Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Triin Bulõgina

# Mikrobioomil põhinevad ennustusmudelid soolevähi diagnoosimiseks

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja MSc Oliver Aasmets

Tartu 2021

**MIKROBIOOMIL PÕHINEVAD ENNUSTUSMUDELID SOOLEVÄHI DIAGNOOSIMISEKS**

Bakalaureusetöö

Triin Bulõgina

**Lühikokkuvõte**

Inimesega seotud mikroorganismide ehk mikrobioomi uurimisvaldkond on muutunud väga oluliseks personaalmeditsiini arengus. Sellest lähtuvalt uuritakse bakalaureusetöös mikrobioomi kasutamist soolevähi diagnoosimiseks. Bakalaureusetöö eesmärk on luua soolevähi ennustusmudelid, kasutades lisaks klassikalistele soolevähi riskiteguritele soolestiku mikrobioomi andmeid. Mudelite koostamisel tutvustatakse *Selbal* metoodikat, mis on väljatöötatud mikrobioomi andmete analüüsimiseks. Meetodit kasutades luuakse soolevähi ennustusmudelid viie erineva populatsiooni andmetel ning hinnatakse prognoosimudelite ennustusvõimet soolevähi diagnoosimiseks ning mudelite üldistusvõimet teistesse populatsioonidesse.

**CERCS teaduseriala:** P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

**Märksõnad:** Soolevähk, Ennustusmudelid, Mikrobioom, ROC-kõver.

**COLORECTAL CANCER PREDICTION MODELS USING GUT MICROBIOME AS A PREDICTOR**

Bachelor thesis

Triin Bulõgina

**Abstract**

The field of research on human microorganisms, or microbiome, has become very important in the development of personal medicine and the present work focuses on the study of the possibility of using the microbiome in the prognosis of colorectal cancer. The aim of this bachelor's thesis is to create prediction models for colorectal cancer, using microbiome data in addition to classical cancer risk factors. Selbal methodology, which is developed for microbiome data analysis, is introduced to develop prediction models. Colorectal cancer risk models are based on data from five different populations to demonstrate the importance of the microbiome in the prognosis of colorectal cancer, and the generalizability of the prediction models in external populatons is further assessed.

**CERCS research specialisation:** P160 Statistics, operations research, programming, financial and actuarial mathematics.

**Key Words:** Colorectal cancer, Prediction models, Microbiome, ROC-curve.

# Sisukord

# Sissejuhatus

Inimese mikrobioom ehk erinevate mikroorganismide kogum täidab mitmeid organismi toimimiseks vajalikke funktsioone ning mikrobioomi tasakaalutuse ehk düsbioosiga seostatakse haiguste nagu diabeedi, ärritunud soole sündroomi, ülekaalulisuse ja depressiooni teket. Samuti on täheldatud mikroobioomi koosluse muutumisel seost vähkkasvajate kõrgema tekkeriskiga. Kuivõrd inimese mikrobioom on kõige mitmekesisem just soolestikus, uuritakse bakalaureusetöös mikrobioomi kasutamise võimalusi soolevähi diagnoosimiseks.

Soolevähk on pahaloomuline kasvaja, millel on nii pärilikud kui ka mittepärilikud tekkepõhjused. Teadaolevad riskitegurid vähkkasvaja tekkeks on inimese vanus, toitumisharjumused, pärilikkus, ülekaalulisus ja suitsetamine. Bakalaureusetöö eesmärk on leida soolevähki prognoosivad mudelid, kasutades lisaks klassikalistele riskiteguritele soolestiku mikrobioomi andmeid. Töös tutvustatakse mikrobioomi andmete spetsiifikal põhinevat metoodit *Selbal*, mille abil luuakse prognoosimudelid soolevähi tuvastamiseks ning hinnatakse mudelite üldistusvõimet erinevates populatsioonides.

Töö jaguneb teoreetiliseks ja praktiliseks osaks. Teoreetilise osa esimeses peatükis selgitatakse mikrobioomi olulisust ja selle uurimise vajalikkust. Teine peatükk keskendub soolevähi riskiteguritele, lisaks antakse ülevaade soolevähi haigestumusest Eestis ning kirjeldadakse mikrobioomi kasutamise võimalusi soolevähi diagnoosimisel. Kolmandas peatükis antakse ülevaade mikrobioomi andmete kogumisest ja nende tõlgendamisega seotud probleemidest. Neljas peatükk käsitleb ennustusmudelite loomiseks kasutatava *Selbal* metoodika teoreetilist kirjeldust ja ennustusmudelite interpreteerimist. Lisaks selgitatakse mudeli ennustusvõime hindamiseks kasutatud näitajat ROC-kõvera alusest pindalast.

Praktilises osas kirjeldatakse kasutatud andmeid ja ennustusmudelisse kaasatavaid tunnuseid. Seejärel rakendatakse *Selbal* mudelit erinevate populatsioonide andmetele ning hinnatakse saadud mudelite üldistusvõimet. Lisaks analüüsitakse saadud

ennustusmudelite parameetrite hinnanguid ja olulisust ning võrreldakse erinevates populatsioonides mudelisse kaasatud bakteriliike.

Töö autor tänab juhendajat Oliver Aasmetsa suunamise, tähelepanelike paranduste ja pühendatud aja eest.

Töö praktilise osa läbiviimiseks ning tulemuste graafilisel kujutamiseks kasutatakse rakendustarkvara R ning töö vormistamiseks tekstitöötlustarkvara LaTeX.

# 1 Inimese mikrobioom ja selle uurimise motivatsioon

Inimese mikrobioomi all mõistetakse inimesega seotud mikroorganismide ehk bakterite, seente ja viiruste geneetilise materjali kogumit [1]. Sealjuures eristatakse erinevate keha piirkondade, nagu hingamisteede, naha, seedetrakti ning suguelundite mikrobioomi, mis kõik erinevad üksteisest liigirikkuse ning koosluse suhtes [2]. Mikrobioomi moodustavaid rakke on inimese kehas vähemalt sama palju kui inimese enda keharakke ning lisaks sisaldab iga mikroorganismi genoom tuhandeid geene, mis kokku muudab mikrobioomi inimese genoomist oluliselt mitmekesisemaks. Näiteks on teada, et mikrobioomi elutegevuse tulemusel toodetakse mitmeid inimese keha funktsioneerimiseks vajalikke molekule, sealhulgas vitamiine. [3]

Mikrobioomi kooslus areneb välja suures osas vastsündinuna - sünnitusprotsessi järgselt algab seedetrakti mikrobioom formuleeruma ning edaspidi on uute liikide lisandumine ja mikrobioomi mitmekesistumine mõjutatud eelkõige toitumisest ning ümbritsevast keskkonnast. Kõige suurem muutus on tingitud lapse rinnapiimalt tahkele toidule üleminekuga. Esimesteks eluaastateks on mikrobioomi mitmekesisus suuresti kogu eluks välja kujunenud. [2]

Täiskasvanueas on mikrobioomi kooslus samuti enim mõjutatud toitumisharjumustest, ümbritsevast keskkonnast, ravimite tarvitamisest ning elustiilist. Näiteks on teada, et antibiootikumidega ravimine mõjutab soolestiku tasakaalu vähendades soolebakterite mitmekesisust, mistõttu langeb inimese immuunsussüsteemi vastupanuvõime, mis omakorda võib soodustada näiteks ülekaalulisuse, astma ja diabeedi teket. Samuti võib stress mõjutada soolebakterite mitmekesisust, soodustades depressiooni ja ärevushäire tekkimist. [2]

