult binaarselt mõõdetud/mitte mõõdetud. Mõõtmistulemusi Coriva andmestikus ei ole. Coriva andmestikus on erinevaid mõõtmisi 178.

Vaatlus Vaatluste (*Observation*) alla märgitakse tähelepanekuid, mis muude tunnuste alla ei mahu. Sellised võivad olla näiteks „kodutu“, „vanem keeldub vaktsineerimisest“, „tubaka kasutaja“ jne. Coriva andmestikus on erinevaid vaatluseid 191.

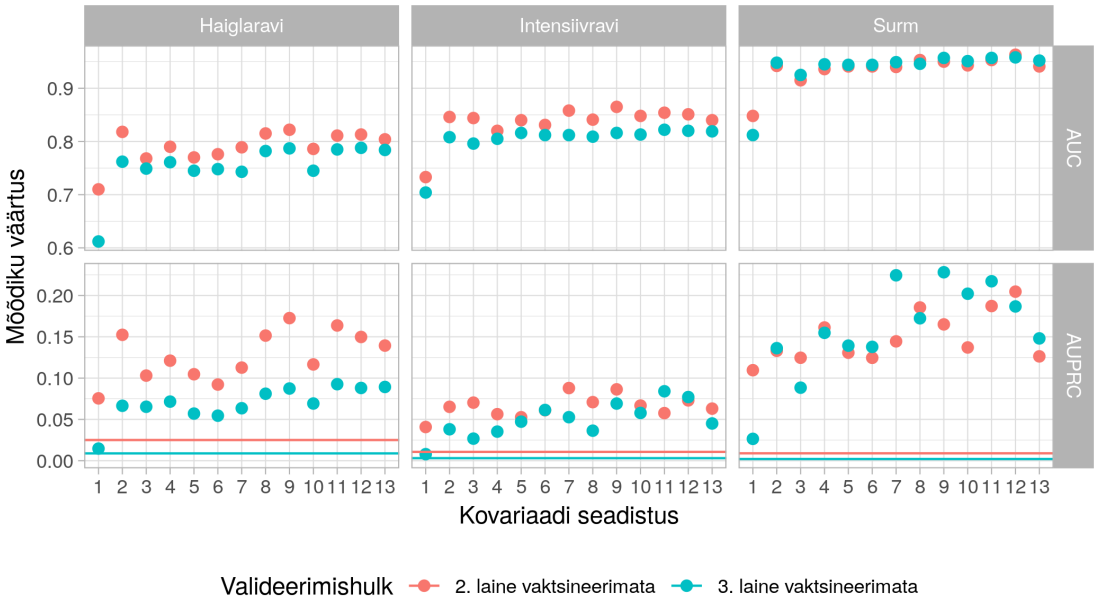
Meditsiiniseade Tunnus näitab meditsiiniseadmeid (*Device*) mida patsient on kasutanud. Coriva andmestikus on erinevaid meditsiiniseadmeid 7.

Samas pole kindel, et kõiki neid tunnuseid on vaja heade COVID-19 haiguskäigu ennustuste saamiseks. Uurimaks erinevate tunnuste mõju mudelile, koostati seadistusi järjest kasvava hulga tunnustega. Kõik treenitud seadistused on välja toodud tabelis 6 lisas I. Uuring seadistati välja viskama tunnuste väärtused, mida esines vähem kui 0.1% patsiendil populatsioonist, kuna nii harvad tunnused ei iseloomusta üldist populatsiooni [10]. Eeldati, et kui patsiendil puudus kirje haigusseisundi, ravimi vms kohta, siis patsiendil seda tunnust ei esine. Mudeli rakendamisel tuleb kindlustada, et analoogsed puuduvad andmed ei ole tingitud süstemaatilisest erinevusest andmekogumisel [23].

Kõigi seadistustega treeniti iga tulemi kohta mudel, kokku 39 mudelit. Mudeli algoritmina kasutati juhumetsa vaikehüperparameetritega. Jällegi treeniti mudelid ajalise valideerimise mõttes teise laine vaktsineerimata COVID-19 haigete andmetel. Mudelite treenimisel kasutatakse 5-kordset ristvalideerimist. Saadud mudelid valideeriti kolmanda laine vaktsineerimata patsientide andmetel.

3.3.1 Tulemused

Eri kovariaadiseadistustega mudelite valideerimiste tulemused on toodud joonisel 7.



Joonis 7. Erinevate kovariaadiseadistustega mudelite saavutatud AUC ja AUPRC mõõdikud kõigi tulemite korral. AUPRC baasväärtused on toodud vastava valideerimishulga värvidega.

