TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Marc David

COVID-19 ennustavate riskimudelite rakendatavuse hindamine Eesti terviseandmetel

Bakalaureusetöö (9 EAP)

Juhendaja: Raivo Kolde, PhD

Tartu 2021

COVID-19 ennustavate riskimudelite rakendatavuse hindamine Eesti terviseandmetel

Lühikokkuvõte:

COVID-19 leviku tõttu alates aastast 2019 on suurenenud erinevate tervishoiusüsteemide koormus maailmas. Selleks, et haiglate piiratud ressursside kasutust optimeerida, saab kasutada erinevaid riski ennustavaid mudeleid, mis võimaldavad patsientide terviseandmete põhjal ennustada, kui raskeks võib kujuneda patsiendi haiguse kulg. Piisavalt täpset mudelit saab kasutada näiteks patsiendi vaktsiini või ravi vajaduse hindamiseks. Üks mudeli tõhusust mõjutav tegur on kasutatud treeningandmete hulk. Kuna Eesti terviseandmeid on suhteliselt vähe ning nendel uue mudeli treenimine on keeruline, siis on efektiivsem leida juba eelnevalt treenitud mudel ning seda väliselt valideerida Eesti terviseandmetel.

Käesolevas töö eesmärk on maailmas häid tulemusi näidanud mudeleid väliselt valideerida Eesti terviseandmetel ning seejärel analüüsida, kas need mudelid on piisavalt head praktiliseks kasutuseks meditsiinis. Töö tulemused näitasid, et mudelite diskrimineerimine on hea ning võrreldav teiste mudelitele tehtud väliste valideerimistega, kuid kalibreerimine on kehvem. Mudelid ennustavad Eesti andmetel madalamaid riskitõenäosusi kui reaalsuses patsientidel täheldati. Seda selgitab asjaolu, et valideerimiseks kasutatud andmeid oli võrdlemisi vähe ning need võisid selle tõttu olla kallutatud. Mudelite rakendatavust mõjutab ka see, et mudelid on treenitud ja valideeritud terviseandmetel, mis on pärit COVID-19 pandeemia algusest, mistõttu riski ennustamisel ei arvesta mudelid viiruse uute mutatsioonidega ega vaktsineeritud patsientidega. Lahendusena tuleks treenida uued mudelid, kasutades uuemaid andmeid ning parandada valideerimiseks kasutatud andmestike kvaliteeti.

Võtmesõnad:

COVID-19, prognoosmudelid, andmeteadus, bioinformaatika, riskiskoor

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Evaluating Applicability of Different COVID-19 Predictive Risk Models on Estonian Health Data

Abstract:

Since the start of the spread of the novel COVID-19 disease in 2019, the workload on health care systems in the world has increased. Different risk prediction models could be used to optimize the use of health care resources, as it could predict how severe the course of the disease could be for the patient. If the model is accurate enough, it could be used for multiple things, for instance finding patients who would need a vaccine or hospital care the most. A big factor in model performance is the size of the dataset used to train it. Since Estonia has a relatively small amount of health data, then the developing of a new model is difficult. Externally validating a model that is already trained on bigger datasets is more cost-effective and simpler to do.

The goal of this thesis was to externally validate already existing risk models and analyse the potential of their use in a practical way. The results of the validation show that the models' discrimination, as displayed by their AUC and AUPRC values on the Estonian health data is fairly good. However, calibration was poor, as the model predicted a much lower risk probability for patients compared to the observed probability. This could be explained by a bias in the Estonian health data that was used for validation. A barrier for the models' potential use in a practical setting is the fact that these models and the data used to validate the model are outdated, as they are both from the beginning of the pandemic. This means that the models don't consider newer virus mutations or a patient's vaccination history. A solution to these problems would be to train a new model with newer and better data.

Keywords:

COVID-19, predictive models, data science, bioinformatics, risk score

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

Sissejuhatus

Uue COVID-19 haiguse leviku tõttu maailmas on mitmetes riikides suurenenud tervishoiusüsteemide koormus, mis tähendab, et patsientide ravimisel jääb puudu haiglapersonalist ja varustusest. D. P. Mareiniss [1] kirjutab, et sellise situatsiooni tagajärgedeks on suurema arvu patsientide ravita jätmine ning korraliste meditsiiniprotseduuride edasilükkamine, mis vähendab patsientidele antud ravi kvaliteeti ning tõstab suremust. Terviseameti andmetel [2] jõudis 2020. aasta sügisel Eestisse koroonaviiruse leviku teine laine, mille puhul oli nakatunute arv suurem kui sama aasta kevadel. 2021. aasta kevadeks oli nakatumiste arv veelgi suurenenud, mistõttu oli paljudes haiglates töötajad ülekoormatud. Ülekoormuse vähendamiseks on oluline leida viis piiratud ressursside otstarbekaks kasutuseks tervishoiusüsteemis, et patsiente võimalikult adekvaatselt ravida.

Üks võimalus haiglate koormuse leevendamiseks oleks kasutada haigestumisega seotud riski ennustavaid mudeleid, mis suudaksid patsiendi terviseandmete abil prognoosida kui raskeks võib haiguse põdemine kujuneda. Kui leiduks piisavalt täpne mudel, saaks patsientidele eraldada ressursse neile vastava riski tõenäosusega kooskõlas. Wynants et al [3] arvates on hetkel valdav enamus maailmas treenitud mudeleid liiga ebatäpsed või liiga vähe testitud, et neid saaks tervishoiusüsteemides kasutusele võtta. Ebatäpsused on tingitud andmete puudusest (mitmed mudelid olid treenitud pandeemia alguses, kui andmeid oli vähem) ning sellest, et treenitud mudeleid pole üldse või pole piisavalt väliselt valideeritud ehk mudelit pole uutel andmetel piisavalt testitud.

Bakalaureusetöö eesmärk on leida erinevaid hea täpsusega loodud mudeleid ning neid Eesti COVID-19 andmetel valideerida. Väline valideerimine näitab, kui töökindlad on riski ennustavad mudelid ning kas oleks võimalik neile praktilist väljundit leida. Piisavalt head mudelit, mida on ka välise andmestikuga testitud ja mis on hea tõhususega, saaks teoreetiliselt kasutada näiteks haigla personali ja varustuse suunamisel suurema riskiga patsientidele ning vaktsineerimise eelisjärjekorra väljaselgitamisel.

Peatükis „Taust“ kirjeldatakse teema aktuaalsust, ennustavate riskimudelite ajalugu ning riskimudelite arenduse kui ka välise valideerimise protsessi. „Metoodika“ peatükis kirjeldatakse bakalaureusetöö raames valideeritud mudeleid ning nende välist valideerimist, samuti selle töö jaoks Eesti terviseandmetel treenitud mudeli kirjeldust. Peatükis „Tulemused“ tuuakse välja nii väliste valideerimiste kui ka uue mudeli tulemused. Peatükis

„Järeldused“ arutletakse tulemuste üle ja tehakse järeldusi väliselt valideeritud mudelite tõhususest ja rakendatavusest praktilises kontekstis.

1. Taust

1.1 COVID-19 levik ning selle mõju tervishoiusüsteemile

COVID-19 pandeemia välja kuulutamisest 2020. aasta märtsil on Maailma Terviseorganisatsiooni andmetel [4] sellesse haigusesse nakatunud üle 133 miljoni ning surnud üle 2,8 miljoni inimese. Pandeemia algusest alates on maailmas proovitud erinevaid meetmeid, et viiruse levikut piirata ning hoida hospitaliseerimiste ja surmade arvu võimalikult väiksena. Viiruse leviku piiramiseks on kehtestatud näiteks maskikandmise kohustus või riigi lukustamine, kuid enamikes riikides on neid piirangud rakendatud liiga hilja või on neid kergendatud liiga vara, et viiruse levik oleks püsivamalt aeglustunud [5]. Ebatõhusamate COVID-19 leviku peatamise meetmetega riikides on nakatumiste arv hüppeliselt tõusnud ja sellega on kaasnenud suremuse ning haiglatesse sattunud patsientide hulga suurenemine. Nii on juhtunud ka Eestis, kus vaatamata nakatumiste arvu kiirele tõusule ei kehtestatud piisavalt tõhusaid meetmeid, mistõttu oli Eestis Euroopa kõrgeim nakatumiste suhtarv 2021. aasta märtsi kolmandal nädalal [6].

Meetmete ebatõhusa rakendamise tagajärjel on Eestis hüppeliselt tõusnud hospitaliseeritud COVID-19 patsientide arv ning piirkonniti haiglate ülekoormus [7]. Lisaks arvukatele hospitaliseeritud patsientidele raskendab olukorda ka see, et COVID-19 patsient vajab keskmiselt 6 päeva haiglaravi [8]. Hospitaliseerimise käigus hõivab patsient haigla varustust (voodikoht, ventilaator, ravimeid jne) ja mitme tervishoiutöötaja aega ning suurendab seeläbi haigla töökoormust. Sellest olukorrast on tekkinud vajadus mitmes haiglas uute COVID-19 osakondade avamiseks ning patsientide transportimiseks teistesse haiglatesse üle Eesti [7].

COVID-19 haiguse kulu kohta on teada, et haigestunu eelnev terviseseisund mõjutab seda suurel määral. Riskifaktorid patsiendi haiguse raskekujuliseks kujunemiseks on näiteks vanus (65 või vanem), hüpertensioon ehk kõrgvererõhutõbi ning suurenenud LDH[1], lümfotsüütide ja D-dimeeride kontsentratsioon veres. Kusjuures meestel, kellel esinevad eelnimetatud riskifaktorid ning ka varasem südameveresoonkonna komplikatsioon ja kõrge veresuhkru tase, on suurem risk, et raskekujuline haigus võibolla surmav, kirjutavad Li et al [9]. Ka Ji et al [10] on leidnud, et vanus üle 60 eluaasta, varasemad komplikatsioonid ja kõrge LDH ning lümfotsüütide tase näitavad, et patsientidel on suurenenud risk haigust

[1] Laktaadi dehüdrogenaas

raskelt põdeda. Riskitegureid on veel mitmeid ning need kõik võivad mõjutada patsiendi haiguse kulgu.

1.2 Riskigrupid

COVID-19 viiruse põdemist raskendavate riskifaktorite tundmine on oluline, et moodustada nn riskigrupid. Riskigruppi kuuluvad inimesed, kelle puhul võib COVID-19 viirusesse haigestumine suurema tõenäosusega lõppeda hospitaliseerimise või surmaga. Võib oletada, et kui saaks vähendada riskigruppidesse kuuluvate inimeste haigestumist, väheneks ka haiglatesse suunduvate inimeste arv ja sellest tulenevalt haiglate ülekoormus.

Liigse koormuse vähendamiseks on mitu viisi, eelkõige haigestumise ennetus vaktsineerimise abil, kuid ka piiratud ressursside otstarbekas kasutamine haiglas. Selle jaoks on tähtis moodustada riskigrupid võimalikult täpselt, et kõrgeima riskiga inimesed saaksid kiirema ligipääsu vajalikele ressurssidele, olgu selleks vaktsiin või efektiivsem abi haiglas.

Eestis on riskigrupid määratletud arstide ning epidemioloogide koostöös [11], mis on olnud kasulik suurema populatsiooni riskide üldistamises, kuid mis ei arvesta iga potentsiaalse haigestunu individuaalset haiguslugu [12] ning on ebatäpsem kui patsiendikeskne andmepõhine lähenemine.

1.3 Riskimudelid

Üks variant olekski leida individuaalse patsiendi täpsem riski tõenäosus, kasutades ennustavat mudelit, mis arvestades inimese haiguslugu ja demograafilisi tegureid, suudaks leida seoseid tema andmete ning võimaliku haiguse põdemise keerukuse vahel. Sellist mudelit võib lühidalt kirjeldada terminiga riskimudel. Riskimudelite kasutamisel on mitu eelist. Esiteks on riskimudelitega ennustamisel võimalik iga patsiendi kohta määrata spetsiifiline riskitase, millega saaks suurema riskiga inimeste jaoks rohkem ressursse loovutada [13], näiteks pakkuda kiiremini ravi või varem vaktsineerida.