Üks tähtsamaid suundi mikrobioomi valdkonnas on mikrobioomi rolli uurimine komplekshaiguste puhul. Tänaseks on teada, et mikrobioomi tasakaalutuse ehk düsbioosiga on seotud haigused nagu ärritunud soole sündroom, ülekaalulisus, diabeet,

depressioon ning isegi erinevad vähivormid, näiteks soolevähk. Tänu muutustele mikrobioomis on võimalik senisest paremini haiguseid diagnoosida ning haigusriske hinnata. Kõige selle juures on märkimisväärne, et mikrobioomi kooslust on võimalik lihtsate meetmetega muuta. Näiteks on võimalik toitumisharjumuste muutmisega või probiootikumide kasutamisega mikrobioomi tasakaalu sobivas suunas mõjutada ning haiguste riske vähendada. Seetõttu pakub mikrobioomi uurimine suurt huvi personaalmeditsiini arendamisel. [4]

# 2 Soolevähk

Vähk on pahaloomuline kasvaja, mis tekib kui keharakkudes leiavad aset mutatsioonid ning kasvajarakkude kasv väljub kontrolli alt. Vähi tekkepõhjuseid on mitmeid - see võib olla pärilik, aga ka tingitud väliskeskkonnast või organismis tekkivatest kantserogeensetest ainetest. [5] Jämesoolevähk on pahaloomulise vähkkasvaja vorm. Jämesool koosneb umbsoolest, käär- ja pärasoolest, kus toimub seedeprotsessis tekkinud jääkainete lagundamine ning vee ja mineraalsoolade imendumine [6]. Edaspidi käsitletakse umb-, käär-, ja pärasoole vähivorme üldistatult soolevähina. Soolevähk tekib valdavalt sooleseinas olevatest healoomulistest kasvajatest ehk polüüpidest, millest umbes iga kahekümnes areneb edasi vähiks [6]. Kasvaja teke on järk-järguline protsess ning põhjustatud erinevatest faktoritest ja nende koosmõjudest. Suurimad riskitegurid soolevähi tekkeks on vanus, toitumisharjumused, pärilikkus, ülekaalulisus, erinevad haigused ja suitsetamine. [7] Seega on paljud riskitegurid mittepärilikud ja seotud inimese elustiiliga, mille muutmine tervislikumas suunas võimaldab soolevähi tekkeriski vähendada. Lisaks klassikalistele riskiteguritele on uuritud ka mikrobioomi rolli soolevähi tekkes ning selgunud on bioloogilised mehhanismid, mille abil mikrobioom otseselt vähi teket mõjutab [8].

Soolevähk on Euroopas meeste seas levimuselt kolmas diagnoositud vähivorm, naiste seas teisel kohal. Hinnanguliselt saab vanusevahemikus 0-74 aastat vähidiagnoosi

üks naine 35 naise hulgast, meeste puhul üks vähijuhtum 22 mehe kohta. Eestis on keskmine soolevähki haigestumus 2020. aasta andmete põhjal 100 000 elaniku kohta 74 inimest, mis on kõrgem kui Soomes, Rootsis ja Leedus. Soolevähki suremus on 100 000 elaniku kohta vahemikus 38,1−44,8 inimest, mis tähendab, et umbes 50% diagnoositud vähijuhtumitest lõppevad inimese surmaga. [9]

Soolevähk on üks vähivorme, mida on varajases staadiumis võimalik edukalt ravida. Pahaloomuline kasvaja eemaldatakse tavaliselt kirurgiliselt või laparoskoopiliselt ehk kõhuõõne kaudu ning vajadusel tehakse ka kiiritus- ja keemiaravi. Ravimeetodid on muutunud aastatega olulisemalt edukamaks - taastumine on kiirem ning aina rohkematel juhtudel on võimalik taastada soolepidavus, mis võimaldab inimestel pöörduda igapäevaellu. [6] Paraku kasvaja varajases staadiumis sümptomeid ei esine, mistõttu avastatakse sageli vähirakud alles hilises staadiumis, kus ravivõimalused on piiratud või puudvad. Soolevähi ennetamiseks ning haigestumise pidurdamiseks on Eestis alates 2016. aastast algatatud sõeluuringuprogrammid. Kuna soolevähi areng on pika peiteperioodiga, võimaldab sõeluuring mutatsioone varakult avastada, mis lihtsustab oluliselt patsientide ravi, parandab raviprognoosi ja vähendab kulusid vähiravile. [10]

Eestis kutsutakse sõeluuringusse 60-69 aastaseid mehi ja naisi iga kahe aasta tagant. Esmalt tuvastatakse, kas inimese väljaheite proovis on peitverd ning positiivse vastuse korral suunatakse inimene koloskoopiasse ehk sooleuuringusse, mis võimaldab näha sooles toimunud muutuseid vähkkasvaja tekke avastamiseks. [10] Paraku on kolonoskoopia puhul tegu äärmiselt invasiivse ning inimesele ebamugava protseduuriga, millest paljud peitveretestilt protseduurile suunatud patsiendid loobuvad. 2018. aasta sõeluuringu raames suunati soolevähi lisauuringule 1539 inimest, kuid lisauuringu tulemus on nendest vaid 35%-l kutsututest [11]. Kuigi aastatega on sõeluuringus osalemise aktiivsus tõusnud, on osalus pigem tagasihoidlik. Näiteks 2019. aastal sai kutse sõeluuringule 30 485 naist ja 37 565 meest, kokku 68 050 inimest. Ravikindustatud ja ravikindlustamata isikute jämesoolevähi sõeluuringule kutsutavate hõlmatus oli 53,7% ehk uuringus osales 36 522 inimest, kusjuures

naiste osalus oli 58,6% ja meestel vastavalt 47,6%. [12]

## 2.1 Mikrobioom soolevähi diagnostikas

Kuigi sõeluuringud on aidanud edukalt kaasa soolevähi varajasele avastamisele peamiselt vanuses 50-74 aastat, on endiselt probleeme uuringu metoodika ja võimekusega vähkkasvajat diagnoosida. Hetkel Eestis kasutusel oleval peitveretestil on madal tundlikkus ehk test ei suuda väljaheitest vere tuvastamisel hinnata, kas tegemist võib olla kasvajast põhjustatud peitverega või on positiivse peitvere testi tulemus seotud muude põhjustega. Ligikaudu 30% positiivsetest peitvere proovi tulemustest võib edasises sooleuuringus anda negatiivse tulemuse, mis on nii rahaliselt kui ajaliselt kulukas. Samuti on täheldatud, et olulised soolevähi riskitegurid nagu vanus ja suitsetamine võivad anda sagedasemini valenegatiivse tulemuse, mistõttu võib vähkkasvajaga inimene jääda edasistest uuringutest kõrvale. [13]
Võttes arvesse lisaks eelnimetatud põhjustele ka kolonoskoopia invasiivusest tuleneva uuringust väljalangemise, on oluline leida alternatiivseid meetodeid, mis võiksid aidata soolevähki kindlamini tuvastada juba peitveretesti abil. Üheks võimaluseks on kasutada väljaheiteproovist määratud mikrobioomi. Kuna inimese mikrobioomi mitmekesisus on kõige suurem soolestikus ning on leitud tugevaid seoseid mikrobioomi ning erinevate seedekulga haiguste ja soolevähi vahel, on see potentsiaalne abivahend soolevähi tuvastamiseks. [13]

Paraku on mikrobioomi andmetel spetsiifilisi omadusi, mis raskendavad andmete analüüsimist ning ennustusmudelite loomist. Seni on edukalt kasutatud mikrobioomi andmete põhjal soolevähi diagnoosimiseks masinõppel põhinevaid lähenemisi, mille abil on saadud mikrobioomi põhjal hästi soolevähki prognoosivaid mudeleid, kuid probleemiks on jäänud mudelite üldistusvõime teiste populatsioonide andmetele. Enamasti ühes populatsioonis välja töötatud mudel ei tööta teise populatsiooni andmetel. Üheks võimalikuks põhjuseks on mikrobioomi andmete eripäraga mitte arvestamine. Edasises tutvustatakse mikrobioomi andmete omadusi ning

mikrobioomi andmete eripära silmas pidades välja töötatud meetodit *Selbal*, mida kasutatakse soolevähi ennustusmudelite loomiseks [14].