AUC mõõdikud on kõrged kõigi kolme tulemi korral. Erandina paistab silma seadistus 1, mis võttis ennustavateks tulemiteks ainult soo ja vanusegrupi. Siit võib järeldada, et sugu ja vanusegrupp ei sisalda piisavalt informatsiooni, et ainult nende põhjal inimese COVID-19 haiguskäiku väga hästi ennustada. AUC väärtused on stabiilsed ka ajaliste muudatuste suhtes, kaotades diskrimineerimisvõimes väga vähe. Kallutatud andmete tõttu pole siiski AUC põhjal võimalik eristada, millised kovariaadid pakuvad parimat ennustusvõimet.

AUPRC mõõdiku järgi on märgata trendi, et rohkemate tunnustega seadistuste tulemused on paremad kui lihtsamatel mudelitel. Arvestades, et AUPRC baasväärtus muutub validatsiooniandmestikuga, on märgata, et kolmandal lainel valideerimine annab isegi paremaid tulemusi, kui ristvalideerimine teise laine patsientidel.

Tabelis 3 on väljavõte suuremast tabelist 7 lisas II, kus on näha täpsed AUPRC väärtused. Tulemusi vaadati keskmiselt üle kõigi tulemite, et vältida tunnuste ülesobita-

	AUPRC (Järjestus)			
Seadistus	Haiglaravi	Intensiivravi	Surm	Keskmine järjestus
9	0.087 (4)	0.069 (3)	**0.228** (1)	2.67
11	**0.093** (1)	**0.084** (1)	0.217 (3)	**1.67**
12	0.088 (3)	0.077 (2)	0.187 (5)	3.33
Baasväärtus	0.009	0.003	0.002	

Tabel 3. Kolme parima kovariaadiseadistuse AUPRC mõõdikud ajalisel valideerimisel.

mist konkreetsele tulemile ja hoida riskimudeli loomise protseduur võimalikult lihtne. Väljatoodud seadistused 9, 11 ja 12 näitasid parimaid AUPRC tulemusi üle kõigi tulemite.

Parimate seadistuste kovariaadikoosseis on toodud tabelis 4. Seadistus 9 ja 11 mõlemad kasutavad sugu, vanusegruppi, protseduure, haigusseisundeid ja ravimeid oma ennustustes. Erinevus kahe seadistuse vahel seisneb kuidas haigusseisundid ja ravimid on arvesse võetud, esimesel juhul vaadeldakse nende tunnuste esinemise perioodi, teisel juhul ainult tunnuste esinemist. Kuna haigusseisundi kestvus on arvutatud haigusseisundi esinemise kirjete põhjal, ja ravimi kasutusaeg on arvutatud ravimi tarvitamise kirjete põhjal, siis võib seadistust 11 pidada lihtsamaks, kuna nõuab ennustuse andmiseks vähem sisendandmete eelnevalt töötlust. Seadistus 12 on seadistuse 11 edasiarendus, kasutades lisaks viite tunnust, kuid AUPRC järgi ei parandanud nende tunnuste lisamine mudeli diskrimineerimisvõimet. Seadistus 11 sai üle kõigi tulemite parima AUPRC tulemused ja on kõige lähedamaid tulemusi saanud seadistustest ka lihtsam. Seetõttu valiti seadistuses 11 sisaldunud tunnused lõpliku riskmudelisse.

Tunnuseid, mida kovariaadiseadistuses 11 mudelisse prooviti oli kokku 9374, millest mudelitesse valiti vastavalt tulemile „Haiglaravi“ 1327, „Intensiivravi“ 1230 ja „Surm“ 1118 tunnust. Seadistuse 11 head või väga head diskrimineerivad mõõdikud näitavad, et haigestumisele eelneva nädala terviseandmete välja jätmine ei mõjutanud oluliselt mudeli tulemusi.

	Kovariaat										
Seadistus	Sugu, vanusegrupp	Haigusseisundi kestvus	Ravimi kasutusaeg	Protseduur	Haigusseisundi esinemine	Ravimi tarvitamine	Indekskuu	Mõõtmine	Vaatlus	Meditsiiniseade	Riskiskoorid
9	+	+	+	+							
11	+			+	+	+					
12	+			+	+	+	+	+	+	+	+

Tabel 4. Kolme parima kovariaadiseadistuse koosseis.

3.4 Hüperparameetrid

Eelnevaga veenduti, et juhumetsa algoritmiga annavad ennustusmudelid parimad tulemused. Samas algoritmi ühtegi hüperparameetrit ei muudetud. Viimase treeningu käigus uuriti ka hüperparameetrite mõju riskimudelile. PLP paketis saab ühele algoritmile anda ette mitu hüperparameetrit, mille kõigi kombinatsioonidega ka mudel treenitakse ja pakett tagastab ristvalideerimise põhjal parima. Varem kasutati hüperparameetrite vaikeväärtuseid, järgnevalt on toodud need juhumetsa jaoks [10].