Teine eelis on see, et juhtudel, kus on vaja arvestada kiirelt areneva olukorraga, näiteks koroonaviiruse puhul mutatsioonidega, saab suhteliselt lihtsalt ning odavalt mudeleid täpsustada uute andmetega. See tähendab, et mudeli tõhusus areneks jooksvalt uute andmetega. Veel võib eeliseks pidada seda, et ennustav riskimudel võib arstidele abiks olla patsiendi haiguse kulu ennustamisel [12].

1.4 Masinõpe ning LASSO logistiline regressioon

Masinõppeks nimetatakse tehisintellekti haru, kus püütakse luua mudeleid, mis ise automaatselt leiavad andmetest erinevaid mustreid ning teevad nende abil ennustusi [14]. Tüüpiline masinõppe mudeli loomise protsess näeb välja selline [15]:

- Andmete kogumine ja töötlus.
- Andmete tükeldamine treening- ja testhulgaks.
- Masinõppe algoritmi valimine vastavalt eesmärgile.
- Mudeli treenimine ehk valitud algoritmi iteratiivne jooksutamine treeningandmetel.
- Mudeli valideerimine ehk testimine testandmetel.

Heade tulemuste saavutamiseks on mitu eeldust, kuid eriti tähtsad on andmete kvaliteet ja kvantiteet. Suurema andmehulgaga saavad mudelid treenimisel paremini seoseid leida ning väheneb mudeli üle- või alasobituse risk.

Käesolevas töös käsitletud riskimudelid on masinõppe mudelid. Kasutades erinevaid algoritme, suudavad mudelid leida patsiendi terviseloost saadud tunnuste ja tema meditsiiniliste tulemuste vahel seoseid. Riskimudelid, mida töö raames valideeritakse, on juba eelnevalt treenitud. Samuti on teostatud nendele erineval määral nii sisemist kui ka välist valideerimist. Sisemiseks valideerimiseks peetakse mudeli tõhususe testimist treeningandmete alamhulgaga, mida nimetatakse testhulgaks, ning väliseks valideerimiseks uute, treenimises mittekasutatud andmetega valideerimist. Mida rohkem väliselt valideerida, seda usaldusväärsemaks muutub mudel, sest nii saab teada, kui universaalne on mudel ehk kui hästi saab mudel hakkama praktilises olukorras, kus sisendiks võib olla tundmatu tunnuste kombinatsioon [16].

Selles töös valideeritavad mudelid on kõik treenitud kasutades *Least Absolute Shrinkage and Selection Operator* (LASSO) logistilise regressiooni algoritmi. LASSO tööpõhimõte seisneb selles, et ta tuvastab L1 regularisatsiooni kasutades neid tunnused, mis tulemust rohkem või vähem mõjutavad ning seejärel muudab vastavalt tunnuste koefitsiente suuremaks või väiksemaks [17]. Selle abil muutuvad osade tunnuste koefitsiendid nulliks ning mudeli keerukus ja ülesobituse oht väheneb.

Logistilise regressiooni funktsioon on järgmine [18]:

$$\mathrm{p} = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)\right)},$$

kus exp on eksponentfunktsioon, $n \geq 0$ on tunnuste arv, β_0 on vabaliige, mis ei ole mõjutatud tunnuste väärtustest, β_n on regressiooni koefitsient mida korrutatakse tunnuse väärtusega x_n. Funktsiooniga leitakse tõenäosus *p*, mille väärtus on $0 < p < 1$. Logistilise regressiooni abil on võimalik sisendile ennustada tulem, arvutades sisendi tunnuste ja koefitsientidega tõenäosus ning valitud lävendi järgi määrata sisendi kuulumist tulemisse.

LASSO logistiline regressioon lisab logistilisele regressioonile L1 regularisatsiooni, mis lisab koefitsientide summade absoluutväärtusega võrdse karistuse (ingl *penalty*), mis piirab koefitsientide suurusi [17]. Tunnuste koefitsiendid leitakse minimeerides järgmise valemi väärtust iteratiivselt [19]:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|,$$

kus eelnevalt määratud karistusparameeter (ingl *tuning* parameeter) on $\lambda \geq 0$, *n* on valimihulk, y_i on *i*-ndal sisendil arvutatud väärtus, *p* on *i*-nda sisendi tunnuste arv ning x_{ij} ja β_j on vastavalt *i*-nda sisendi tunnuse väärtus ja selle koefitsient.

Juhul kui $\lambda = 0$, siis karistus on null ja koefitsientide optimeerimisel sellisel juhul regularisatsiooni ei kasutata. Parameetrit λ suurendades võib tulemust vähem mõjutavate tunnuste koefitsiendid muutuda nulliks, mille tõttu mudeli kompleksus väheneb, sest mudel kasutab vähem tunnuseid.

1.5 Tulemusi iseloomustavad näitajad

Käesolevas töös uuritud riskimudelid ennustavad patsiendi riski vahemikus 0–1 ning vastavalt valitud riski lävendile arvestatakse patsient tulemisse või tulemist välja [16]. Valideerides teame ka patsiendi tegelikku haiguskulgu, mistõttu on võimalik leida mudeli poolt ennustatud tõeseid ja valesid positiivsed ning negatiivsed tulemusi. Tabelis 1 on näha erinevate näitajate nimetusi ning nende väärtuste arvutamise valemeid. Valemites esinevad TP ja FP tähistavad vastavalt tõeseid ning valepositiivseid ennustusi, TN ja FN tõeseid ning valenegatiivseid ennustusi. TP ja TN näitavad seda, mitmele patsiendile ennustas mudeli õige tulemi ning FP ja FN näitavad seda, mitmele patsiendile ennustati vale tulem.

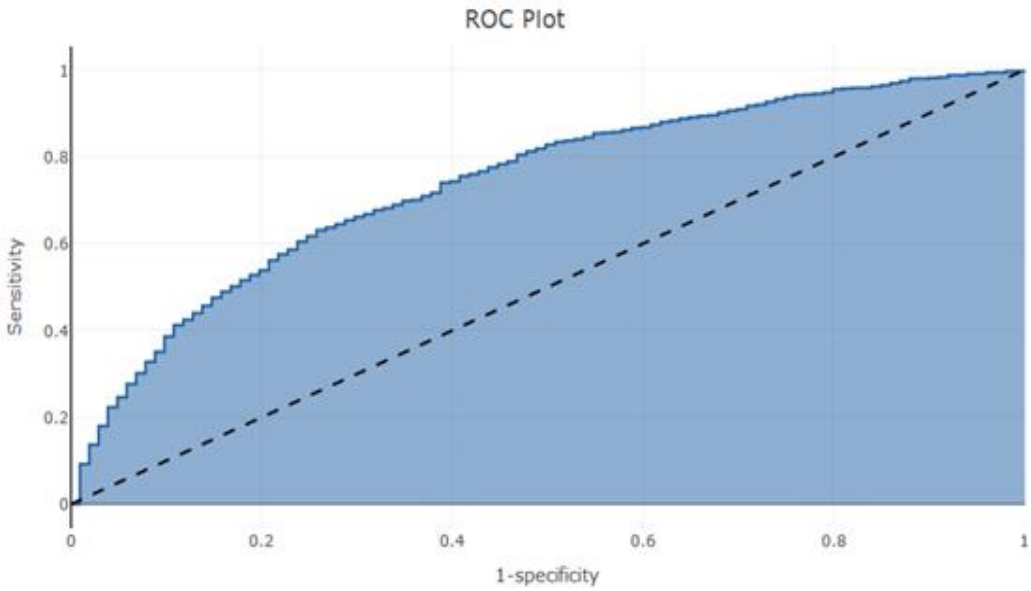
Tabel 1. Näitajate nimetused eesti ja inglise keeles ning valem väärtuse leidmiseks.

Näitaja eesti keeles	**Näitaja inglise keeles**	**Valem näitaja väärtuse leidmiseks**
Õigsus	*Accuracy*	(TP + TN)/(TP + TN + FP + FN)
Tundlikkus	*Sensitivity*	TP/(TP + FN)
Spetsiifilisus	*Specificity*	TN/(TN + FP)
Positiivne ennustuse väärtus	*Positive predictive value*	TP/(TP + FP)

Nende näitajate abil on võimalik leida mudeli diskrimineerimine ning kalibreerimine, mille abil on võimalik mudeli tõhusust analüüsida.

1.5.1 Diskrimineerimine

Diskrimineerimine on mudeli võime määrata kõrgem risk patsientidele, kes kogevad valitud tulemust riskiperioodi jooksul [16]. Selle visualiseerimiseks on *Receiver Operating Characteristics* (ROC) kõvera graafik, mille loomiseks on x-teljele lisatud spetsiifilisus ning y-teljele tundlikkus kõigil võimalikel lävenditel. Näide ROC kõverast on toodud joonisel 1.



Joonis 1. ROC kõvera näide [16].

ROC kõverast on võimalik arvutada AUC[2], mis on näitab, kui hea on mudeli diskrimineerimine. AUC väärtus näitab, kui suure pindala moodustab kõveraalune osa. Mida kõrgem on AUC, seda parem on mudeli diskrimineerimine. Enamikel avaldatud mudelitel on AUC 0,6 kuni 0,8 vahel [16].

Tulemite puhul, mis esinevad valimis haruldaselt, ei pruugi hea AUC näitajaga mudel praktikas kasulik olla, kuna mudel võib siiski valepositiivseid tulemusi ennustada [16]. Kui selliseid mudeleid kasutada ennustusprobleemidega, kus positiivse ennustuse puhul sooritatakse patsiendile kallis või invasiivne protseduur, siis on valepositiivsete ennustamine probleemi lahendamiseks kulukas. Paremaks ülevaateks on mõistlik vaadata peale mudeli AUC ka AUPRC[3] näitajat. AUPRC leitakse sarnaselt AUC näitajale, kuid kõvera loomisel on x-teljel tundlikkus ning y-teljel positiivse ennustuse väärtus iga lävendi puhul [16]. AUPRC näitaja näitab, kui hästi mudel suudab positiivseid juhte ennustada. Kõrgem AUPRC väärtus näitab, et mudel ennustab õigesti positiivseid juhte [20].