# 3 Mikrobioomi andmete omadused

Mikrobioomi andmete kogumine ja analüüsimine on keeruline protsess. Selleks kogutakse inimeselt proov, milleks on üldiselt väljaheiteproov, ning sellest eraldatakse DNA, mille järjestust loetakse ehk sekveneeritakse. Protseduurile järgneb bioinformaatiline analüüs, mille tulemusena saadakse andmetabel, kus ridades on indiviidid, veergudes sekveneerimise teel leitud bakteriliigid ning tabeli väärtusteks lugemite arv, mis näitab, mitu korda vastavasse liiki kuuluvat bakterit antud isikul tuvastati.

Saadud andmetabel on paljuski mõjutatud uuringudisainist, proovide säilitamise tingimustest, sekveneerimise meetodist ning bioinformaatiliseks analüüsiks kasutatud meetoditest. Andmete edasist analüüsi mõjutab enim asjaolu, et bakterite lugemite arv andmetabelis on tingitud sekveneerimisinstrumendi täpsusest ning indiviidi bakterite tegelikke arve organismis ei ole võimalik teada saada. Sellel põhjusel tuleb saadud andmeid tõlgendada suhtarvudena ning selliseid andmeid defineeritakse kui kompositsionaalsed andmed (*compositional data*). [15]

Seega saab mikrobioomi andmeid kujutada järgneva tabelina,

Tabel 1: Näide mikrobioomi andmetest

| Indiviid/Bakter | $x_1$ | $x_2$ | $\dots$ | $x_p$ | $\sum$ |
|---|---|---|---|---|---|
| Indiviid 1 | $x_{11}$ | $x_{12}$ | $\dots$ | $x_{1p}$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| Indiviid n | $x_{n1}$ | $x_{n2}$ | $\dots$ | $x_{np}$ | 1 |

kus $x_{ij}$ tähistab *i.* indiviidi *j.* bakteri suhtelist sagedust *i.* indiviidi proovis, iga $i = 1, \dots, n$, $j = 1, \dots, p$ korral.

Vektor $\mathbf{x} = [x_1, \dots, x_p]$ on defineeritud p-osalise kompositsioonina, kui elemendid $x_i$ on positiivsed reaalarvud, mis kannavad endas suhtelist informatsiooni ning mille

koguarv on piiratud. Siis paiknevad vektori $\mathbf{x}$ elemendid simpleksil (*simplex*) $S^p$, mis on defineeritud kui

$$S^p = \left\{ \mathbf{x} = [x_1, \ldots, x_p] | x_i > 0, i = 1, \ldots, p \sum_i x_i = D \right\} \tag{3.1}$$

kus $\mathbf{x}$ tähistab kompositsioonaalset vektorit, $x_i$ tähistab kompositsiooni elementi, $p$ tähistab kompositsiooni elementide arvu ning $D$ tähistab konstantset summat ehk suhtarvude summat. [16] On selge, et tabeli 1 read vastavad kompositsionaalsete andmete definitsioonile.

Mikrobioomi andmete kompositsionaalsusest tuleneb mitmeid probleeme, mis raskendavad andmete analüüsimist. Näiteks, kuna bakterite osakaalude kogusumma on fikseeritud, võib näha andmetes „valesid“ (*spurious*) korrelatsioone, mida tegelikkuses bakterite arvukuste vahel ei esine. Lisaks eelmainitule on probleemiks alamkompositioonil tehtavate järelduste üldistatavus. Näiteks kompositsioonist alamkompositsiooni eraldamise teel peaksid säilima algandmetel saadud seosed, kuid kuna mikrobioomi andmed on sõltuvad, siis juba ühe bakteri eraldamine andmestikust muudab ülejäänute osakaalusid ja seetõttu ka bakterite omavahelisi korrelatsioone. Seega võivad alamkompositsioonilt tehtavad järeldused tegelikkuses mitte kehtida. [17]

**Näide**

Illustreerimaks mikrobioomi andmete kompositsionaalsusega kaasnevaid probleeme kirjeldatakse järgnevalt näiteandmestiku (Tabel 2) peal kahe indiviidi $S_1$ ja $S_2$ mikrobioomi andmeid. Olgu mõlema indiviidi puhul olemas andmed nende organismis olevate bakterite tegeliku arvukuse kohta. Kuid reaalsete andmete korral saab mikrobioomi andmeid analüüsida ainult liikide suhtelise arvukuse põhjal.

Tabel 2: Bakterite tegelik arvukus ja suhteline arvukus

| | Tegelik arvukus | | | | | Suhteline arvukus | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Indiviid | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $\sum$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $\sum$ |
| $S_1$ | 10 | 30 | 60 | 100 | 200 | 5% | 15% | 30% | 50% | 100% |
| $S_2$ | 10 | 30 | 60 | 400 | 500 | 2% | 6% | 12% | 80% | 100% |

Vaadates esmalt bakterite tegelikku arvukust, on indiviidil $S_1$ ja $S_2$ bakterite $X_1 - X_3$ arv ühesugune ning erinev on vaid bakteri $X_4$ arvukus, mis on indiviidil $S_2$ neli korda suurem kui indiviidil $S_1$. Kui oleks täiendavalt teada, et indiviid $S_2$ on haige ja $S_1$ terve, siis võiks haiguse tekkepõhjuseks olla potentsiaalselt bakteri $X_4$ oluliselt kõrgem sisaldus võrreldes terve inimesega. Kuid klassikalisel sekveneerimisel põhineva mikrobioomi analüüsi puhul ei ole võimalik teada indiviidide bakterite üldarvu, mistõttu ei ole praktikas võimalik selliseid andmeid koguda.

Bakterite suhtelist arvukust võrreldes on bakterite osakaalud sõltuvad iga inimese bakterite koguarvust. Kuigi mõlemal indiviidil oli bakterite $X_1 - X_3$ tegelik arvukus ühesuur, ei ole nende bakterite proportsioonid organismis enam ühesugused, mis võib viia eksiarvamusele, et kõigi bakterite arvukus tervel ja haigel inimesel on erinev. See on üheks peamiseks põhjuseks, miks on mikrobioomi andmete analüüsimine ning tulemuste põhjal järelduste tegemine raskendatud.