- Puu sügavus: 4, 10, 17
- Puude arv metsas: 500
- Tunnuste arv ühes puus: ruutjuur kõigi tunnuste arvust
- Teosta enne puude treenimist kasulike tunnuste valimine: jah

Mudeli ja kovariaatide valimise sammudel realiseerusid kõigil mudelitel vaikeväärtuste seast komplekt (sügavus: 17, puid: 500, tunnuseid puus: ruutjuur, eelvalik: jah). Uurimaks, kuidas hüperparameetrid mõjutavad riskimudeli tulemusi, laiendati hüperparameetrite valikuid.

	Hüperparameeter			
	Puu sügavus	Puude arv	Tunnuste arv	Tunnuste eelvalimine
Varasem parim	17	500	Ruutjuur	Jah
Haiglaravi	25	1000	10	Jah
Intensiivravi	25	1000	10	Jah
Surm	25	1000	Ruutjuur	Jah

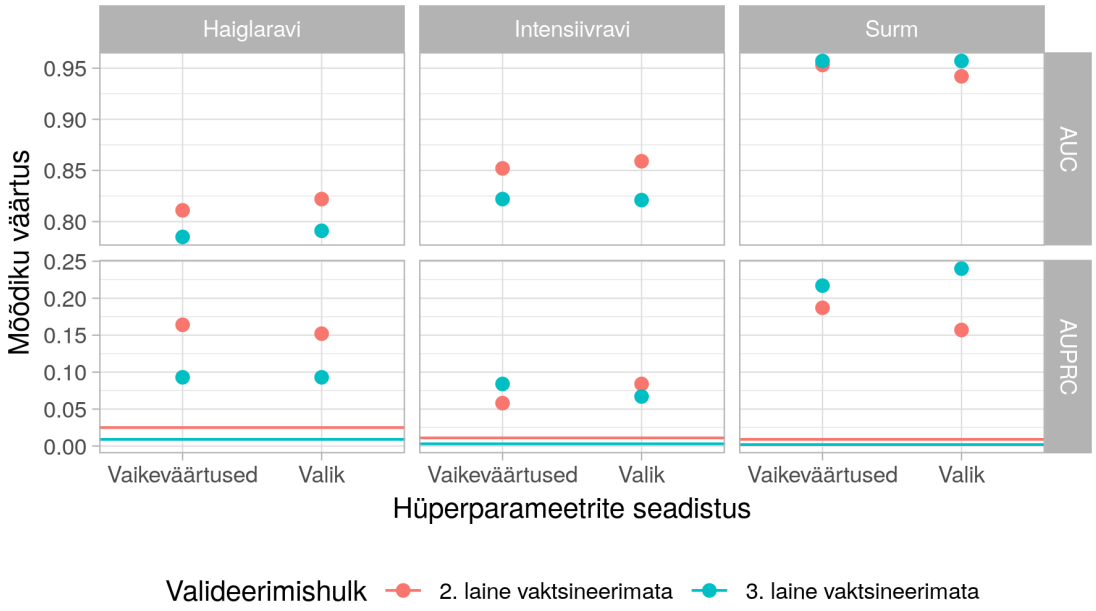
Tabel 5. Realiseerunud hüperparameetrid laiendatud valikust.

- Puu sügavus: 4, 10, 17, 25
- Puude arv metsas: 100, 500, 1000
- Tunnuste arv ühes puus: ruutjuur kõigi tunnuste arvust, 5, 10, 50, 200
- Teosta enne puude treenimist kasulike tunnuste valimine: jah, ei

Mudelid treeniti iga tulemi jaoks, kokku kolm mudelit. Ennustavad tunnused valiti samad, mis peatükis „Kovariaadid“ parimaks osutusid. Mudelid treeniti teise laine vaktsineerimata COVID-19 haigete andmetel, et ajaline valideerimine läbi viia kolmanda laine vaktsineerimata haigete andmetel. Mudeli sisemine valideerimine viidi läbi 5-kordse ristvalideerimisega.

3.4.1 Tulemused

Laiendatud valiku seast realiseerunud hüperparameetrid on ära toodud tabelis 5. Ainuke parameeter, mis jäi kõigi mudelite jaoks samaks oli „Tunnuste eelvalimine“. Hüperparameeter „Puu sügavus“ osutus kõigi kolme tulemi mudelite jaoks 25, mis on suurem kui varasemalt maksimaalselt võimalik olnud 17. Ka puude arv kasvas kõigil mudelitel 500 pealt 1000 puuni. Puude arvu ja sügavuse suurendamine võivad viia ülesobitamisele, mida algoritmi hüperparameetrite valikus on võib-olla kompenseeritud siis väiksema arvu tunnustega puudes 10. Ainult tulemi „Surm“ mudelis jäi tunnuste arv samaks - ruutjuur kõigist võimalikest tunnustest ehk $\sqrt{9374} \approx 97$.