AUC puhul on selle tõlgendamine lihtne, sest lähtepunktiks on 0,5, millest suuremad väärtused näitavad head ja väiksemad kehva diskrimineerimist. AUPRC tõlgendamine on aga raskem, sest igal mudelil on erinev lähtepunkt, mis leitakse positiivsete juhtude protsendi järgi [20]. Näiteks, kui tulemisse kuulub 10% patsientidest, siis AUPRC

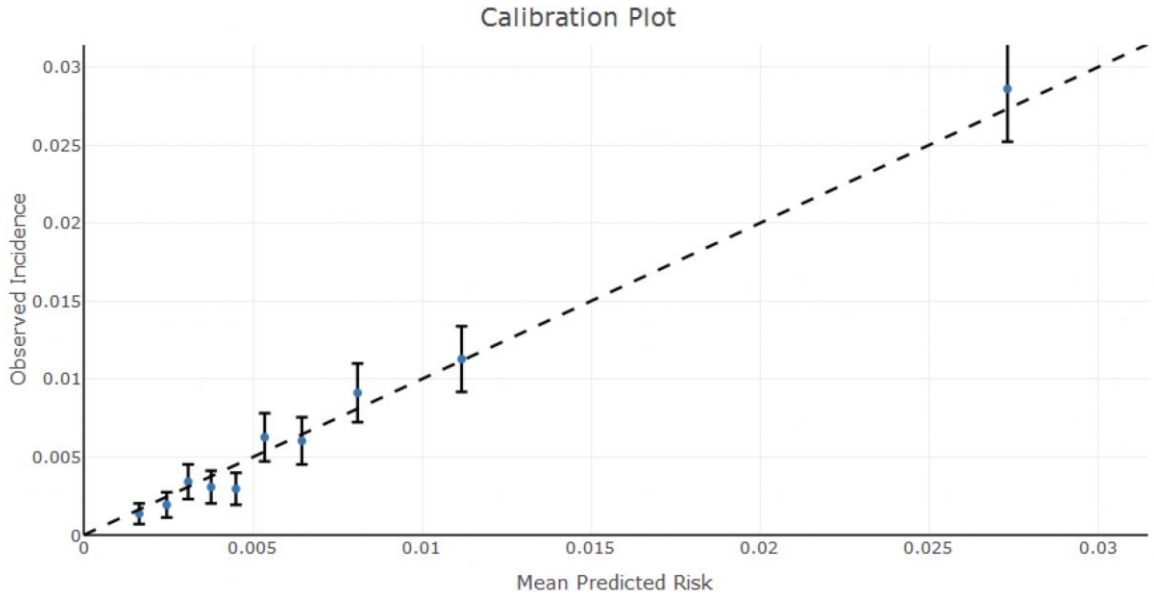
[2] *Area Under Curve*

[3] *Area Under the Precision-Recall Curve*

väärtused, mis on suuremad kui 0,1 on head ja väiksemad on kehvad. Mudelil on väga hea diskrimineerimine, kui AUC ja AUPRC näitajate väärtused on mõlemad kõrged.

1.5.2 Kalibreerimine

Kalibreerimine on mudeli võime määrata tegelikkusega sarnane risk [16]. Kalibreerimist arvutatakse enamasti valimit osadeks jagades ning seejärel iga loodud osa keskmise ennustatud riski ning reaalsuses täheldatud risk suhte abil. Tõenäosused kantakse graafikule nii, et x-teljel on ennustatud tõenäosus ning y-teljel tegelik tõenäosus. Perfektse kalibreerimise puhul langevad graafikule kantud punktid x=y graafile, sest mudel on õigesti ennustanud ning tegelik ja ennustatud tõenäosus on võrdsed [16]. Kõrgema asetusega punktid näitavad, et mudel määrab väiksema riskitõenäosuse, kui tegelikult on. Madalama asetusega punktid näitavad vastupidist ehk mudel hindab riski suuremaks, kui see tegelikult on. Joonisel 2 on toodud näide kalibreerimise graafikust, kus punktid on x=y joonele lähedal ehk tegemist on hästi kalibreeritud mudeliga.



Joonis 2. Näide kalibreerimise graafikust [16].

Ennustatud ning tegeliku riski tõenäosuse suhte paremaks visualiseerimiseks tehakse LOESS graafik, mis luuakse eelnevalt kalibreerimise graafikule kantud punktide koordinaatide abil [21].

Mudeli kvaliteeti ei saa vaid ühe näitajaga otsustada, vaid peab arvestama kõiki eelnevalt mainitud näitajaid ning valede ennustuste kulu praktilises kasutuses.

1.6 Riskimudelite kasutus tänapäeval

1.6.1 Maailmas üldiselt

Riskimudeleid on loodud ka teistel eesmärkidel, mis pole COVID-19 pandeemiaga seotud. Kuigi riskimudelite abil saaks vähendada koormust tervishoiusüsteemile [3], siis praegu ei kasutata andmepõhiseid riskimudeleid väga tihti. Põhjuseid on mitmeid, kuid eelkõige mõjutab kasutamist mudelite ebausaldusväärsus vähese testimise tõttu ning ka see, et mudelite kasulikkust on praktikas vähe uuritud [12]. Sellegipoolest on mõnedes tervishoiusüsteemides sellised riskimudelid kasutusel. Suurbritannia tervishoiusüsteem NHS kasutab kahte ennustavat riskimudelit: *Patients-at-Risk-of-Rehospitalisation* (PARR) ja *Combined Predictive Model* (CPM) [13]. PARR ennustab patsiendi järgneva 12 kuu korduvhospitaliseerimise tõenäosust ja veidi uuem CPM ennustab taashospitaliseerimise asemel hospitaliseerimist üldisemalt ning kasutab suuremat andmete hulka. Walesis ja Šotimaal on kasutusel nendele mudelitele sarnased PRISM (*Predictive Risk Stratification Model)* ja SPARRA (*Scottish Patiens At Risk of Readmission and Admission)* mudelid.

Kasutusel on ka lihtsamaid riskimudeleid, millest üks tuntumaid on APGARi skoor. Selle lõi Dr. Virginia Apgar aastal 1953 ning selle abil määratakse vastsündinule viie kriteeriumi järgi skoor 0 – 10ni [22]. APGARi skoori võib vaadelda kui riskimudelit, kus tunnusteks on pulss, vastsündinu välimus, reageerimine välisele stiimulile, hingamine ning lihastoonus ning neil tunnustel on kindlad koefitsiendid.

1.6.2 COVID-19 riski ennustavad riskimudelid

Tänaseks on arendatud välja mitu COVID-19 ennustavat riskimudelit, kuid suurel osal neist on mõned puudused, mis muudavad mudelid ebausaldusväärseks juba enne välist valideerimist [3]. Suurim probleem nende mudelite puhul on kõrge kallutatus, mis tekib mitme asjaolu tõttu. Wynants et al [3] järeldasid, et uuritud 232st ennustavast riskimudelist olid kõik kõrge või ebaselge kallutatusega.

Esiteks võib mudeli treenimiseks kasutatud treening- ja valideerimisandmete kvaliteet või kvantiteet olla liiga madal. Wynantsi et al [3] analüüsis maailmas leiduvate riskimudelite kohta on avastatud, et 30% uuritud riskimudelitest kasutasid ebasobivaid andmeid. Uuritud mudelitest 27% kasutasid sobimatuid kriteeriumeid valimi loomisel, mistõttu mudeli arendusel kasutatud andmestik ei esindanud õigesti ennustamise valimit. Veel toodi välja,

et 75% mudelitest kasutasid ainult ühe riigi patsientide andmeid ning ainult 18% mitmest riigist pärinevaid andmeid.

Teiseks probleemiks on mudeli arendusprotsessi puudulik kirjeldus. 73% samas metaanalüüsis uuritud uurimistöödest kasutasid ebasobilikke või ebaselget metoodikat andmete kogumisel. Veel leiti, et paljudel mudelitel oli ülesobitamise risk ehk mudelid ei ennusta piisavalt üldiselt, et uute andmestikega hästi töötada.

Wynants et al [3] arvates ei ole hetkel loodud mudelite kasutus soovitatud, sest enamike mudelite arendus on halvasti raporteeritud või kõrge kallutatusega, mistõttu head tulemused võivad olla petlikud. Mudeleid on ka vähe väliselt valideeritud, mistõttu ei saa kindel olla mudelite tõhususes erinevate populatsioonide seas.

1.7 Mudelite arenduse standardiseerimine ning OHDSI

Peale mudelite vähese välise valideerimise on probleemiks ka see, et paljud riskimudelid ei ole arendatud kindla standardi järgi, mis muudab mudeli välise valideerimise ning tõhususe analüüsi keerulisemaks ning ebatäpsemaks [16]. Nende probleemide lahenduseks loodi Ameerika Ühendriikides 2007. aastal OMOP[4], mis oli avaliku- ja erasektori vaheline koostöö ning mille ülesandeks oli leida viise, kuidas tervishoiu andmeid kasutada meditsiiniliste toodete turvalisuse analüüsimiseks [23]. OMOPi meeskond sai kiiresti aru, et analüüside tegemine erinevates formaatides andmestike peal on aeganõudev protsess. Lahenduseks loodi OMOP-CDM (edaspidi CDM[5]), mis standardiseerib tervisandmete esitusformaati. Tänu sellele saab statistilise analüüsi koodi taaskasutada erinevate andmekogude peal, näiteks CDM-kujul andmehulgal treenitud mudelit saab väga lihtsalt teiste CDM-kujul andmetega valideerida [24].

Andmekogu CDM-formaati üleviimine ei ole lihtne protsess, sest esialgsed andmed on kogutud erinevatel eesmärkidel, mistõttu võivad kindlad andmeväljad olla ebakvaliteetselt täidetud või puudulikud. Selle tõttu ei pruugi iga patsiendi kohta samad andmed olla saadaval, mis vähendab mudelite ennustamiste tõhusust.

Peale OMOP projekti lõppemist loodi OHDSI[6] kommuun, mis soovis jätkata OMOPi missiooni muuta meditsiiniinformaatika uurimist läbipaistvamaks ning standardiseerituks.

[4] *Observational Medical Outcomes Partnership*

[5] *Common Data Model*

[6] *Observational Health Data Sciences and Informatics*

OHDSI eesmärgiks on kogu teadusliku uurimise protsess muuta avalikuks, lihtsustades sellega koostööd teadlaste vahel, kes püüavad erinevaid tervishoiu aspekte analüüsida. Selle nimel on OHDSI loonud mitu raamistikku ning abistavat tarkvara, et soodustada terviseandmete analüüsi ning nende põhjal tehtavaid ennustusi.

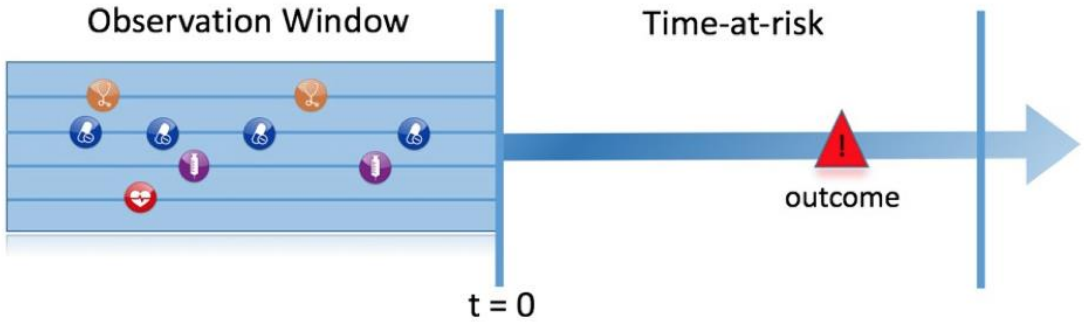
1.8 Masinõppe mudeli loomise protsess OHDSI keskkonnas

Käesolevas töös kasutatakse OHDSI uurijate poolt arendatud riskimudeleid, mille tõttu on tarvis lähemalt kirjeldada mudeli loomise protsessi.

Esimeseks sammuks on ennustusprobleemi sõnastamine. Selleks, et ennustusprobleemi defineerida, peab arvestama nelja tegurit:

- valimi valikut,
- tulemi valikut,
- ennustava mudeli ja tunnuste valikut,
- ennustuse eesmärgi sõnastamist.

Valimiks (ingl *target cohort*) nimetatakse patsientide hulka, kelle peal soovitakse ennustust sooritada. Valimi kohordi loomiseks on vaja defineerida kindlad kriteeriumid ning valimisse võetakse kõik need patsiendid, kes nendele kriteeriumitele vastavad [25]. Kriteeriumeid võibolla üks või mitu. Üheks kindlaks kriteeriumiks on joonisel 3 välja toodud jälgimisaeg (ingl *observation window*), mis määrab ära selle, kui kaua peab kindla patsiendi kohta olema andmeid andmestikus, et teda valimisse lisada [16].



Joonis 3. Illustratsioon ennustusprobleemist [16].

Tulem (ingl *outcome cohort)* on patsientide hulk, kellel on kindla riskiperioodi ajal esinenud soovitud tulemus. Riskiperioodiks (ingl *time-at-risk*) loetakse aega peale indekskuupäeva t ehk kuupäeva, mil patsienti loetakse valimisse (illustreeritud joonisel 3).