Lisaks eelmainitule raskendab andmete analüüsimist asjaolu, et mikrobioomi andmetes on sageli erinevate vaadeldud liikide arv märkimisväärselt suurem kui uuringus osalejate arv. Seetõttu pööratakse palju tähelepanu just mikrobioomi andmete analüüsimiseks mõeldud meetodite arendamisele, mis arvestaksid eelnimetatud probleemidega. Järgnevalt tutvustatakse üht spetsiaalselt mikrobioomi andmete analüüsimiseks välja töötatud meetodit, mis on implementeeritud R paketis *Selbal*. [14]

# 4 Metoodika

Töö eesmärk on luua mudelid soolevähi prognoosimiseks, kasutades lisaks klassikalistele riskiteguritele ka soolestiku mikrobioomi andmeid ning hinnata saadud mudelite üldistusvõimet erinevates populatsioonides. Mudelite loomiseks kasutatakse meetodit *Selbal*, mis kujutab endast logistilise regressiooni mudelit, kus üheks kovariaadiks on mikrobioomi andmetelt leitud tunnus [14].

## 4.1 Logistiline regressioon

Siinses alapeatükis kirjeldatud metoodika tugineb autorite James jt raamatul „An Introduction to Statistical Learning“, kui ei ole teisiti viidatud [18].

Eesmärgiga eristada soolevähki põdevaid inimesi tervetest, on uuritav tunnus binaarne ehk kahe võimaliku väärtusega. Olgu Y uuritavale tunnusele vastav binaarne juhuslik suurus:

$$Y = \begin{cases} 1 & , \text{ inimesel on soolevähk tõenäosusega } \pi \\ 0 & , \text{ inimesel ei ole soolevähki tõenäosusega 1-}\pi \end{cases}$$

Sellisel juhul on $Y$ Bernoulli jaotusega juhuslik suurus parameetriga $\pi$: $\mathrm{Y} \sim Be(\pi)$, $0 \leq \pi \leq 1$. Bernoulli jaotuse puhul avaldub uuritava tunnuse keskväärtus sündmuse toimumise tõenäosusena ehk $EY = P(Y = 1) = \pi$.

Olgu $\mathbf{X} = (X_1, \ldots, X_m)$ argumenttunnused, mille abil soovime uuritavat tunnust prognoosida. Üks võimalus uuritava tunnuse modelleerimiseks on kasutada lineaarse regressiooni mudelit:

$$E[Y|X_1 = x_1, \ldots, X_m = x_m] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{4.1}$$

Kus $\beta_0$ on mudeli vabaliige, $x_1, \ldots, x_m$ on mudeli argumendid ning $\beta_1, \ldots, \beta_m$ on mudeli kordajad. Paraku ei tarvitse sellisel juhul uuritava tunnuse tinglik keskväärtus $E[Y|\mathbf{X}]$ olla tõkestatud lõigus $[0, 1]$. Samas, kuna uuritav tunnus Y on Bernoulli jaotusega, on tema keskväärtus võrdne sündmuse toimumise tõenäosusega. Sellisel juhul lõigust $[0, 1]$ välja jäävad hinnangud ei ole loogilised. Selle probleemi lahendamiseks kasutatakse binaarse uuritava tunnuse puhul mudeli loomisel seosefunktsioone, mille muutumispiirkonnaks on reaalarvude hulk. Üheks selliseks funktsiooniks on *logit* funktsioon:

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) \tag{4.2}$$

Kus suurust $\prod = \frac{\pi}{1-\pi}$ nimetatakse sündmuse esinemise šansiks (*odd*), mis on defineeritud sündmuse esinemise ja vastandsündmuse toimumise suhtena. Logistilise regressiooni mudel avaldub kujul:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \tag{4.3}$$

Uuritava tunnuse $Y$ toimumise tõenäosus $\pi$ on avaldatav järgnevalt:

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m}} \tag{4.4}$$

Mudeli interpreteerimiseks kasutatakse šansside suhet (*odds ratio*), mis on defineeritud kahe indiviidi $i$ ja $j$ šansside suhtena

$$OR = \frac{\prod_i}{\prod_j} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}}.$$

Olgu mudelis interpreteeritav argument $X_1$, mille kordaja on $\beta_1$. Öeldakse, et teiste argumentide samaks jäädes, muudab argumendi $X_1$ $k-$ühikuline muutus indiviidide $i$ ja $j$ šansside suhet $e^{k\cdot\beta_1}$ korda.

Šansside suhte muutuse suund sõltub interpreteeritava argumendi kordaja ees olevast märgist - positiivne kordaja argumendi ees viitab samapidisele seosele uuritava tunnuse ja argumendi vahel, negatiivne kordaja vastupidisele seosele.

Töös kasutatav *selbal* algoritm leiab soolevähki ennustava logistilise regressiooni mudeli, mille üheks kovariaadiks on mikrobioomi andmetelt leitud liige (edaspidi MIL). Lisaks mikrobioomi komponendile kasutatakse argumenttunnustena kehamassiindeksit ning vanust. Seega avalduvad mudelid kujul:

$$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot VANUS + \beta_2 \cdot KMI + \beta_3 \cdot MIL$$

kus $\beta_0$ on mudeli vabaliige, $\beta_1, \ldots, \beta_3$ on mudeli argumentide kordajad ning MIL on mikrobioomi andmetest moodustatud arvuline tunnus.

## 4.2 Mudeli headuse hindamine ROC-kõvera abil

Prognoosimudelite loomisel on oluline kuidagi hinnata mudelite ennustustäpsust. Ennustustäpsus aitab võrrelda erinevaid algoritme valimaks välja efektiivseima, kuid samuti on vaja mudelite headust hinnata optimaalse mudelikuju leidmisel. Töös kasutatakse selleks ROC-kõvera alust pindala ehk AUC (*Area Under the Curve*) väärtust, mille olemust järgnevalt tutvustatakse. Siinses alapeatükis kirjeldatud metoodika tugineb autori Tanel Kaardi õpiobjektil „Binaarsete tunnuste analüüsimeetodid“, kui ei ole teisiti viidatud [19].

Olgu uuritav tunnus binaarne, mis kirjeldab inimese soolevähi prognoosi. Tabel 3 kirjeldab võimalikke tulemusi prognoosimudeli rakendamisel.

Tabel 3: Prognooside klassifitseerimine

| | Tegelik diagnoos | |
|---|---|---|
| Mudeli prognoos | Negatiivne | Positiivne |
| Negatiivne | Tõene negatiivne (TN) | Valenegatiivne (FN) |
| Positiivne | Valepositiivne (FP) | Tõene positiivne (TP) |

*Positiivsena käsitletakse juhtu, kus inimesel on prognoositud soolevähk, vastasel korral on tegemist negatiivse juhtumiga.*

- tõene positiivne (*true positive, TP*) - juhtumid, kus vähkkasvajaga inimesel on prognoositud vähkkasvaja
- tõene negatiivne (*true negative, TN*) - juhtumid, kus terve inimese vähkkasvaja prognoos on negatiivne
- valepositiivne (*false positive, FP*) - juhtumid, kus terve inimene saab positiivse vähkkasvaja prognoosi
- valenegatiivne (*false negative, FN*) - juhtumid, kus vähkkasvajaga inimene ei saa kasvaja prognoosi.