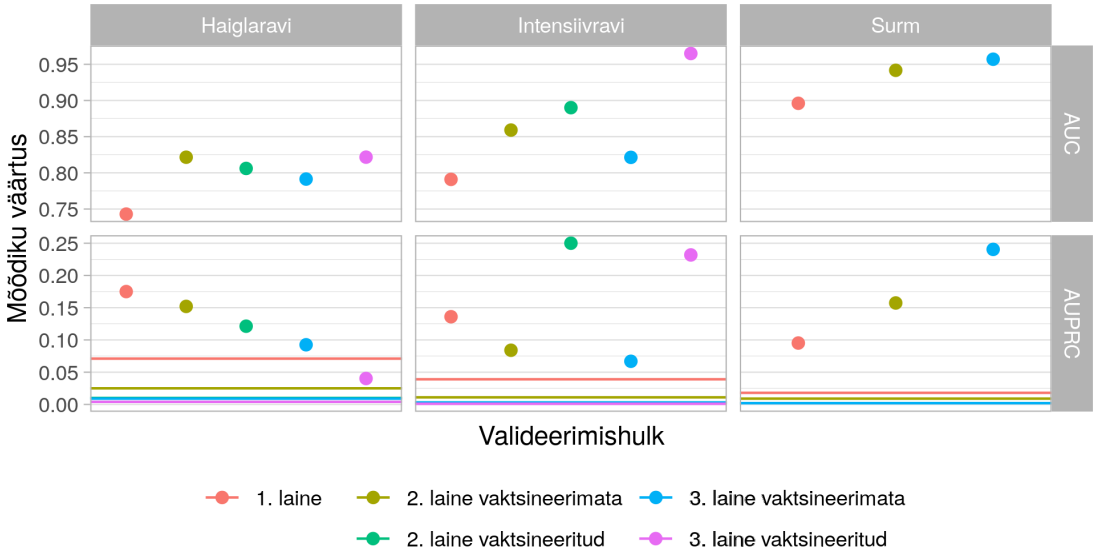


Joonis 8. Hüperparameetrite seadistamise valideerimise tulemused võrreldes vaikeväärtustega. AUPRC baasväärtused on toodud horisontaaljoontega.

Kuidas hüperparameetrite muutus mudelite tulemusi mõjutas on näha jooniselt 8. Mõned uute hüperparameetritega saadud tulemused on ristvalideerimisel teise laine haigetel kehvemad kui varasemalt saadud tulemused. Kuna erinevused on pigem väikesed siis võib arvata, et siin on tegu juhuslikkuse mõjuga masinõppe protsessis. Ajalise valideerimise põhjal hüperparameetrite seadistamisega mudeli headus oluliselt ei kasvanud.

3.5 Lõplikud mudelid

Kokkuvõtvalt osutus uuringu tulemusel parimaks masinõppe algoritmiks juhumetsad, hüperparameetrite seadistus sõltuvalt mudeli tulemist on toodud tabelis 5. Ennustavateks tunnusteks mudelites jäid valikusse sugu (2 väärtust), vanusegrupp (22 väärtust), protseduurid (kuni 410 väärtust), haigusseisund (kuni 6563 väärtust) ja ravimid (kuni 915 väärtust). Protseduuride, haigusseisundite ja ravimite esinemist vaadeldi aasta ja kuu jooksul enne COVID-19 haigestumist. Oluliseks osutusid tulemile „Haiglaravi“ 1327,



Joonis 9. Lõpliku mudeli diskrimineerimismõõdikud kõigil andmestiku valideerimishulkadel. AUPRC baasväärtused on näidatud horisontaaljoontega.

„Intensiivravi“ 1230 ja „Surm“ 1118 tunnust. Lõpliku mudeli kirjeldus on nähtaval veebirakenduses `http://omop-apps.cloud.ut.ee/ShinyApps/COVID-19_FinalModel/`.

3.5.1 Diskrimineerimine

Lisaks teise ja kolmanda laine vaktsineerimata patsientidele, valideeriti lõplik mudel ka esimese laine ning vaktsineeritud patsientidel. Kõigil võimalikel valideerimishulkadel saavutatud diskrimineerimismõõdikud on toodud joonisel 9.