Tähtis on ka mudeli valik. Vastavalt ennustuse eesmärgile on mõistlik kasutada erineva keerukusega masinõppe algoritme. Mudeli treenimiseks on ka tunnuste valik oluline, kuna sellest sõltub mudeli komplekssus ja tõhusus on sellest mõjutatud [26]. Veel võivad tunnuste valikut mõjutada kasutatavad andmed, sest andmestikus mingi tunnuse puudumisel ei saa seda tunnust kasutada.

Mudeli loomisel on oluline eesmärgi defineerimine, sest ilma praktilise kasuta mudelit ei ole otstarbekas luua. Eesmärkideks võibolla näiteks arstide aitamine meditsiiniliste otsuste tegemisel.

Mudelite arendust ja analüüsimist lihtsustab protsessi avatud raporteerimine, mille jaoks on loodud TRIPOD nimekiri, kus on kirjas aspektid, mida mudelite arendusel ning uurimisel silmas pidada [27]. Nimekirjas on toodud välja näiteks algandmete, mudeli arenduse protsessi, tulemuste ning ka uuringul ilmnenud piirangute kirjeldamine. Nende juhiste järgimine tõstab arendatud mudeli usaldusväärsust, sest täpse ning avatud arendusprotsessi kirjelduse tõttu on võimalik veenduda mudeli tulemuste ehtsuses.

2. Metoodika

2.1 Valideeritavate mudelite valik

Käesolevas töös valideeritakse OHDSI uurijate Williams et al [26] poolt välja arendatud COVID-19 riskimudeleid. Valik sai tehtud järgmistel põhjustel:

- OHDSI mudelitel on ennustamisel head tulemused.
- Mudelitele on tehtud juba eelnevalt erinevatel andmestikel välist valideerimist ehk selles töös sooritatud valideerimine ei oleks esmakordne.
- Mudelid on treenitud CDM-formaadis andmetega ning kindlate defineeritud kohortidega, mis muudab valideerimise protsessi lihtsamaks, kuna ei pea andmetöötlusega tegelema.
- Tänu OHDSI läbipaistvusele on lihtsam erinevaid mudeleid analüüsida ning võrrelda.
- Kasutatud on TRIPOD nimekirjas välja toodud suuniseid uuringu läbipaistvuse tagamiseks.

Arendatud mudeleid on mitu ning need ennustavad erinevaid tulemusi, selle lõputöö raames valideeriti OHDSI poolt arendatud COVER skoori arvutavaid mudeleid. Mudelid ennustavad iga patsiendi kohta kolme tulemust [26]:

1. Hospitaliseerimine kopsupõletikuga 30 päeva peale indekskuupäeva.
2. Hospitaliseerimine kopsupõletikuga, mis nõudis intensiivravi või lõppes surmaga 30 päeva peale indekskuupäeva.
3. Surm 30 päeva peale indekskuupäeva.

Intensiivravi alla on arvestatud ventileerimist, intubeerimist, trahheostoomiat ning kehavälist membraanoksügenatsiooni [26]. Neid ennustavaid mudeleid on kolme tüüpi: andmepõhine mudel, vanuse/soo mudel ning andmepõhisest mudelist tuletatud lihtsustatud COVER mudelid. Indekskuupäevaks on perearsti, erakorralise meditsiini või ambulatoorse ravi vastuvõtu kuupäev.

Mudelid on treenitud kuue erineva riigi terviseandmetega: Ameerika Ühendriigid, Lõuna-Korea, Hispaania, Austraalia, Jaapan ning Madalmaad. Andmed on võetud perearstide ja haiglate vastuvõttudest, kindlustusnõuetest ning elektroonilistest terviseandmetest [26]. Seejärel on need viidud üle CDM-formaati, sest standardiseeritud formaadi kasutamine lihtsustab nii mudelite arendust kui ka välist valideerimist [26]. Treenimisel kasutatud

patsientide andmed olid anonüümsed ning isikuga mitte seotavad ja seetõttu ei olnud vaja nende nõusolekut andmete kogumisel. Suurem osa andmekogumitest ei sisalda COVID-19 patsiente, kuid Williams et al [26] oletasid, et sarnaste sümptomite tõttu võib treenimisel kasutada ka gripipatsientide andmeid[7]. See tuli eriti kasuks selle tõttu, et mudeli treenimise ajal (pandeemia esimestel kuudel) oli COVID-19 patsientide andmeid veel ainult väikestes kogustes. Valitud strateegia tõttu oli võimalik mudeleid kordades rohkemate andmetega treenida. Kõikidest andmetest võeti treenimiseks valimist juhuslikult 150 000 patsiendi suurune andmehulk, et efektiivselt mudeleid treenida, kuid säilitades tulemi proportsioone [26].

Mudeli arenduseks ning valideerimiseks loodi erinevad kohordid. Arenduseks võeti kohorti patsiendid, kes olid vanemad kui 18 aastat; olid tervishoiuasutuses gripisümptomitega vastuvõtul vähemalt 365-päevase jälgimisajaga ja indekskuupäevale eelneval 60 päeval ühegi sümptomita. Valideerimiseks loodud kohorti võeti patsiendid, kes olid samuti üle 18-aastased ja COVID-19 diagnoosiga ning täitsid treenimise kohordiga samasid jälgimisperioodi tingimusi [26].

Mudelite välist valideerimist sooritati kahte moodi, esiteks gripihaigetega andmekogude peal, et testida mudelite tõhusust treeningandmetega sarnaste andmetega, ning eraldi COVID-19 andmekogude peal. Teist tüüpi valideerimisel saadud tulemused olid võrreldavad grippi põdenud patsientide andmetega, mis tõestas ka Williamsi et al hüpoteesi, et mudelite sooritus COVID-19 andmetel on võrreldav gripihaigete andmetega [26].

2.1.1 Andmepõhine ja vanus/sugu mudelid

Esimene mudel on andmepõhine mudel (ingl *data-driven model*), mis võtab sisendiks patsiendi terve haigusloo [26]. Selle tõttu on tunnuste arv väga suur- hospitaliseerimise ennustamisel 521, intensiivravi puhul 349 ning surma puhul 205 tunnust. Suur tunnuste arv võib mudeli täpsust tõsta, kuid praktilises kasutuses ei pruugi iga patsiendi kohta kõik vajalikud tunnused andmestikus olemas olla. Sellepärast oleks mudeli erinevate maailma tervishoiusüsteemidega integreerimine keeruline. Mudeli komplekssuse vähendamiseks loodi ka kaks vähemate tunnustega mudelit.

Teist tüüpi mudel ennustab ainult kahe tunnuse alusel: vanuse ning soo järgi. Kuna haiguse kulgu mõjutab palju patsiendi vanus [9], siis on mudeli tõhusus võrreldav komplekssema

[7] Kasutati andmeid Optum© De-Identified Clinformatics© Data Mart andmebaasi [26].

andmepõhise mudeliga. Tulemused tulid siiski halvemad ning komplekssuse ja mudeli täpsuse kompenseerimiseks loodi lihtsamad COVER[8] mudelid.

Joonisel 4 on illustreeritud andmepõhise mudeli arendusprotsessi. Sarnane protsess on nii andmepõhisel kui ka vanus/sugu mudelil, kus ainuke erinevus on tunnuste loomises.

Joonis 4. Mudeli arenduse protsess.

Pärast mudeli treenimist luuakse automaatselt valideerimispakett, mis sisaldab endas mudelite sätteid (ennustamiseks vajalike tunnuste koefitsiendid) ning valideerimiseks vajalikke kohorte. Välise valideerimise protsess on lähemalt kirjeldatud peatükis 2.4.

[8] *COVID-19 Estimated Risk*

2.1.2 COVER mudelid

COVER skoori ennustavad mudelid määravad üheksa patsiendi tunnuse järgi riskiskoori, mis vastavalt mudelile ennustab patsiendi hospitaliseerimise (COVER-H), intensiivravi vajamise (COVER-I) või suremise (COVER-F) tõenäosuse.

Mudel arvestab järgmisi ennustajaid:

- Vanus
- Sugu
- Vähk
- Krooniline obstruktiivne kopsuhaigus
- Diabeet
- Südame-veresoonkonna haigused
- Hüpertoonia ehk kõrgvererõhutõbi
- Hüperlipideemia
- Neeruhaigused

Mudel töötab loogikal, et igal tunnusel on mingi koefitsient ning patsiendile kehtivate tunnuste koefitsientide summad moodustavad COVER-H, COVER-I ning COVER-F skoorid. Riskiskoori teisendamiseks vastavaks tõenäosuseks p kasutatakse logistilist funktsiooni

$$p = \frac{1}{1 + \exp\left(\frac{score - 93}{10}\right)},$$

kus *score* on patsiendile kehtivate ennustajate järgi summeeritud väärtus ning exp on eksponentfunktsioon.

COVER mudeli ennustajate skoori väärtused on loodud andmepõhise mudeli ning arstide koostööga. Andmepõhiselt treenitud mudelit analüüsisid arstid, kes tuvastasid need ennustajad, mis kõige rohkem tulemust mõjutasid [26]. Seejärel loodi üldistavad kategooriad, et vähendada ennustavate tunnuste arvu COVER mudeli jaoks ning treeniti originaalsete andmetega LASSO logistilise regressiooni mudel. Tulemuseks saadud mudelist võeti tunnuste koefitsiendid, mis korrutati kümnega ning ümardati lähima täisarvuni. Need täisarvud ongi COVER skoori loomisel ennustajate väärtusteks.

Sellel lihtsamal mudelil on mitu eelist. Esiteks on mudeli treenimine palju kiirem ning nõuab vähem arvutusressurssi, kuna ennustatavaid tunnuseid on ainult üheksa. Samas on mudel täpsem kui eelnevalt mainitud kahe tunnusega [26]. Teine eelis on ka mudeliga ennustuste tegemise lihtsus. Skooridel põhinevaid mudeleid saab ka ilma arvutita kasutada ning haigla personal saab kiiresti skoori arvutades teada patsiendi riski.

2.2 Valideerimiseks kasutatud Eesti terviseandmed

Valideerimiseks kasutatud andmed on saadud Terviseameti nakkushaiguste infosüsteemist NAKIS, Eesti Haigekassast ja Eesti surmapõhjuste registrist. Andmed on kogutud Uusküla et al [28] jooksva projekti raames. Valimi hulgas on nii laboratoorselt (RNA testi alusel) kui ka kliiniliselt (haigustunnuste esinemisel või kontakt COVID-19 haigega) diagnoositud COVID-19 juhud. Kokku on andmestikus umbes 2600 COVID-19 diagnoosiga patsienti ning 6400 COVID-19 diagnoosita kontrollisikut. Andmestikus on patsiendid ajavahemikus 25.02.2020–14.09.2020. Andmed on eelnevalt anonümiseeritud ning üle viidud CDM-formaati andmekogu koostaja poolt.

2.3 Valideerimiseks kasutatud kood

Mudelite valideerimiseks on OHDSI loonud kaks R programmeerimiskeskkonna projekti, mille leiab OHDSI „Covid19PredictionStudies“ Githubi repositooriumist „CovidSimpleModels“ ja „HospInOutpatientVal“ kaustadest [29]. „CovidSimpleModels“ kaustas on lihtsamad 9 ennustatava tunnusega mudelid ning kaustas „HospInOutpatienVal“ leiduvad andmepõhine ning vanus/sugu mudelid.

Kood on kirjutatud programmeerimiskeeles R (versioon 3.6.3) ning mudelite valideerimiseks on kasutatud OHDSI poolt välja töötatud PLP[9] paketti (versioon 3.0.16) [26]. Koodi jooksutamiseks on kasutatud RStudio versioon 1.3.1056 keskkonda.