Tasakaalustatud valimiga on võimalik mudeli täpsust hinnata klassikaliselt, jagades õigesti prognoositud juhtude arvu kogu uuringus osalejate arvuga. Kui valim ei ole tasakaalustatud ehk näiteks kontrollgrupis on inimesi oluliselt rohkem kui haigusjuhtumiga patsientide grupis, võib klassikaline täpsus anda vale tulemuse, sest klassikalisel juhul ei arvestata tõeste negatiivsete ja valepositiivsete juhtudega. [18] Sellest lähtuvalt pakuvad rohkem huvi mudeli tundlikkus (*sensitivity*) ja spetsiifilisus (*specificity*). Tundlikkus näitab kui suur on soolevähiga patsientide osakaal, kellele prognoositi mudeli põhjal vähkkasvaja ehk milline on tõeste positiivsete prognooside määr:

$$\text{Tundlikkus} = \frac{TP}{TP + FN}.$$

Spetsiifilisus näitab, kui suure osakaalu moodustavad terved inimesed, kes said negatiivse soolevähi diagnoosi ehk:

$$\text{Spetsiifilisus} = \frac{TN}{TN + FP}.$$

Logistilise regressiooni mudel väljastab iga indiviidi kohta tõenäosuse soolevähi esinemise kohta. Selleks, et inimest klassifitseerida haigeks või terveks, tuleb saadud tõenäosused mingi lävendi alusel jagada kahte klassi. Lävendist suurema tõenäosuse puhul määratakse inimene soolevähi diagnoosiga patsiendiks ning lävendist madalama väärtuse korral loetakse inimene terveks. Selleks, et kõige täpsemini eristada vähkkasvajaga patsiente tervetest, tuleb vaadata ja võrrelda mitmeid lävendeid.

Erinevate lävendite (*threshold*) korral leitud tundlikkuse ja spetsiifilisuse kombinatsioonide graafilisel kujutamisel saadakse ROC-kõver (*Receiver Operating Characteristic curve*). Iga lävendi korral jaotatakse andmestikus olevad indiviidid kahte gruppi - soolevähiga ja terved inimesed. Saadud tulemusi võrreldakse reaalselt teadaolevate andmetega, mille abil on võimalik tabeli 3 alusel leida fikseeritud lävendiga mudeli tundlikkus ja spetsiifilisus. ROC-kõvera $x$-teljel kujutatakse tõeste positiivsete prognooside määrasid ehk tundlikkuse väärtuseid ning $y$-teljel valepositiivsete määrasid ehk (1-spetsiifilisus) väärtuseid.

ROC-kõveraga mudeli headuse hindamiseks on oluline arvestada mudeli spetsiifilisusega, et mudel suudaks terveid ja vähkkasvajaga patsiente võimalikult täpselt eristada. Ideaalse tulemuse korral suudab mudel 100%-lise täpsusega prognoosida haigel inimesel soolevähki, tervel inimesel mitte. Sellisel juhul läbib ROC-kõver punkti (0;1). Kui tundlikkus ja spetsiifilisus on võrdsed väärtusega 0,5, on positiivse vähidiagnoosi määramine juhuslik ehk mudel ei sobi vähkkasvaja prognoosimiseks.

Mudeli üldistusvõime hindamiseks kasutatakse näitajat ROC-kõvera alusest pindalast ehk AUC (*Area Under the Curve*). Näitaja maksimaalne väärtus on 1, mis avaldub, kui mudel suudab 100%-lise täpsusega eristada terveid inimesi haigetest, st spetsiifilisus ja tundlikkus on mõlemad 100% ja seega valepositiivsete määr (1-spetsiifilisus) on null. Kui AUC väärtus on 0,5, siis soolevähi diagnoosi saanud inimeste osakaal on võrdne haiguse diagnoosi saanud tervete inimeste osakaaluga ehk mudel ei suuda terveid ja vähkkasvajaga inimesi piisavalt täpselt eristada.

## 4.3 Mikrobioomi esindava argumenttunnuse leidmine

Järgnevates alapeatükkides kirjeldatud mikrobioomi tunnuse väärtuse leidmine tugineb autorite Rivera-Pinto jt artiklil „Balances: a New Perspective for Microbiome Analysis“, kui ei ole teisiti viidatud [14].

Olgu indiviidi mikrobioomi arvukused tähistatud vektoriga $\mathbf{X} = (X_1, \ldots, X_p)$, kus $p$ tähistab unikaalsete bakterite arvu. *Selbal* metoodika kasutab logistilise regressiooni mudelisse mikrobioomi komponendi (MIL) arvutamiseks järgnevat valemit:

$$MIL(X_+, X_-) = \sqrt{\frac{k_+ k_-}{k_+ + k_-}} \log \frac{(\prod_{i \in I_+} X_i)^{1/k_+}}{(\prod_{j \in I_-} X_j)^{1/k_-}} \tag{4.5}$$

Kus $X_+$ ja $X_-$ on vektorist $\mathbf{X}$ *Selbal* metoodika poolt leitud kaks lõikumatut bakterite alamhulka indekseeritud vastavalt $I_+$ ja $I_-$ ning $k_+$ ja $k_-$ tähistavad vastavalt alamhulkade elementide arve. Tähis „+“ viitab murru lugejas olevatele elementidele ning „−“ murru nimetajas olevatele elementidele. Seega on mikrobioomi kaasav tunnus defineeritud normaliseeritud log-suhtena kahe bakterite grupi arvukuste geomeetrilisest keskmisest. Sellist konstruktsiooni nimetatakse kompositsionaalsete andmete teoorias „tasakaaluks“ (*balance*) [20].

**Näide**

Illustreerimaks MIL arvutamist valemi (4.5) abil, oletame, et tabeli 2 andmete põhjal osutus soolevähi prognoosimismudel kõige täpsemaks bakterite $X_1$, $X_3$ ja

$X_4$ korral, kusjuures lugejasse $X_+$ valitakse bakterid $X_1$ ja $X_4$ ning nimetajasse $X_-$ bakter $X_3$.

Indiviidi $S_1$ korral, kui $k_+ = 2$, $k_- = 1$ ja vastavad indeksite hulgad $I_+ = \{1, 4\}$, $I_- = \{3\}$, siis MIL väärtus absoluutse arvukuse korral avaldub valemi (4.5) põhjal järgnevalt:

$$\begin{aligned} MIL(X_+, X_-) &= \sqrt{\frac{2}{3}} \left( \frac{1}{2} (\log 10 + \log 100) - \log 60 \right) \\ &= \frac{\sqrt{6}}{6} (\log 10 + \log 100) - \sqrt{\frac{2}{3}} \log 60 \approx -0{,}52 \end{aligned}$$

ning suhtelise arvukuse korral

$$\begin{aligned} MIL(X_+, X_-) &= \sqrt{\frac{2}{3}} \left( \frac{1}{2} (\log 5 + \log 50) - \log 30 \right) \\ &= \frac{\sqrt{6}}{6} (\log 5 + \log 50) - \sqrt{\frac{2}{3}} \log 30 \approx -0{,}52 \end{aligned}$$

Analoogiliselt saab leida MIL väärtused indiviidi $S_2$ andmetel.

Tabel 4: MIL arvutamine bakterite arvukuse ja suhtelise arvukuse põhjal

| | MIL absoluutne arvukus | MIL Suhteline arvukus |
|---|---|---|
| Indiviid | $MIL([X_1, X_4], X_3)$ | $MIL([X_1, X_4], X_3)$ |
| $S_1$ | $\frac{\sqrt{6}}{6} (\log 10 + \log 100) - \sqrt{\frac{2}{3}} \log 60 \approx -0,52$ | $\frac{\sqrt{6}}{6} (\log 5 + \log 50) - \sqrt{\frac{2}{3}} \log 30 \approx -0,52$ |
| $S_2$ | $\frac{\sqrt{6}}{6} (\log 10 + \log 400) - \sqrt{\frac{2}{3}} \log 60 \approx 0,04$ | $\frac{\sqrt{6}}{6} (\log 2 + \log 80) - \sqrt{\frac{2}{3}} \log 12 \approx 0,04$ |

Tabelist 4 on näha, et mõlema indiviidi puhul säilitab MIL väärtuse arvutamiseks kasutatav valem (4.5) bakterite suhted nii absoluutse kui ka suhtelise bakterite arvukuse korral, mis on väga oluline mikrobioomi andmete analüüsimisel. Sellist tüüpi transformatsioone, kus kompositsiooni elemente või elementide funktsioone omavahel jagatakse, tuntakse kompositsionaalsete andmete teoorias üldise nimega log-suhte (*log-ratio*) transformatsioonidena [21]. Transformatsioonid aitavad toime

tulla sekveneerimisega kaasnevate piirangutega, kuid nende probleemiks on nullide olemasolu andmestikus, mille korral ei ole elementide jagatis ning logaritm arvutatavad.