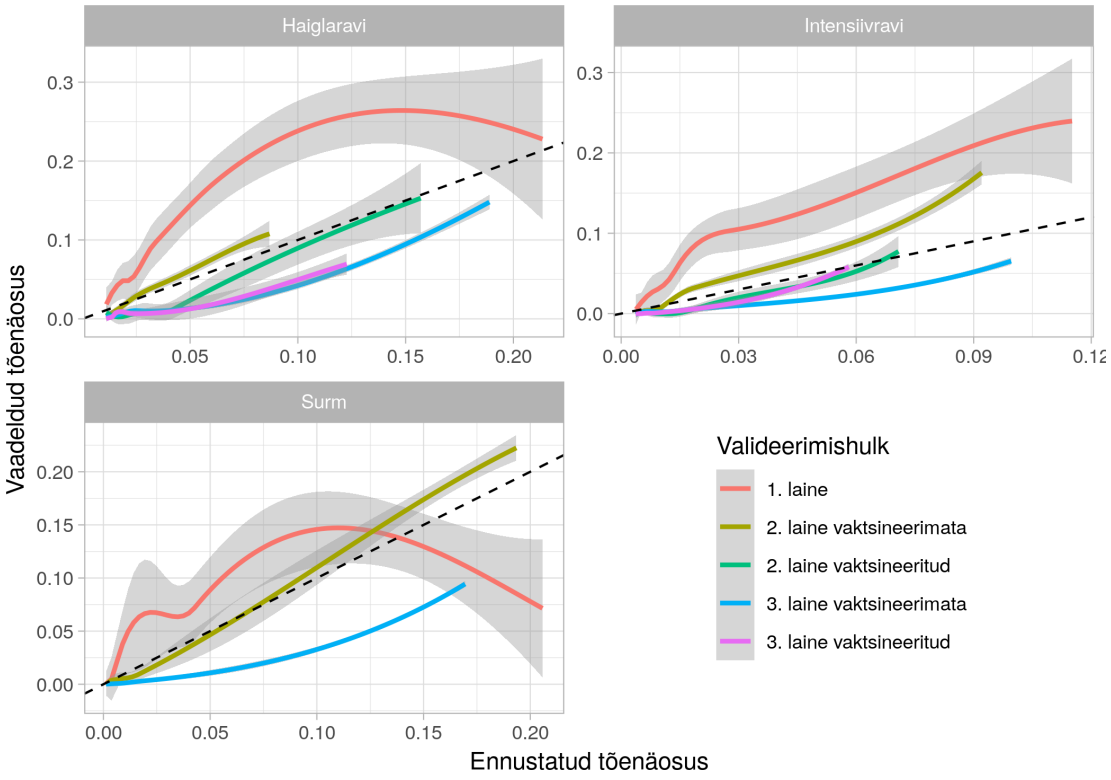
Esimese laine patsientidel jäävad mudeli AUC väärtused madalamaks, kui teistel valideerimishulkadel. See võib tuleneda esimese ja teise laine vahelisel ajal tehtud muutustest ravi ja diagnostika praktikas.

Vaktsineeritud patsientide seas esines tulemit „Surm“ vähem kui 5 inimesel, mistõttu diskrimineerimise mõõdikuid nendel hulkadel välja ei arvutatud. AUC mõõdiku järgi olid lõpliku mudeli ennustused vaktsineerimata patsientidele sarnased või täpsemad kui vaktsineerimata patsientidel. AUPRC mõõdik näitab haiglaravi ennustamisel veidi

kehvemaid tulemusi ja intensiivravi korral palju täpsemaid ennustusi vaktsineeritud inimestele. Arvestades ka tulemiga vaktsineeritud patsientide vähesust võib arvada, et mudel vastab nii vaktsineeritud kui ka vaktsineerimata patsientidele rahuldavad.

3.5.2 Kalibratsioon

Selgitamaks, kas mudeli ennustatud tõenäosused COVID-19 haiguskulu kohta vastavad ka vaadeldud tõenäosustele, vaadati kalibreerimiskõveraid joonisel 10.



Joonis 10. Kalibreerimiskõverad kõigi valideerimiskõverate korral. Musta katkendliku joonega on toodud ideaalse kalibratsiooni joon.

Esimese laine kalibratsioonikõverad on kõige laiemate usaldusvahemikega, mida seletab valideerimishulga väiksus ning valideerimishulka sattumise tingimuste suur erinevus. Treenitud mudelid arvestasid üldisema populatsiooniga, kui esimese laine ajal patsiendid COVID-19 diagnoosi said, mistõttu ka tulemused on vähem kindlad. Siiski,

nii haiglaravi kui intensiivravi vajadust ennustasid mudelid vähem kui tegelikkuses ette tuli.

Kalibratsioonikõverad on üpris lähedased ideaalsele teise laine mõlemal hulgal, mida oli ka oodata, kuna ühel neist hulkadest oli mudel treenitud, ja teine oli ajalises mõttes kõige sarnasem valideerimishulk.