2.4 Koodi jooksutamise kirjeldus

Koodi jooksutamiseks vajalik RStudio ja Eesti terviseandmetega andmebaas asetsesid Tartu Ülikooli serveris. Eelnevalt mainitud „Covid19PredictionStudies“ repositoorium [29] klooniti serverisse ning avati „HospInOutpatientVal“ ja „CovidSimpleModels“ kaustades olevad .Rproj failid. Kõik „Covid19PredictionStudies“ olevad kaustad järgivad PLP

[9] *Patient-Level Prediction*

valideerimispakettide projektide struktuuri, mis on automaatselt genereeritud mudeli arenduse jooksul [16].

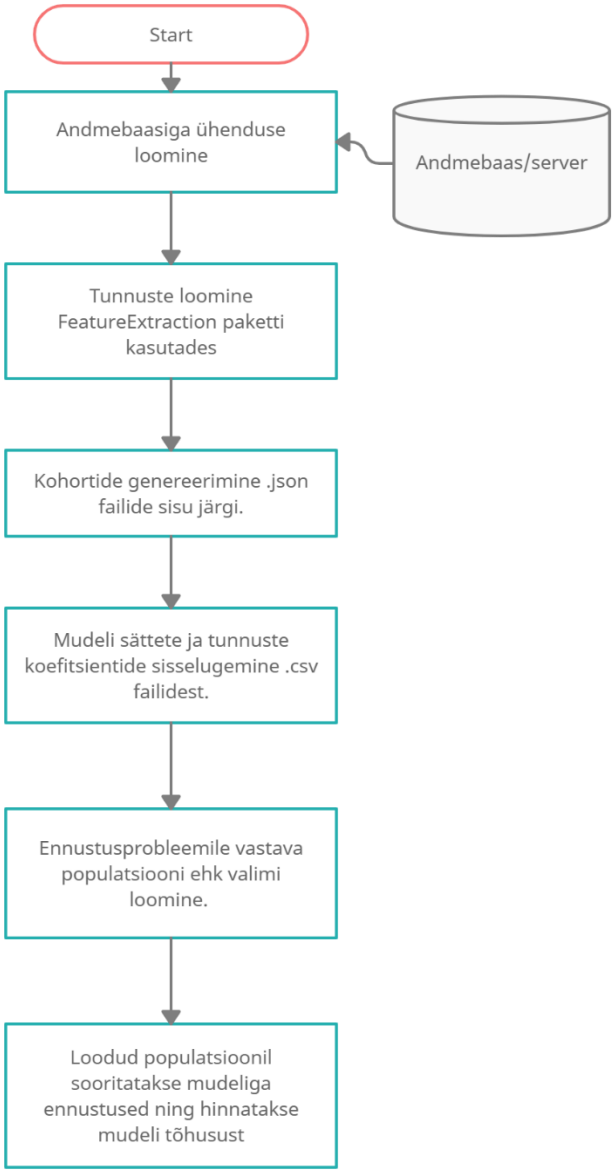
Igas projektis leidub kaustas „extra“ fail nimega „codeToRun.R“, mille jooksutamisel saab mudelite valideerimist käivitada. OHDSI juhiste järgi pidi lisama koodile andmebaasiga ühendamiseks vajaliku konfiguratsiooni selleks ettenähtud muutujatele.

2.4.1 Valideerimise protsess koodis

Kahes projektis on erinev koodi toimimine, kuna lihtsamal COVER mudelitel on mudeli sätted CSV-failides, mitte PLP paketi *plpResult* objektina. See on selletõttu, et CSV-failid kasutavad vähem mälu kui *plpResult* objektid, mis on vajalik „CovidSimpleModelsi“ COVER mudelite komplekssuse vähendamise eesmärk täitmiseks. „CovidSimpleModelsi“ „codeToRun.R“ failis leiduvat *execute* funktsiooni käivitades toimub järgmine protsess:

1. Käivitades eeldatakse, et failis olevad muutujad on väärtustatud korrektselt OHDSI juhiste järgi.
2. Kuna mudel loetakse sisse CSV-failist, siis puudub sellel tunnuste kohta informatsioon, mis muidu oleks *plpResult* objektist kättesaadav. Seetõttu jooksutatakse enne kohortide loomist PLP paketis leiduv „FeatureExtractioni“ paketi *createCovariateSettings* funktsioon, millega luuakse vajalikud tunnuste sätted.
3. Seejärel jooksutatakse *createCohorts* funktsiooni, mis loob serverisse valideerimiseks vajalikud kohordid. Kohortide loomist kirjeldavad JSON-failid, mis asuvad „inst/cohorts“ kaustas. Andmebaasist andmete kättesaamiseks vajalikud päringud on kaustas „inst/sql“.
4. Järgmisena loetakse mudelite sätted kaustast „inst/settings“ CSV-failidest sisse.
5. Seejärel luuakse kohortide abil valideeritava populatsiooni objekt ning ennustatakse sellel tulemusi.
6. Valideerimistulemused salvestatakse muutuja *outputFolder* järgi määratud kausta.

Seda protsessi illustreerib joonis 5.



Joonis 5. Mudeli väliseks valideerimiseks sooritatud sammud.

Projekti „HospInOutpatienVal“ mudelite puhul on protsess sarnane, kuid kuna mudelid loetakse sisse *plpResultist* objektist, kus on tunnused ja mudeli sätted olemas, siis nende loomise samme 2 ja 4 ei läbita.

Valideerimistulemusi saab visualiseerida OHDSI „Covid19CoverPrediction“ repositooriumis leiduva Shiny rakenduse projekti kasutades [30]. Selles projektis leiduvad teistel andmestikel sooritatud välised valideerimised, mille erinevad tõhususe näitajad on visualiseeritud. Lisades Eesti terviseandmetel sooritatud valideerimiste tulemused kausta „data“, visualiseeritakse samuti nende tõhususe näitajad ning luuakse erinevaid graafikuid, näiteks LOESS graafik kalibreerimise ja ROC kõvera diskrimineerimise analüüsimiseks.

2.5 Eesti terviseandmetel treenitud mudel

Töö eesmärgiks on väliselt valideerida juba OHDSI poolt treenitud COVID-19 ennustavaid mudeleid, kuid kuna OHDSI „Covid19PredictionStudies“ repositooriumis on olemas ka vajalikud PLP projektid uue mudeli mugavaks treenimiseks, siis on valideerimistulemustega võrdlemiseks treenitud ka täiesti uus mudel, kasutades ainult Eesti terviseandmeid. Täpsemalt asub mudeli treenimise R projekt „HospitalizationInSymptomaticPatients“ kaustas. Treenimise protsessi on kirjeldatud peatükis 2.1.1. Treenitud mudelitüüp on andmepõhine mudel.

Mudeli valimiks on COVID-19 või gripi sümptomitega patsiendid. Treenimiseks on võetud kogu valimist 75% ning sisemiseks valideerimiseks 25%. Valideerimisel on kasutatud ristvalideerimist (ingl *cross-validation*), mis tähendab seda, et valim on tükeldatud võrdseteks alamhulkadeks ning iga alamhulga puhul testitakse ülejäänud osadel treenitud mudelit [16]. Ristvalideerimine on hea selleks, et väiksemal andmehulgal vähendada mudeli ülesobitust, kuna sellega väheneb võimalus, et treenimishulka satuvad erandlike tunnustega patsiendid. Mudel ennustab peatükis 2.1. kirjeldatud tulemeid.

Kuna loodud mudel on treenitud väiksel andmehulgal ning väliselt seda valideeritud ei ole, siis mudelit on mõistlik kasutada ainult selle töö kontekstis valideerimistulemustega võrdlemiseks.

3. Tulemused

3.1 Mudelite valideerimiste tulemused

Kokku sooritati nelja valimi, kolme mudeli ning kolme tulemiga 36 ennustust. Mudeli tõhususe analüüsimiseks on kasutatud peatükis 1.5. mainitud AUC ja AUPRC näitajaid ning kalibreerimist. Näitajate valik on tehtud selle tõttu, et Shiny rakenduse näitajate visualiseerimist on mugav analüüsimiseks kasutada ja neid näitajaid kasutasid ka Williams et al [26] COVER mudelite tõhususe analüüsimises.

Tabelitest 2, 3 ja 4 on näha, et COVER mudelite AUC tulemused on vahemikus 0,746–0,930, andmepõhise mudeli puhul 0,736–0,874 ja vanus/sugu mudelil 0,746–0,932. AUPRC on vastavalt 0,047–0,29, 0,041–0,315 ning 0,036–0,312. Selliste AUC ja AUPRC näitajatega on mudel võrreldav nii sisemise valideerimisega kui ka teistel andmestikel tehtud väliste valideerimistega ning suuri erinevusi pole [26].

Märkimisväärseks tulemuseks on see, et andmepõhise mudeli tõhusus pea iga erineva ennustuse puhul on sama või vähem tõhus lihtsamate COVER ja vanus/sugu mudelitega võrreldes. Williams et al [26] poolt mudelite arendamisel sooritatud sisemisel valideerimisel olid tulemused vastupidised, nimelt oli AUC ning AUPRC kõrgem just andmepõhise mudeli puhul. Ka välistes valideerimistes on üldiselt andmepõhise mudeli tulemused pigem paremad.

Huvitav on ka see, et COVER mudeli ning vanus/sugu mudeli näitajad on üksteisele väga lähedased. Sellest võib järeldada, et tulemust mõjutavad kõige rohkem vanus ning sugu ja valimisse kaasatud patsientidel ei olnud palju kaasnevaid haiguseid. Sellegipoolest on COVER mudeli diskrimineerimine teiste mudelitega võrreldes keskmiselt parem.

Tabel 2. Mudelite tulemused patsiendi kopsupõletikuga hospitaliseerimise ennustamisel.

Valimi kirjeldus	**Inimeste arv valimis**	**Tulemisse ennustatud inimeste arv**	**Ennustatud inimeste protsent (%)**	**COVER-H mudel**		**Andmepõhine mudel**		**Vanus/sugu mudel**	
				AUC	**AUPRC**	**AUC**	**AUPRC**	**AUC**	**AUPRC**

Patsiendid COVID-19 või gripi sümptomitega	3443	271	7,871	0,795	0,266	0,784	0,238	0,791	0,269
Patsiendid COVID-19 või gripi sümptomitega aastal 2020	1648	178	10,801	0,755	0,283	0,745	0,263	0,755	0,307
Patsiendid COVID-19 diagnoosi või sümptomitega aastal 2020	1587	176	11,09	0,75	0,29	0,74	0,264	0,75	0,312
Patsiendid COVID-19 diagnoosiga aastal 2020	1565	174	11,118	0,746	0,279	0,736	0,256	0,746	0,308

Tabel 3. Mudelite tulemused patsiendi intensiivravi vajaduse ennustamisel.

Valimi kirjeldus	**Inimeste arv valimis**	**Tulemisse ennustatud inimeste arv**	**Ennustatud inimeste protsent (%)**	**COVER-I mudel**		**Andmepõhine mudel**		**Vanus/sugu mudel**	
				AUC	**AUPRC**	**AUC**	**AUPRC**	**AUC**	**AUPRC**
Patsiendid COVID-19 või gripi sümptomitega	3443	25	0,7261	0,906	0,047	0,839	0,041	0,889	0,036
Patsiendid COVID-19 või gripi sümptomitega aastal 2020	1648	21	1,2743	0,885	0,066	0,795	0,054	0,862	0,050

Patsiendid COVID-19 diagnoosi või sümptomitega aastal 2020	1587	21	1,3233	0,883	0,068	0,793	0,054	0,859	0,051
Patsiendid COVID-19 diagnoosiga aastal 2020	1565	21	1,3419	0,883	0,069	0,793	0,057	0,859	0,051

Tabel 4. Mudelite tulemused patsiendi surma ennustamisel.