### 4.3.1 Nullväärtuste asendamine

Valemi (4.5) rakendamiseks asendatakse esimese sammuna *Selbal* meetodika põhjal andmestiku nullväärtused. Vastavas paketis on implementeeritud mitmeid nullide asendamise meetodeid. Vaikimisi asendab *selbali* algoritm andmestikus esinevaid nullväärtuseid geomeetrilise Bayesi multiplikatiivse asendamise meetodil ($GBM$), mis säilitab andmestikus olevate bakterite suhted. Bakalaureusetöös kasutati nullide asendamiseks konstandiga asendamist, milleks oli minimaalseim positiivne väärtus andmestikus.

### 4.3.2 Samm 1: Kahe komponendi põhjal optimaalsete bakterite valimine

Esimese sammuna hindab *selbal* algoritm kõiki võimalikke „tasakaale“, mis on leitud ainult kahe bakteri põhjal. Seega valemist (4.5) tulenevalt vaadatakse kõikvõimalikke MIL väärtuseid kujul:

$$MIL_{ij} = \sqrt{\frac{1}{2}}[\log(X_i) - log(X_j)]\forall i, j \in \{1, \ldots, p\}, i \neq j \tag{4.6}$$

Kus $X_i$ on „tasakaalu“ lugejas olev bakter ning $X_j$ nimetajas olev bakter. Iga leitud $MIL_{ij}$ väärtuse põhjal testitakse selle sobivust soolevähi ennustamiseks. Valituks osutub kahe bakteri kombinatsioon, mille korral on AUC väärtus suurim ehk kombinatsioon, mis võimaldaks kõige täpsemini haigeid ja terveid inimesi klassifitseerida. Kui ühegi kombinatsiooni lisamine mudelisse ei paranda mudeli ennustusvõimet, jäetakse mikrobioomi kaasav tunnus mudelist välja.

Kuna algoritm kasutab kahe optimaalse bakteri leidmiseks kõikvõimalikke kahe bakteri kombinatsioone, kontrollib algoritm lisaks valemile (4.6) ka vastupidiseid jagatisi, kus nimetaja ja lugeja on vahetuses, ehk

$$MIL_{ji} = \sqrt{\frac{1}{2}}[\log(X_j) - log(X_i)] \forall i, j \in \{1, \ldots, p\}, i \neq j \tag{4.7}$$

Oluline on märkida, et valemist (4.6) erineb lugejat ja nimetajat vahetades saadav väärtus vaid märgi poolest ning *selbali* algoritm valib juhu, mille korral on kordaja logistilise regressiooni mudelis positiivne.

### 4.3.3 Samm 2: Täiendavate bakterite lisamine mikrobioomi tunnusesse

Kui eelnevas sammus leitakse kaks bakterit, mille põhjal arvutatud „tasakaalu" lisamine logistilise regressiooni mudelisse võimaldab soolevähki prognoosida, otsitakse sellele järgneval sammul baktereid, mille lisamine „tasakaalu" arvutamisse aitaks veelgi mudeli ennustusvõimet parandada. Sealjuures, eelnenud sammul valitud bakterid jäävad fikseerituks.

Olgu $MIL^1$ sammul 1 valitud „tasakaal", mis koosneb elementidest $X_i$ ja $X_j$:

$$MIL^1 = \sqrt{\frac{1}{2}}[\log(X_i) - log(X_j)], i \neq j \tag{4.8}$$

Siis saab tähistada sammul 1 „tasakaalu" mittevalitud tunnuseid vektoriga $X_L$, kus $X_L = \mathbf{X} \setminus \{X_i, X_j\}$. Teisel sammul lisatakse bakter $X_l$ hulgast $X_L$ esmalt „tasakaalu" lugejasse:

$$MIL_l^{2+} = \sqrt{\frac{1 \cdot 2}{2+1}} \cdot \left\{ \frac{1}{1+1} \left[ \log(X_i) + \log(X_l) \right] - \frac{1}{1} \log(X_j) \right\} \tag{4.9}$$

Analoogselt sammule 1 arvutatakse „tasakaal" nii, et bakter lisatakse murru nimetajasse:

$$MIL_l^{2-} = \sqrt{\frac{1 \cdot 2}{2+1}} \cdot \left\{ \frac{1}{1} \log(X_i) - \frac{1}{1+1} \left[ \log(X_j) + \log(X_l) \right] \right\} \tag{4.10}$$

Selliselt luuakse iga hulga $X_l$ tunnuse jaoks kaks „tasakaalu“ $MIL_l^{2+}$ ning $MIL_l^{2-}$, mida testitakse ennustusmudelis. Mikrobioomi liikme MIL väärtus, mis maksimeerib optimiseerimise kriteeritumit ehk mille põhjal arvutatud mudeli täpsus ehk AUC on maksimaalne, defineerib uue „tasakaalu“ $MIL^2$. Algoritm lõpetab töö juhul, kui mudeli täpsust kirjeldavat AUC väärtust ei ole võimalik parandada täiendava bakteri lisamisel lugejasse või nimetajasse.

### 4.3.4 Optimaalse bakterite arvu valimine ristvalideerimise abil

*Selbal* meetodis rakendatakse ristvalideerimist mudeli ehitamise protsessis kahel eesmärgil. Esiteks, et leida optimaalne elementide arv „tasakaalu“ arvutamiseks ning teiseks, et hinnata selle robustsust andmetes. Selleks kasutatakse iteratiivset $k$-kordset ristvalideerimist. Optimaalne bakterite arv tagab mudeli parema täpsuse, bakterite arvu piiramine aitab ära hoida mudeli ala- või ülesobitamist. Vaikimisi on „tasakaalu“ arvutamisel maksimaalne kaasatavate bakterite arv 20.

Ristvalideerimise põhimõte on esmalt jaotada juhuslikult kogu andmestik kaheks lõikumatuks alamandmestikuks ehk test- ja treeningandmestikuks. Andmestiku treening ja testandmestikuks jagamist tehakse $M \times k$ korda, kus $k$ näitab, mitmeks osaks treeningandmestik jagatakse ning $M$ näitab iteratsioonide arvu erinevate $k$-kordsete jaotuste tegemiseks. Iga treening- ja testandmestikuks jagamise korral leitakse esmalt treeningandmetel „tasakaalud“, mis moodustavad alates kahel bakteril põhinevast tasakaalust täiendavate bakteri lisandumisel jada $MIL(2), \ldots, MIL(20)$. Seejärel rakendatakse saadud regressioonmudeleid koos vastavate tasakaaludega testandmetele ning arvutatakse iga $MIL(j)$ korral AUC väärtus: $AUC(2), \ldots, AUC(20)$. Protseduuri korratakse $M \times k$ korda. Optimaalseks bakterite arvuks loetakse väärtus $t$, mille korral keskmine $AUC(t)$ väärtus üle kõi-

kide testandmestike on ühe standardhälbe kaugusel maksimaalsest keskmisest AUC väärtusest.