Kolmanda laine vaktsineerimata hulga jaoks said ennustatud tõenäosused mõnevõrra suuremad kui reaalelus vaadeldud tõonäosused. Seda võib seletada näiteks muutused diagnostikas, kui kolmanda laine ajal diagnoositi rohkem mittesümptomaatilisi patsiente tänu kiirtestide laialdasemale kasutusele.

Kalibratsioonikõverad koos näitavad, et kuigi mudeli diskrimineerimisvõime jäi läbi kõigi lainete heaks, siis praktilise kasutamise jaoks tuleks jälgida mudelite kalibratsiooni ajaliste muutuste suhtes ning vajadusel mudelid ümber kalibreerida.

3.5.3 Ennustavad tunnused

Lõplikud mudelid võtavad arvesse mõnevõrra erinevaid ennustavaid tunnused ning samu tunnuseid ka erineva olulisusega. Kuigi kõik välja toodud riskitunnustest ega kaitsvatest tunnustest ei pruugi olla põhjuslikus seoses COVID-19 haiguskuluga, siis nende seosed tulemitega olid piisavalt tugevad, et olla ennustamisel kasulikud.

Haiglaravi vajadust ennustav mudel määras kõige olulisemateks riskitunnusteks viimase aasta jooksul esinenud haigusseisundeid kõrgvererõhutõbe, südamerabandust ning südamerütmihäireid. Protseduuridest olid suurima kaaluga riskifaktorid viimase aasta jooksul tehtud kompuutertomograafia uuringud, PCR testid ja EKG uuringud. Ravimite seast olid olulisemad torsemide, metoprolol ja pantoprazole, mis on vastavalt kasutusel kõrgvererõhutõve, südamehaiguste ja maohaavandite ravis. Vanusegrupp 80-84 ja meessugu olid vastavatest tunnustest suurima kaaluga riskitegurid, kuid nende tunnuste olulisus jäi alla haigusseisundite, ravimite ja protseduuride olulisusele. Üllataval kombel kõige tugevama kaitseefektiga tunnus oli protseduur "COVID-19 vaktsineerimisest

loobumine". Veel olid kaitsva toimega naissugu, vanuserühmad 45-59, günekoloogilised läbivaatused ning ravimi cetirizine (allergiaravim) tarvitamine.

Intensiivravi vajaduse ennustamisel olid riskitunnusteks samuti viimase aasta jooksul esinenud haigusseisundid kõrgvererõhutõbi, südamerabandus ja südamerütmihäired. Lisaks olid riskiteguriteks ravimite torsemide, metoprolol kasutamine, mis on vererõhuravimid, ja allopurinol, mida kasutatakse podagra ravis. Protseduuridest olulisim riskitunnus oli antibiootikumide kuuri läbimine ja kõrgeima riskiga oli vanusegrupp 80-84. Ka intensiivravi korral oli COVID-19 vaktsiinist loobumine tugevaim kaitsev tunnus, lisaks naissugu ja günekoloogilised läbivaatused.

Surma ennustava mudeli olulisimad näitajad olid vanusegrupid 80-95, kusepõie kateteriseerimine ja dementsus. Kõrgel kohal olid esindatud ka teistes mudelites olulised riskitunnused kõrgvererõhutõbi ja selle ravimid, südamerabandus ja -rütmihäired. Samamoodi oli ka surma ennustavate mudelites kaitsetunnusteks COVID-19 vaktsiinist loobumine, naissugu ja günekoloogilisel läbivaatused.

Kõigi mudelites olid riskitunnused 3-5 kordades olulisemad näitajad kui kaitsvad tunnused. See oli oodatav, kuna enamus võimalike tunnuseid käivad haigusseisundite kohta, mis on tihedamini nõrgendavad organismi ja immuunsust, kui teevad tugevamaks. Samuti oli varasemalt kovariaatide valikul näha, et ainult sugu ja vanus ei sisalda piisavalt infot heade mudelite saamiseks, mistõttu ka lõplikutes mudelites nende olulisus ei olnud väga suur.

4 Kokkuvõte

Käesoleva töö eesmärgiks oli luua Eesti terviseandmetel COVID-19 haiguskulgu ennustavad riskimudelid. Uuringu käigus sooviti ennustada COVID-19 haige inimese tõenäosuseid haigestumisest 30 päeva jooksul sattuda haiglaravile, vajada intensiivravi või surra. Riskimudelite loomiseks kasutati Coriva projekti raames väljastatud ja CDM kujule standardiseeritud Haigekassa registriandmeid vahemikust veebruar 2017 kuni november 2021.