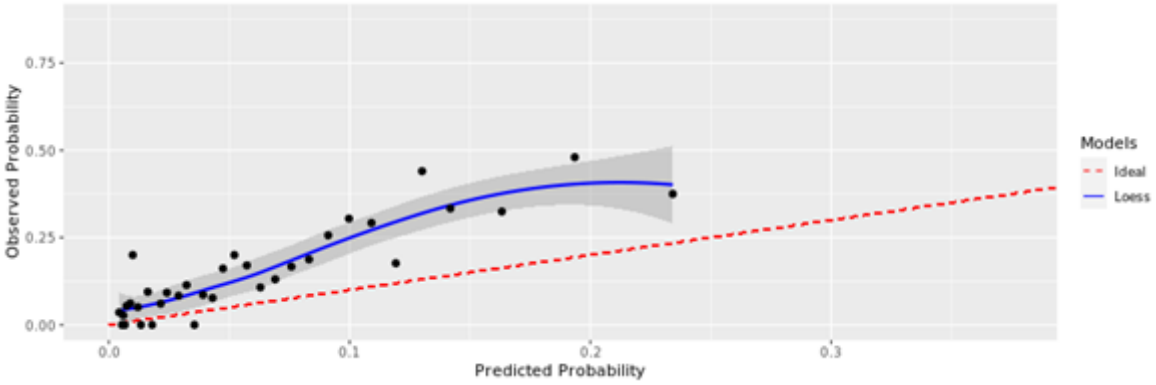
Valimi kirjeldus	**Inimeste arv valimis**	**Tulemisse ennustatud inimeste arv**	**Ennustatud inimeste protsent (%)**	**COVER-F mudel**		**Andmepõhine mudel**		**Vanus/sugu mudel**	
				AUC	**AUPRC**	**AUC**	**AUPRC**	**AUC**	**AUPRC**
Patsiendid COVID-19 või gripi sümptomitega	3443	30	0,8713	0,930	0,073	0,874	0,078	0,932	0,077
Patsiendid COVID-19 või gripi sümptomitega aastal 2020	1648	27	1,638	0,904	0,091	0,846	0,095	0,907	0,097
Patsiendid COVID-19 diagnoosi või sümptomitega aastal 2020	1587	27	1,701	0,902	0,092	0,845	0,097	0,906	0,099

Patsiendid COVID-19 diagnoosiga aastal 2020	1565	27	1,725	0,902	0,093	0,844	0,099	0,905	0,1

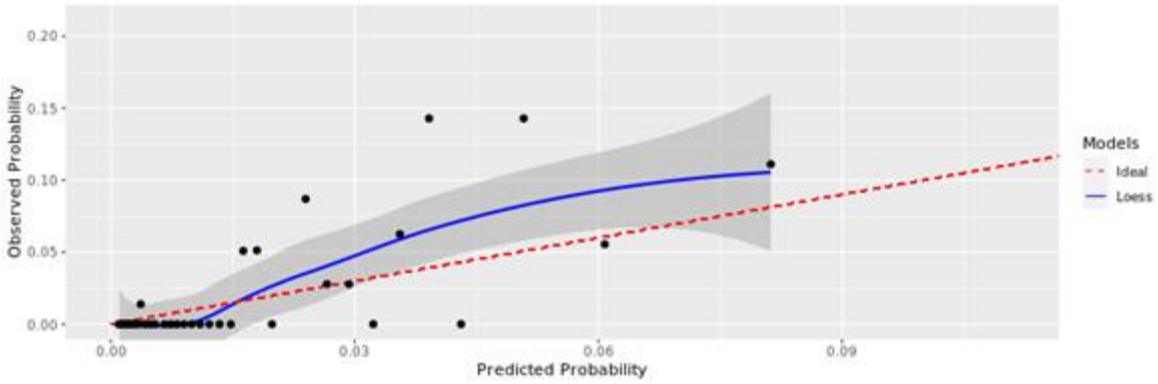
Kuna käesoleva töö eesmärgiks oli väliselt valideerida mudelit, mis teoreetiliselt võiks leida kasutust praktikas, siis on mõistlik lähemalt analüüsida ainult COVER mudelite tõhusust. Seda selletõttu, et andmepõhine mudel on praktiliseks kasutamiseks liiga kompleksne ning vanus/sugu mudel erineb COVER mudelist vaid 7 tunnuse võrra, kuid diskrimineerimine on üldiselt parem. Valimeid, millel COVER mudeleid rakendati, on neli, kuid praktilise kasutuse kohta annab kõige täpsema ülevaate valim, kuhu kuuluvad COVID-19 diagnoosiga patsiendid. Nende põhjuste tõttu ülejäänud mudelite ning valimite väliseid valideerimisi lähemalt käesolevas töös ei analüüsita.

Joonistel 6, 7 ja 8 on näha erinevate COVER mudelite kalibreerimine eelnevalt defineeritud valimil. Jooniste x-teljel on mudeli poolt ennustatud risk ja y-teljel patsientide tegelik risk. Sinisel graafikul on kujutatud LOESSi graafik ning punane katkendlik joon näitab perfektset kalibreerimist.

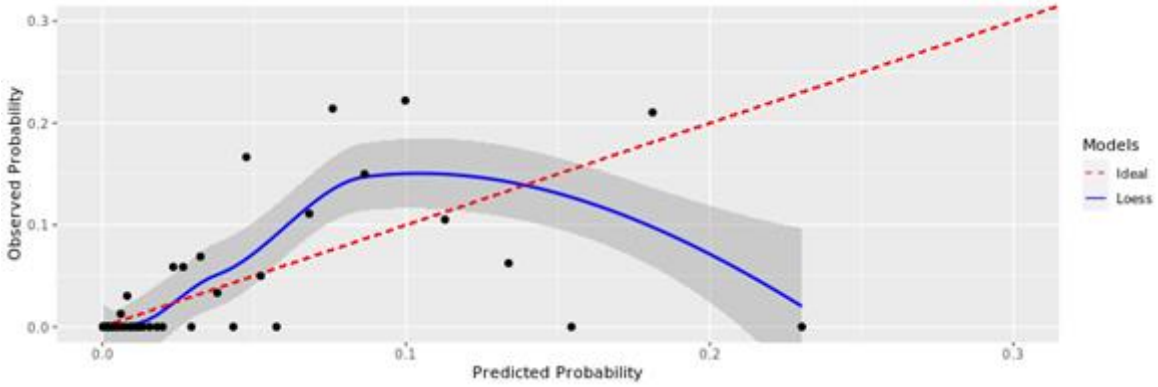
Nii COVER-H kui ka COVER-I kalibreerimise graafikutest on näha, et need mudelid ennustavad madalamat riski kui tegelikkuses risk kujunes. COVER-F aga ennustab väiksemate ennustatud väärtuste juures tegelikkusest väiksemat riski, kuid suuremate ennustatud väärtuste puhul tegelikkusest suuremat riski. Varieeruvus võibolla tingitud sellest, et tulemisse ennustati väike patsientide hulk ehk mõni erand võis kalibreerimist tugevalt mõjutada.



Joonis 6. COVER-H mudeli kalibreerimine Eesti terviseandmetel.



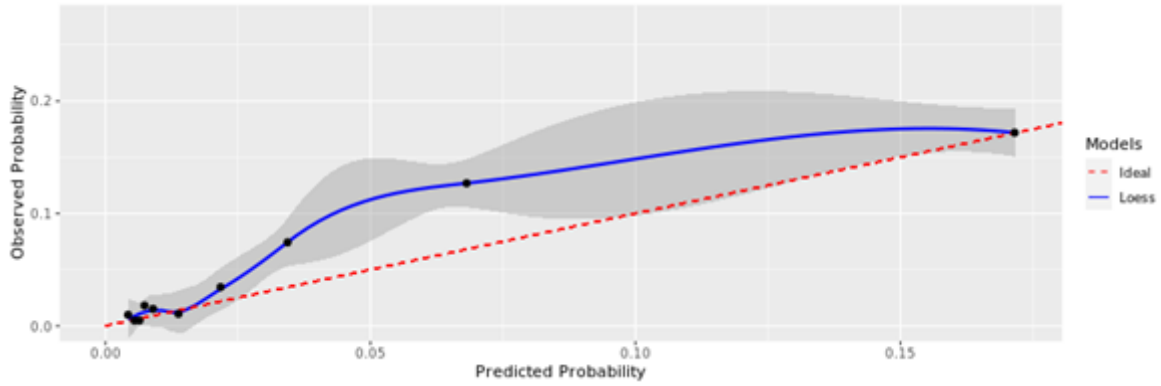
Joonis 7. COVER-I mudeli kalibreerimine Eesti terviseandmetel.



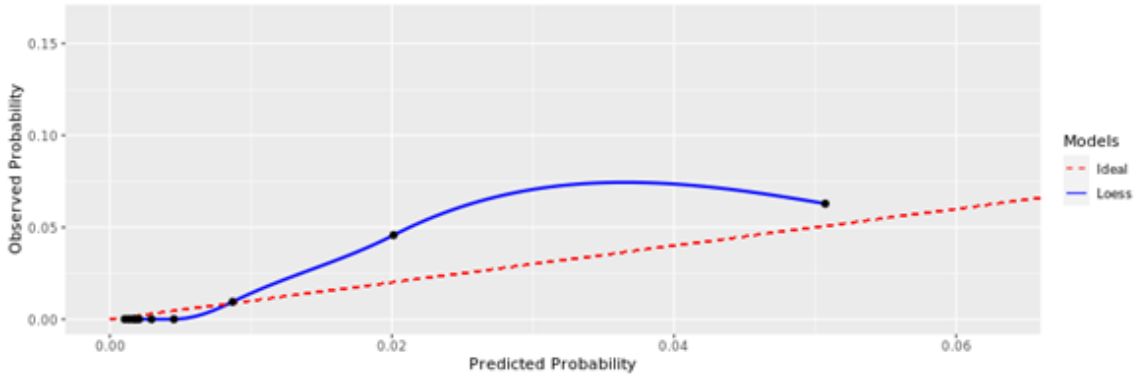
Joonis 8. COVER-F mudeli kalibreerimine Eesti terviseandmetel.

COVER mudeleid on väliselt valideeritud ka teiste COVID-19 diagnoosiga patsientidega andmekogude peal. Välisel valideerimisel saavutas parima diskrimineerimise Lõuna-Korea

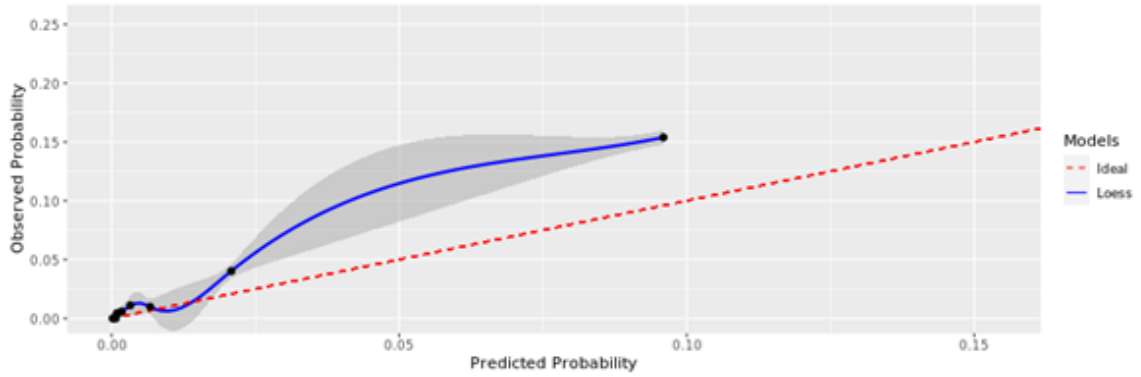
tervisekindlustuse nõuetest koosnev HIRA[10] andmekogu [26]. Joonistel 9, 10 ja 11 on COVER mudelite kalibreerimised HIRA andmestikul.