# 5 Andmete analüüs

Töös kasutatud andmed pärinevad autorite Wirbel jt artiklist „Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer“. Artiklis analüüsitakse viie populatsiooni soolevähi patsientide ja kontrollide andmeid - Prantsusmaa, Austria, Saksamaa, Hiina ja USA. Artikli eesmärgiks oli tuvastada, millised bakteriliigid on seotud soolevähiga ning kui hästi töötavad soolevähki ennustavad mudelid erinevates populatsioonides. [22]

Uuringus osalenute andmed on kogutud vaadeldavates riikides erinevalt - erinevused on valikuuringu disainis, väljaheite proovide säilitamises kui ka DNA sekveneerimiseks kasutatud meetodites. Kõikide uuringus osalenud soolevähi diagnoosiga patsientide mikrobioomi andmed on kogutud enne vähiraviga alustamist, vältimaks ravimeetodite mõju soolebakterite mitmekesisusele. [22]

Austria ja Prantsusmaa uuringus on väljaheite proovid kogutud enne kolonoskoopia uuringut. USA andmetes on proov kogutud kõigil uuringus osalejatel peale kolonoskoopia uuringut, Saksamaa kontrollgrupi proovid on kogutud enne sooleuuringut ja soolevähi patsientidel peale uuringut ning Hiina uuringus on 60%-l osalejatest mikrobioomi andmed kogutud peale sooleuuringut.

Kolonoskoopia võib mõjutada inimese soolestiku mikrobioomi mitmekesisust, kuna uuringueelselt toimub soole täielik puhastamine. Lisaks on oluline märkida, et USA andmete kogumise puhul on teada, et väljaheite proovid olid kuivkülmutatud 25 aastat enne DNA eraldamise ning sekveneerimise protseduuri, mis võib samuti mõjutada sekveneerimise tulemusel saadud bakterite liigirikkust ning mikrobioomi struktuuri. [22]

Ennustusmudelite koostamisel kasutati kvalitatiivseid ja kvantitatiivseid tunnuseid.

Kvalitatiivseteks tunnuseks on soolevähki kirjeldav tunnus (CRC - on soolevähk, CTR - ei ole soolevähki) ja inimese sugu (mees, naine). Arvulisteks tunnusteks on inimese vanus täisaastates, DNA sekveneerimisel saadud bakterite andmed suhtarvudena, kokku 849 bakteriliiki ning kehamassiindeks (KMI), mis on arvutatud valemiga

$$KMI = \frac{kehakaal(kg)}{pikkus^2(m)}. \tag{5.1}$$

## 5.1 Kirjeldav analüüs

Uuritavas andmestikus oli kokku 575 inimese andmed. Naiste osakaal uuringus on 40% ehk 230 on naised ja 60% osalenutest on mehed, koguarvuga 345. Soolevähi diagnoosiga (CRC) on nendest ligikaudu $49,6\%$ ehk 285 patsienti ning kontrollgrupi (CTR) moodustavad $50,4\%$ uuringus osalenutest (vt Tabel 12).

Keskmine uuringus osaleja on 63-aastane, kehamassiindeksiga 25,3. Andmestikus on kõige noorem indiviid 25-aastane ja kõige vanem 90-aastane, mediaanvanus on 64 aastat. Kehamassiindeksi (edaspidi KMI) minimaalne väärtus on 13,3 ja maksimaalne väärtus on 40, KMI mediaan on 24,7. Vanuse standardhälve on 11 aastat ja kehamassiindeksi hajuvus on 3,9 ühikut (vt Tabel 13 ). Seega on hajuvus ümber keskmise vanuse ja kehamassiindeksi üsna suur. Kuigi KMI on soolevähi riskiteguriks, siis mediaanväärtuse põhjal võib öelda, et vähemalt 50%-l uuringus osalenutel on KMI väiksem kui 24,7, mis viitab sellele, et vähemalt pooled osalenutest ei ole ülekaalulised.

Kontrollrühma (CTR) moodustavad terved inimesed, kellest 42,4% moodustavad naised ja 57,6% mehed. Soolevähi diagnoosiga (CRC) meeste osakaal on suurem kui naiste - positiivse diagnoosi saanud patsientidest moodustavad naised $37,5\%$ ja mehed vastavalt 62,5% (vt Tabel 16). Kõige noorem soolevähi diagnoosi saanud patsient on 31-aastane, kõige vanem on 90-aastane. Mediaanvanus vähkkasvajaga

inimesel on 65 aastat (vt Tabel 14).

Keskmise vanuse erinevus vähkkasvajaga ja tervetel inimestel on kaks aastat, mediaanvanuse erinevus üks aasta. Seega vähidiagnoosi saanud patsientide keskmine vanus on mõnevõrra kõrgem tervete inimeste keskmisest, kuid erinevus on väike (vt Tabel 15). Kehamassiindeksi mediaan on vähkkasvajaga inimestel 25, seega 50% positiivse diagnoosi saanutest on ülekaalulised ja 50% juhtudest on indeks väiksem kui 25. Kontrollgrupis on indeksi mediaan 25,3 ning seega ei ole gruppidevahelised erinevused märkimisväärselt varieeruvad.

Järgnevalt kirjeldatakse uuringus osalenute keskmist vanust, kehamassiindeksit ning soolist jaotuvust riigiti (Tabel 5).

Tabel 5: Kirjeldav statistika riikide ja diagnooside jaotuses

| | N | | Vanus | | KMI | | Naiste osakaal (%) | |
|---|---|---|---|---|---|---|---|---|
| Riik | CTR | CRC | CTR | CRC | CTR | CRC | CTR | CRC |
| Austria | 63 | 46 | 67 (6, 37) | 67 (10, 91) | 27, 6 (3, 74) | 26, 5 (3, 54) | 26 (41) | 18 (39) |
| Hiina | 54 | 74 | 62 (5, 67) | 66 (10, 60) | 23, 5 (2, 96) | 24, 0 (3, 16) | 21 (39) | 26 (35) |
| Prantsusmaa | 61 | 53 | 61 (11, 39) | 67 (10, 88) | 24, 7 (3, 17) | 25, 6 (5, 17) | 33 (54) | 24 (45) |
| Saksamaa | 60 | 60 | 58 (11, 08) | 63 (12, 64) | 24, 9 (3, 20) | 26, 2 (4, 00) | 28 (47) | 24 (40) |
| USA | 52 | 52 | 61 (11, 03) | 62 (13, 58) | 25, 3 (4, 28) | 24, 9 (4, 25) | 15 (29) | 15 (29) |

*Diagnoos CRC - soolevähiga patsient, CTR - terve inimene. Kehamassiindeksi KMI ning vanuse jaoks on raporteeritud keskmine ja standardhälve.*

Kõige rohkem indiviide pärines Hiinast ($n = 128$), kõige väiksema valimimahuga oli USA ($n = 104$). Saksamaa ja USA andmetes on soolevähi diagnoosiga patsientide ja kontrollgrupi osakaalud valimis võrdsed, Austria ja Prantsusmaa uuringuandmetes on kontrollgrupis inimesi rohkem kui vähkkasvaja diagnoosiga ning Hiina uuringus on soolevähiga patsientide osakaal kontrollgrupist oluliselt suurem.