Riskimudelite loomiseks läbiti kolmesammuline protsess, alustades riskimudeli algoritmi valikuga, liikudes edasi ennustavate tunnuste valimisega ning lõpetades parima algoritmi ja ennustavate tunnuste kombinatsiooni jaoks optimaalsete hüperparameetrite valikuga. Igal sammul teostati mudelitele ajaline valideerimine, mille põhjal teostatud valikud lubasid parimat üldistuvust viiruse ajas muutuva loomuse suhtes. Mudelite loomisel ei piirdutud lihtsate lineaarsete mudelitega ega limiteeritud ennustavate tunnuste arvu.

Parimaks algoritmiks osutus juhumets, mis on mõnevõrra keerukama tõlgendatavusega kui riskimudelite jaoks enimlevinud logistiline regressioon, kuid näitas paremat üldistavust ajalisel valideerimisel. Ennustavaid tunnuseid kaasati kõigi tulemite korral üle 1000, mida on rohkem kui varasemates COVID-19 riskimudelites. Vahetult enne haigestumist saabunud terviseandmete väljajätmine vältis andmeleket tulevikusündmuste kohta kuid ei mõjutanud mudeli täpsust.

Lõplikud riskimudelid saavutasid ajalisel valideerimisel vaktsineerimata patsientidel haiglaravi korral diskrimineerimisvõime AUC 0.791, intensiivravi korral 0.821 ja surma korral 0.957. Mudelite AUC mõõdikud on tugevad, kuid mudelite kalibreeritus näitab, et viiruse mutatsioonidest tulenevad muutused haiguskulus mõjutavad mudelite ennustuste täpsust. Seetõttu esineb selge vajadus riskimudelite kliinilisel kasutamisel mudelite kalibreeritust pidevalt kontrollida ja parandada.

Kuigi praeguseks on avalik huvi COVID-19 haigestumise uurimiseks ja ennusta-

miseks raugenud, saab käesolevas töös kirjeldatud riskimudelite loomise protsessi järgida tulevaste uute nakkushaiguste levimisel. Saadavate riskimudelite rakendamine võib anda näiteks lisainformatsiooni personaalmeditsiini põhimõtetel riikliku vaktsiinipoliitika välja töötamiseks.

Viidatud kirjandus

[1] Mari-Liis Allikivi and Ardi Tampuu. Tehisintellekti algkursus (LTAT.TK.013) 2019/20 kevad. https://courses.cs.ut.ee/2020/Tehisintellekti_algkursus/spring/Main/PARTIIJuhendatud, May 2022. (17.05.2022).

[2] Alain Bernard. Clinical prediction models: a fashion or a necessity in medicine? *Journal of Thoracic Disease*, 9(10):3456, 2017.

[3] Mary E. Charlson, Peter Pompei, Kathy L. Ales, and C.Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5):373–383, 1987.

[4] Ash K Clift, Carol A C Coupland, Ruth H Keogh, Karla Diaz-Ordaz, Elizabeth Williamson, Ewen M Harrison, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ*, 371, 2020.

[5] Marc David. COVID-19 ennustavate riskimudelite rakendatavuse hindamine Eesti terviseandmetel. Bakalaureusetöö, Tartu ülikool, 2021.

[6] Suzanne Ekelund. Precision-recall curves—what are they and how are they used. https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used, 2017. (17.05.2022).

[7] A. Cecile J.W. Janssens and Forike K. Martens. Prediction research - an introduction. http://www.cecilejanssens.org/wp-content/uploads/2021/04/PredictionManual2.3.pdf, 2020. (17.05.2022).

[8] Neil K. Kaneshiro. Apgar score. https://medlineplus.gov/ency/article/003402.htm, 2020. (17.05.2022).

[9] Tatjana Meister, Heti Pisarev, Raivo Kolde, Ruth Kalda, Kadri Suija, Lili Milani, et al. Clinical Characteristics and Risk Factors for COVID-19 Infection and Disease Severity: A Nationwide Observational Study in Estonia. *Available at SSRN 3955730*, 2021.

[10] OHDSI. *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI, 2019.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Fabio Sigrist. Demystifying ROC and precision-recall curves. `https://towardsdatascience.com/demystifying-roc-and-precision-recall-curves-d30f3fad2cbf`, 2022. (17.05.2022).

[13] Sotsiaalministeerium. Kuus fakti koroonaviiruse mutatsioonide ja vaktsiinide kohta. `https://vaktsineeri.ee/blogi/kuus-fakti-koroonaviiruse-mutatsioonide-ja-vaktsiinide-kohta/`. (17.05.2022).

[14] Sotsiaalministeerium. RHK-10 diagnoosikoodid. `https://rhk.sm.ee/`. (17.05.2022).