Joonis 9. COVER-H mudeli kalibreerimine HIRA andmestikul.



Joonis 10. COVER-I mudeli kalibreerimine HIRA andmestikul.



Joonis 11. COVER-F mudeli kalibreerimine HIRA andmestikul.

HIRA andmestikul valideeritud COVER-H ja COVER-I mudelite kalibreerimised on sarnased Eesti terviseandmetel valideeritud mudelite kalibreerimisega, kus samamoodi

[10] *Health Insurance and Review Assessment*

ennustatakse liiga madalat riski. Erinevus on aga COVER-F kalibreerimistel, sest HIRA andmestikul ennustab COVER-F mudel ainult reaalsest riskist madalamat riski. Selle põhjuseks võibolla see, et andmed on treeningandmetele suhteliselt sarnased ning ei esine paari erandliku patsiendi poolt põhjustatud kallutatust.

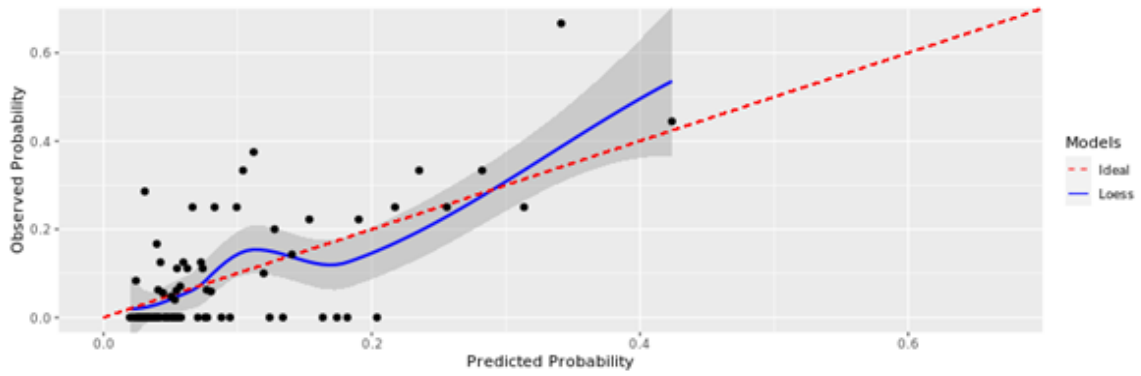
3.2 Eesti terviseandmetega treenitud mudeli tulemused

Eesti terviseandmetega treenitud mudeli diskrimineerimist erinevate tulemite ennustamisel on välja toodud tabelis 5. Valimi kohorti kuulus 3433 patsienti. Näitaja AUC kohta on sulgudes välja toodud erinevate ristvalideerimiste tulemused. Ristvalideerimiste keskmine tulemus jääb sulgudest välja. Eesti terviseandmetega treenitud mudeli AUC ja AUPRC väärtused on väga head, kuid kuna tegemist on väikse treening- ja valideerimishulgaga, siis on võimalik, et mudel on ülesobitunud ning välise valideerimise puhul oleks tulemused halvemad.

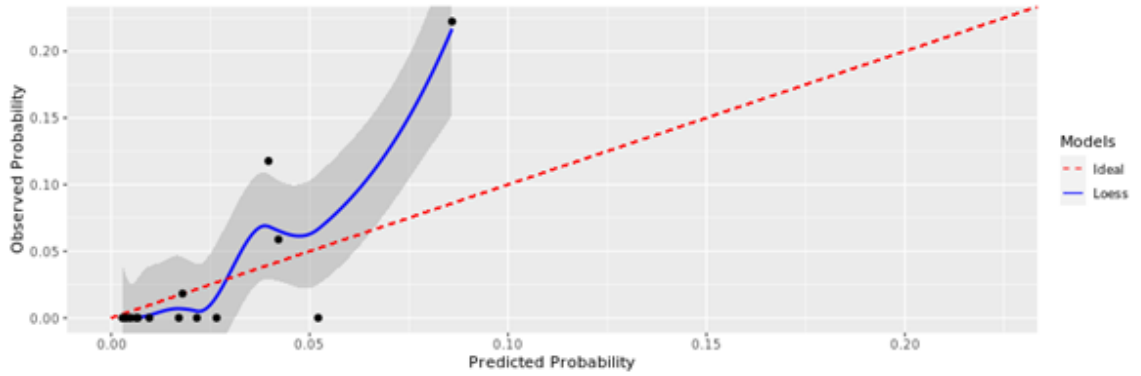
Tabel 5. Eesti terviseandmetel treenitud mudeli sisemise valideerimise tulemused.

Tulemi kirjeldus	**AUC**	**AUPRC**	**Tulemisse ennustatud inimeste arv**	**Ennustatud inimeste protsent (%)**
Kopsupõletikuga hospitaliseerimine	0,754 (0,692-0,816)	0,253	269	7,836
Intensiivravi vajadus	0,961 (0,929-0,993)	0,173	26	0,757
Surm	0,828 (0,638-1,02)	0,081	30	0,874

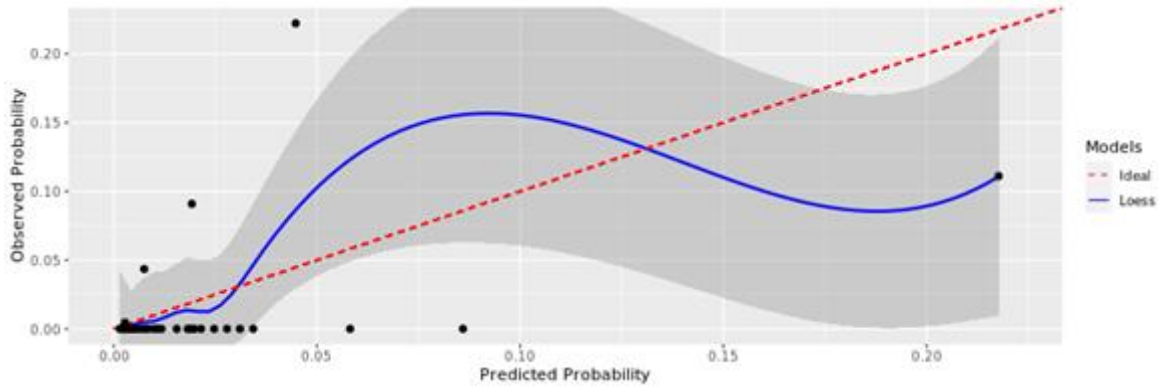
Mudeli kalibreerimine on kujutatud joonistel 12, 13 ja 14.



Joonis 12. Eesti terviseandmetel treenitud andmepõhise mudeli kalibreerimine hospitaliseerimise ennustamisel.



Joonis 13. Eesti terviseandmetel treenitud andmepõhise mudeli kalibreerimine intensiivravi vajaduse ennustamisel.



Joonis 14. Eesti terviseandmetel treenitud andmepõhise mudeli kalibreerimine patsiendi surma ennustamisel.

Võrreldes OHDSI poolt treenitud mudeli välise valideerimiste kalibreerimistele on näha, et töö käigus treenitud mudeli kalibreerimine on täpsem, kuid peab ka arvestama fakti, et Eesti

andmetega treenitud mudel on andmepõhine ning arvestab rohkemate tunnustega, mis võib tulemusi mõjutada. Eesti andmetel treenitud mudeli tulemused näitavad parimat võimalikku tõhusust taolise valimi ja ennustavate tulemitel, mistõttu on tulemuste parandamiseks variant valimit muuta, et kaasata peale gripihaigete patsientide ka COVID-19 diagnoosiga patsiente. Sellise valimi puhul suudaks mudel paremini arvestada COVID-19 ja gripi haiguse kulgude vahega ning ennustatakse täpsem risk (kalibreerimine muutuks paremaks).

4. Järeldused

Tulemustes on näha, et suurimaks probleemiks on OHDSI mudelite valideerimisel Eesti terviseandmetel kehv kalibreerimine. Kuigi AUC ja AUPRC näitajad on head, siis mudel ennustab reaalse tõenäosusega võrreldes liiga väikse riski tõenäosuse. Praktikas tähendaks see seda, et patsiendid, kes vajaksid kõrgema prioriteediga haiglaravi, eelisjärjekorras vaktsiini või muud ennetavaid meetmeid, jääksid mudeli ennustuste järgi piisava tähelepanuta. Mudeli kalibreerimise ebapiisavust võib selgitada see, et mudelit treeniti gripipatsientide andmetel, kuid valideeritakse COVID-19 andmetel. Selletõttu alahindab mudel patsiendi riski, kuna gripp on vähem tõsine haigus kui COVID-19 [31].

Veel võib kallutatust mõjutada andmestiku CDM-formaati viimine. Originaalsed terviseandmed on saadud erinevatest allikatest, mistõttu võisid osa andmeväljadest tühjaks jääda. See selgitaks ka seda, miks andmepõhise mudeli AUC näitaja on madalam kui teistel mudelitel.

Tulemusi mõjutas ka see, et valideerimiseks loodud kohordid olid väiksed ning tulemit ennustati väiksel protsendil valimist. See võis põhjustada valideerimistulemustes kallutatust, sest iga patsient mõjutas proportsionaalselt rohkem valideerimistulemusi.

Mudeli praktikasse kasutusele võtmist mõjutab ka asjaolu, et mudel on arendatud pandeemia alguses. Treenimisel ei kasutatud üldse COVID-19 patsientide andmeid ning valideeriti väiksematel COVID-19 andmestikel, mis olid loodud pandeemia esimestel kuudel.

Mudeli treenimise ajal oli COVID-19 andmeid liiga vähe, et piisava tõhususega mudel treenida. Tänaseks on aga COVID-19 juhtumite arv palju suurem ning oleks tõenäoliselt võimalik suurema COVID-19 andmete hulgaga mudel treenida. COVID-19 andmete kasutamine treenimisel muudaks tunnuste koefitsiendid selliselt, et mudeli ennustamine COVID-19 patsientide andmetel oleks täpsem.

Pandeemia algusest on muutunud ka see, et viirus on muteerunud, mille tõttu võivad OHDSI mudeli tunnuste koefitsiendid olla aegunud [32]. Käesolevas töös kasutatud terviseandmed on kogutud enne uute viiruse mutatsioonide sattumist Eestisse ehk valideerimised ei pruugi enam aktuaalsed olla [33]. Lisaks ei arvesta mudel vaktsiinide ja paranenud ravimeetodite olemasoluga, mis mõjutavad väga tugevalt haiguskulgu. Aktuaalsema ülevaate saamiseks on vajalik uuemate andmete kogumine ning nendega mudelite valideerimine.

Käesolevas töös tehtud valideerimiste tulemuste järgi võib järeldada, et OHDSI COVID-19 mudeleid ei ole mõistlik praktikas kasutusele võtta. Valideerimiseks kasutatud Eesti terviseandmed ja ka mudelid ise on aegunud, sest nad ei arvesta uute koroonaviiruse mutatsioonide, paremate ravimeetodite ja vaktsiinide olemasoluga.

5. Kokkuvõte

Käesoleva töö eesmärgiks oli väliselt valideerida OHDSI kommuuni uurijate poolt arendatud COVID-19 haigusekulgu ennustavad riskimudelid. OHDSI on avatud teadust tegev kommuun, mis püüab muuta terviseinformaatika valdkonda läbipaistvamaks muuta ning selle kaudu ülemaailmset koostööd teadlaste vahel tõhustada.

Valideeriti andmepõhine, vanus/sugu ning COVER mudelid, mis ennustasid patsiendi hospitaliseerimise, intensiivravi ja surma tõenäosust 30 päeva peale indekskuupäeva. Andmepõhine mudel kasutas tulemi ennustuseks kuni 521 patsiendi terviseloo tunnust. Vanus/sugu mudel, nagu nimi viitab, kasutab tunnustena ainult patsiendi vanust ning sugu. COVER mudel on andmepõhisest mudelist tuletatud mudel, mis kasutab tunnustena vanust ja sugu ning seitset potentsiaalset kaasnevat haigust.

Valideerimiseks kasutati Eesti terviseandmetest loodud CDM-formaadis andmestikku, mis sisaldas endas COVID-19 diagnoosiga patsiente. Väline valideerimine oli tehtud selleks, et uurida, kas teistel andmetel treenitud mudel suudaks ka Eesti terviseandmetel täpseid tulemusi ennustada. Eesti kontekstis on mõistlik kasutada eelnevalt suurematel andmekogudel treenitud mudeleid, kuna Eesti terviseandmete hulk on liiga väike, et treenida tõhusat mudelit.

Mudelite välise valideerimiste AUC vahemikud on järgmised: andmepõhise mudeli jaoks 0,736–0,874, vanus/sugu mudeli jaoks 0,746–0,932 ning COVER mudeli jaoks 0,746–0,930. Antud väärtused on pigem head, kuid mudelite kalibreerimine on kehv. Mudelid ennustavad patsiendi tegeliku riski tõenäosusega võrreldes liiga madala või kõrge riski tõenäosuse.

Pandeemia olemuse kiirete muutuste tõttu on probleemiks nii andmete kui ka mudelite aegumine ehk pandeemia hetkeseisu ei kajastata täpselt. Praeguseks on viirus muteerunud ning tekkinud on uued tüved, mis mõjutavad patsientide haiguskulgu teistmoodi võrreldes pandeemia algusega. Lisaks sellele on nüüdseks välja arendatud COVID-19 vaktsiinid, mis mõjutavad suurel määral haiguse põdemist.

Mudelite kehva kalibreerimise ja aegunud andmete tõttu ei ole mõistlik töös uuritud COVID-19 riskimudeleid meditsiinilises kontekstis kasutusele võtta. Töös analüüsitud puuduste parandamiseks oleks võimalik luua uuemad mudelid, mis kasutavad nii

treenimiseks kui ka valideerimiseks hiljutisemaid terviseandmeid, mis kajastavad paremini pandeemia hetkeolukorda.

Viidatud kirjandus

[1] Mareiniss D. P. The impending storm: COVID-19, pandemics and our overwhelmed emergency departments. *The American Journal of Emergency Medicine*. 2020. https://doi.org/10.1016/j.ajem.2020.03.033

[2] Terviseamet. Igaüks saab koroonaviiruse teise laine kiiret tõusu kontrolli all hoida. 2020. https://www.terviseamet.ee/et/uudised/igauks-saab-koroonaviiruse-teise-laine-kiiret-tousu-kontrolli-all-hoida (10.12.2020)

[3] Wynants L., Van Calster B., Collins G. S., Riley R. D., Heinze G., Schuit E., Bonten M. M. J., Dahly D. L., Damen J. A. A., Debray T. P. A., de Jong V. M. T. De Vos M., Dhiman P., Haller M. C., Harhay M. O., Henckaerts L., Heus P., Kreuzberger N., Lohmann A., Luijken K., Ma J., Martin G. P., Navarro C. L. A., Reitsma J. B., Sergeant J. C., Shi C., Skoetz N., Smits L. J. M., Snell K. I. E., Sperrin M., Spijker R., Steyerberg E. W., Takada T., Tzoulaki I., van Kuijk S. M. J., van Royen F. S., Verbakel J. Y., Wallisch C., Wilkinson J., Wolff R., Hooft L., Moons K. G. M., van Smeden M. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *The BMJ*. 2020. https://doi.org/10.1136/bmj.m1328

[4] World Health Organization. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/ (12.04.2021)

[5] Kook U. Lutsar: komandanditundi me ei soovita, rangemaid piiranguid küll. *ERR*. 2021. https://www.err.ee/1608119944/lutsar-komandanditundi-me-ei-soovita-rangemaid-piiranguid-kull (13.04.2021)

[6] European Centre for Disease Prevention and Control. Weekly COVID-19 country overview. https://www.ecdc.europa.eu/en/covid-19/country-overviews (13.04.2021)

[7] Põhja-Eesti Regionaalhaigla. Haiglaravi vajavate COVID-patsientide juurdevool põhja regiooni haiglatele on ületamas siinset ravivõimekust. 2021. https://koroonakriis.ee/haiglaravi-vajavate-covid-patsientide-juurdevool-pohja-regiooni-haiglatele-uletamas-siinset (13.04.2021)

[8] Wright H. Graphic: Half of hospitalized covid patients are being treated in Tallinn. *ERR*. 2021. https://news.err.ee/1608142858/graphic-half-of-hospitalized-covid-patients-are-being-treated-in-tallinn (13.04.2021)

[9] Li X., Xu S., Yu M., Wang K., Tao Y., Zhou Y., Shi J., Zhou M., Wu B., Yang Z., Zhang C., Yue J., Zhang Z., Renz H., Liu X., Xie J., Xie M., Zhao J. Risk factors for severity and morrtality in adult COVID-19 inpatients in Wuhan. *The Journal of allergy and clinical immunology*. 2020. https://doi.org/10.1016/j.jaci.2020.04.006

[10] Ji D., Dawei Z., Xu J., Chen Z., Yang T., Zhao P., Chen G., Cheng G., Wang Y., Bi J., Tan L., Lau G., Qin E. Prediction for Progression Risk in Patients With COVID-19 Pneumonia: The CALL Score. *Clinical Infectious Diseases*. 2020. https://doi.org/10.1093/cid/ciaa414

[11] Sotsiaalministeerium. COVID-19 vaktsineerimise plaan. 2020. lk 1-2. https://www.sm.ee/sites/default/files/news-related-files/covid-19_vaktsineerimise_plaan_14.12.2020.pdf (13.04.2021)

[12] Salisbury A., Spertus J. Realizing the Potential of Clinical Risk Prediction Models. Where Are We Now and What Needs to Change to Better Personalize Delivery of Care? *Circ Cardiovasc Qual Outcomes*. 2015. https://doi.org/10.1161/CIRCOUTCOMES.115.002038

[13] Panattoni L., Vaithianathan R., Ashton T., Lewis G. Predictive risk modelling in health: options for New Zealand and Australia. *Australian Health Review*. 2011. https://doi.org/10.1071/AH09845

[14] Expert.ai. What is Machine Learning? A Definition. 2020. https://www.expert.ai/blog/machine-learning-definition/ (06.05.2021)

[15] IBM Cloud Education. Machine Learning. 2020. https://www.ibm.com/cloud/learn/machine-learning (23.04.2021)

[16] Rijnbeek P., Reps J. Chapter 13 Patient-Level Prediction. *The Book of OHDSI*. 2019. https://ohdsi.github.io/TheBookOfOhdsi/PatientLevelPrediction.html (06.05.2021)

[17] Glen, S. Lasso Regression: Simple Definition. 2015. https://www.statisticshowto.com/lasso-regression/ (24.04.2021)

[18] Molnar C. Interpretable Machine Learning. 2019. lk 71-78. https://christophm.github.io/interpretable-ml-book/logistic.html (06.05.2021)

[19] Narvik P. Kant- ja lassoregressioon ning nende rakendamine müügiskoori loomiseks Creditinfo Eesti AS andmetel. Tartu. 2017. lk 11-13.

[20] Draelos R. L. B. The Complete Guide to AUC and Average Precision. *Towards Data Science*. 2020. https://towardsdatascience.com/the-complete-guide-to-auc-and-average-precision-cf1d4647efc3 (04.05.2021)

[21] Glen S. Lowess Smoothing in Statistics: What is it. 2013. https://www.statisticshowto.com/lowess-smoothing/ (05.05.2021)

[22] Finster M., Wood M. The Apgar score has survived the test of time. *Anesthesiology*. 2005. https://doi.org/10.1097/00000542-200504000-00022

[23] Stang P. E., Ryan P. B., Racoosin J. A., Overhage J. M., Hartzema A. G., Reich C., Welebob E., Scarnecchia T., Woodcock J. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med*. 2010. https://doi.org/10.7326/0003-4819-153-9-201011020-00010

[24] Ryan P., Hripcsak G. Chapter 1 The OHDSI Community. *The Book of OHDSI*. 2019. https://ohdsi.github.io/TheBookOfOhdsi/OhdsiCommunity.html (12.04.2021)

[25] Kostka K. Chapter 10 Defining Cohorts. *The Book of OHDSI*. 2019. https://ohdsi.github.io/TheBookOfOhdsi/Cohorts.html (06.05.2021)

[26] Williams R. D., Markus A. F., Yang C., Salles T. D., DuVall S. L., Falconer T., Jonnagaddala J., Kim C., Rho Y., Williams A., Alberga A., An M. H., Aragón M., Areia C., Choi Y. H., Drakos I., Abrahão M. T. F., Fernández-Bertolín S., Hripcsak G., Kaas-Hansen B. S., Kandukuri P. L., Kors J. A., Kostka K., Liaw S. T., Lynch K. E., Machnicki G., Matheny M. E., Morales D., Nyberg F., Park R. W., Prats-Uribe A., Pratt N., Rao G., Reich C. G., Rivera M., Seinen T., Shoaibi A., Spotnitz M. E., Steyerberg E. W., Suchard M. A., You S. C., Zhang L., Zhou L., Ryan P. B., Prieto-Alhambra D., Reps J. M., Rijnbeek P. R. Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. *medRxiv*. 2020. https://doi.org/10.1101/2020.05.26.20112649

[27] Equator Network. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. https://www.equator-network.org/reporting-guidelines/tripod-statement/ (06.05.2021)

[28] Uusküla A., Kolde R., Piirsoo M. COVID-19 haigusjuhtumite analüüs ja riskirühmade väljaselgitamine Eestis. 2020. https://www.etis.ee/Portal/Projects/Display/13fb7703-84db-44a1-a44a-9323ee99650e (06.05.2021)

[29] Reps J., Williams R., Rijnbeek P. Development and validation of complex and simple patient-level prediction models for predicting various outcomes in COVID patients: a rapid network study to inform the management of COVID-19. *Github*. 2020. https://github.com/ohdsi-studies/Covid19PredictionStudies (29.12.2020)

[30] OHDSI. Covid19CoverPrediction. *Github*. 2020. https://github.com/OHDSI/ShinyDeploy/tree/master/Covid19CoverPrediction (06.05.2021)

[31] World Health Organization. Coronavirus disease (COVID-19): Similarities and differences with influenza. 2020. https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-similarities-and-differences-with-influenza (06.05.2021)

[32] Bollinger R., Ray S. New Variants of Coronavirus: What You Should Know. *John Hopkins Medicine*. 2021. https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/a-new-strain-of-coronavirus-what-you-should-know (06.05.2021)

[33] Raestik T. Eestis on tuvastatud kolm Ühendkuningriigist sissetoodud COVID-19 uue agressiivse mutatsiooni põdejat. Tervisegeenius. 2021. https://tervise.geenius.ee/rubriik/uudis/eestis-on-tuvastatud-kolm-uhendkuningriigist-sissetoodud-covid-19-uue-agressiivse-mutatsiooni-podejat/ (06.05.2021)

Tänusõnad

Käesoleva töö valmimisel oli suureks abiks töö juhendaja, Raivo Kolde, kes aitas töö jooksul esinevate probleemide lahendamisega. Suur tänu läheb ka OHDSI kommuunile ning Ross D. Williamsile, kes oli nõus aitama ning vastas mitmele tööga seotud küsimusele.

Lisad

I. Litsents

Mina, Marc David,

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

COVID-19 ennustavate riskimudelite rakendatavuse hindamine Eesti terviseandmetel, 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.

Marc David

06.05.2021