[15] Statistikaamet. Rahvaarv. `https://www.stat.ee/et/avasta-statistikat/valdkonnad/rahvastik/rahvaarv`. (17.05.2022).

[16] Ewout W Steyerberg. *Clinical prediction models*. Statistics for Biology and Health. Springer, New York, NY, 2009 edition, December 2009.

[17] Tartu ülikooli peremeditsiini ja rahvatervishoiu instituut. Projekt CORIVA. https://tervis.ut.ee/et/sisu/projekt-coriva. (17.05.2022).

[18] Graham Teasdale and Bryan Jennett. Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, 304(7872):81–84, 1974.

[19] Terviseamet. Mis on COVID-19 ja kuidas sellest hoiduda? https://www.terviseamet.ee/et/mis-covid-19. (17.05.2022).

[20] Terviseamet. Sünnitus, raseda ja vastusündinu jälgimine COVID-19 epideemia muutunud tingimustes. https://www.terviseamet.ee/sites/default/files/Nakkushaigused/Juhendid/COVID-19/sunnitus_raseda_ja_vastsundinu_jalgimine_covid19_epideemia_muutunud_tingimustes.pdf, 2021. (17.05.2022).

[21] Valitsuse kommunikatsioonibüroo. Esmaspäevast tagatakse eelkõige erakorralise ja vältimatu abi andmine. https://www.kriis.ee/uudised/esmaspaevast-tagatakse-eelkoige-erakorralise-ja-valtimatu-abi-andmine, 2021. (17.05.2022).

[22] Helen Ward, Matthew Whitaker, Barnaby Flower, Sonja N Tang, Christina Atchison, Ara Darzi, et al. Population antibody responses following COVID-19 vaccination in 212,102 individuals. *Nature communications*, 13(1):1–6, 2022.

[23] Ross D. Williams, Aniek F. Markus, Cynthia Yang, Talita Duarte Salles, Scott L. DuVall, Thomas Falconer, et al. Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. *medRxiv*, 2020.

[24] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*, 369, 2020.

Lisad

I. Ennustavate tunnuste valikud

	Seadistus												
Ennustav tunnus	1	2	3	4	5	6	7	8	9	10	11	12	13
Sugu (*Gender*), vanusegrupp (*Age group*)	+	+	+	+	+	+	+	+	+	+	+	+	+
Üldistatud haigusseisundi kestvus (*ConditionGroupEra*)		+		+				+					+
Ravimiklassi kasutusaeg (*DrugGroupEra*)			+	+				+					+
Haigusseisundi kestvus (*ConditionEra*)					+		+		+				
Ravimi kasutusaeg (*DrugEra*)						+	+		+				
Protseduur (*Procedure*)								+	+		+	+	+
Haigusseisund (*ConditionOccurence*)										+	+	+	
Ravim (*DrugExposure*)										+	+	+	
Indekskuu (*IndexMonth*)												+	+
Mõõtmine (*Measurement*)												+	+
Vaatlus (*Observation*)												+	+
Meditsiiniseade (*Device*)												+	+
Riskiskoorid (*IndexScores*)												+	+

Tabel 6. Ennustavate tunnuste nimekiri ning märge, millises seadistuses see sisaldus.

II. Kovariaatide otsingu AUPRC tulemused

	AUPRC (Järjestus)			
Seadistus	Haiglaravi	Intensiivravi	Surm	Keskmine järjestus
1	0.015 (13)	0.008 (13)	0.026 (13)	13.00
2	0.066 (8)	0.038 (9)	0.136 (11)	9.33
3	0.065 (9)	0.027 (12)	0.088 (12)	11.00
4	0.072 (6)	0.035 (11)	0.155 (7)	8.00
5	0.057 (11)	0.047 (7)	0.139 (9)	9.00
6	0.055 (12)	0.061 (4)	0.138 (10)	8.67
7	0.064 (10)	0.053 (6)	0.224 (2)	6.00
8	0.081 (5)	0.036 (10)	0.172 (6)	7.00
9	0.087 (4)	0.069 (3)	**0.228 (1)**	2.67
10	0.069 (7)	0.058 (5)	0.202 (4)	5.33
11	**0.093 (1)**	**0.084 (1)**	0.217 (3)	**1.67**
12	0.088 (3)	0.077 (2)	0.187 (5)	3.33
13	0.089 (2)	0.045 (8)	0.148 (8)	6.00
Baasväärtus	0.009	0.003	0.002	

Tabel 7. Juhumetsa algoritmiga treenitud mudelite AUPRC tulemused erinevate kovariaatseadistustega.

III. Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Sille Habakukk**,

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

   COVID-19 ennustavate riskimudelite loomine Eesti terviseandmete põhjal,

   mille juhendaja on Raivo Kolde,

   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.

Sille Habakukk

17.05.2022