Riikide võrdluses on kõrgeim keskmine kehamassiindeks Austrias (27,1), mis viitab üsna tugevale ülekaalule. Samuti on uuringus osalenud riikidest Austrias kõrgeim keskmine vanus - 67 aastat. Hiinas on keskmine kehamassiindeks erinevalt teistest riikidest madalam kui 25 ehk keskmine uuringus osalenu ei ole ülekaaluline. Üldiselt

on naiste osakaal riikide võrdluses sarnane. Naiste osakaal on valdavalt väiksem kui meeste, jäädes mõlema grupi puhul vahemikku 35−47%. Erandlikult on USA uuringus naiste osakaal vaid 29% ning Prantsusmaa kontrollgrupis on naiste osakaal 54%.

# 6 Selbal mudeli rakendamine

Ennustusmudelite loomine jaguneb kaheks: esmalt treenitakse logistilisel regressioonil põhinev mudel mikrobioomi komponenti sisaldava tunnusega ühe populatsiooni andmestiku peal, seejärel rakendatakse saadud mudelit ülejäänud nelja populatsiooni andmetele. Mudeli sobitamiseks teistele populatsioonidele arvutatakse esmalt testpopulatsioonis mikrobioomi tunnus (MIL) kasutades selleks mudeli poolt treeningpopulatsioonis leitud baktereid.

## 6.1 Mudeli koostamisel kasutatud paketid

Mudelite loomiseks kasutati rakendustarkvara R paketi *Selbal* funktsiooni *Selbal.cv* eesmärgiga leida optimaalne bakterite arv ja bakteriliikide loend individuaalse mikrobioomi tunnuse (MIL) väärtuse arvutamiseks.

Funktsiooni sisend on mikrobioomi andmetest koosnev maatriks, soolevähki indikeeriv vektor ning ülejäänud soovitud kovariaatidest koosnev maatriks. Iteratiivse $k$-kordse ristvalideerimise parameetritena valiti iteratsioonide arvuks $M = 4$ ja korduste arvuks $k = 5$. Parameetriks $C$, mis kirjeldab maksimaalset „tasakaalu“ arvutamiseks valitud bakterite arvu, määrati $C = 20$.

*Selbal* funktsioonis on implementeeritud mitmed algoritmi tööks vajalikud nullide asendamise võimalused. Bakalaureusetöös kasutatakse selleks nullide väikese väärtusega asendamist. *Selbal.cv* väljastab pärast algoritmi töö lõppu sobitatud logistilise regressiooni mudeli objekti, ristvalideerimisel saadud AUC väärtused ja mikrobioomi tunnuse arvutamiseks vajaliku bakterite loendi ning märke, kas bakter kuulus MIL väärtuse arvutamisel „tasakaalu“ lugejasse ($X_+$) või nimetajasse ($X_-$).

Mudeli täpsuse ja prognoosivõime hindamiseks kasutati ROC-kõvera alust pindala ehk AUC väärtust. ROC-kõvera kujutamiseks vajalike suuruste arvutamiseks kasutati R paketti *pROC* funktsiooni *roc*, mille parameetriteks on mudeli põhjal

prognoositud soolevähi esinemise tõenäosused ning indiviidi tegelik soolevähi diagnoos. ROC-kõvera aluse pindala ehk AUC väärtuse arvutamiseks kasutati analoogselt paketi *pROC* funktsiooni *auc*. Tulemuste visualiseerimiseks kasutati R pakette *ggplot2* ning *ggpubr*.

## 6.2 Soolevähi ennustusmudelid

Järgnevalt kirjeldatakse viit *Selbal* metoodika abil loodud logistilise regressiooni mudelit kasutades Austria, Hiina, Prantsusmaa, Saksamaa ja USA patsientide andmeid. Lisaks kirjeldatakse iga riigi jaoks bakteriliikide loetelu, mida *Selbal* meetod kasutas mudelitesse mikrobioomi liikme MIL väärtuse arvutamiseks. Mudelite parameetrite olulisuse hindamiseks kasutatakse olulisusnivood $\alpha = 0{,}05$.

Mudeli interpreteerimine põhineb peatükil 4. Illustreerimaks mudeli abil soolevähi esinemise tõenäosuse arvutamist ja mudeli argumentide interpreteerimist kasutatakse Austria andmetel loodud mudeli hinnanguid. Teiste populatsioonide mudelite tõlgendamine on analoogiline.

**Austria**

Tabelis 6 on toodud *Selbal* meetodi abil saadud logistilise regressiooni mudeli väljund Austria andmetel.

Tabel 6: Austria andmelt saadud Selbal mudeli väljund

| Koefitsent | Hinnang | Std | $e^{\text{Hinnang}}$ | p-väärtus |
|---|---|---|---|---|
| VABALIIGE | $-1,52$ | $2,82$ | $0,22$ | $0,5896$ |
| VANUS | $0,05$ | $0,03$ | $1,05$ | $0,1258$ |
| KMI | $-0,17$ | $0,09$ | $0,84$ | $0,0538$ |
| MIL | $0,78$ | $0,15$ | $2,18$ | $< 4 \cdot 10^{-7}$ |

Austria andmetel loodud soolevähi ennustusmudel avaldub kujul:

$$Logit(\pi) = -1,52 + 0,05 \cdot VANUS - 0,17 \cdot KMI + 0,78 \cdot MIL \qquad (6.1)$$

Mudelis osutus olulisuse nivool $\alpha = 0,05$ statistiliselt oluliseks ainult mikrobioomi tunnus MIL. Mikrobioomi kaasav tunnus MIL sisaldab lugejas baktereid *Gemella morbillorum*, *Prevotella copri*, klassifitseerimata liiki 1 perekonnast *Clostridiales* ja nimetajas klassifitseerimata liiki 2 perekonnast *Clostridiales*. Järgnevalt kirjeldatakse Austria mudeli abil soolevähi esinemise tõenäosuse arvutamist ja mudeli MIL argumendi interpreteerimist.

Näiteks indiviidil, kes on 70-aastane, kehamassiindeksiga 25 ning mikrobioomi tunnuse väärtusega 4,72, on soolevähi positiivse diagnoosi tõenäosus arvutatav järgnevalt:

$$\pi = \frac{e^{0,05 \cdot 70 - 0,17 \cdot 25 + 0,78 \cdot 4,72 - 1,52}}{1 + e^{0,05 \cdot 70 - 0,17 \cdot 25 + 0,78 \cdot 4,72 - 1,52}} = \frac{e^{1,412}}{1 + e^{1,412}} = 0,804$$

Mudeli parameetrite interpreteerimine tugineb peatükile 4. Kuivõrd mudelis osutus oluliseks vaid MIL tunnus, ei paku teiste argumentide interpreteerimine niivõrd huvi. Seega kui võrreldakse kaht indiviidi, kellest ühel on MIL väärtus ühe ühiku võrra suurem, siis teiste argumentide samaks jäädes, on sellel inimesel šanss soolevähki haigestuda 2,18 korda suurem võrreldes teise indiviidiga (vt Tabel 6).

**Hiina**

Tabelis 7 on toodud *Selbal* meetodi abil saadud logistilise regressiooni mudeli väljund Hiina andmetel.

Tabel 7: Hiina andmelt saadud Selbal mudeli väljund

| Koefitsent | Hinnang | Std | $e^{\text{Hinnang}}$ | p-väärtus |
|---|---|---|---|---|
| VABALIIGE | $-3,29$ | $2,49$ | $0,04$ | $0,187$ |
| VANUS | $0,05$ | $0,03$ | $1,05$ | $0,109$ |
| KMI | $0,11$ | $0,09$ | $1,12$ | $0,193$ |
| MIL | $0,57$ | $0,10$ | $1,77$ | $< 6,5 \cdot 10^{-4}$ |

Hiina treeningandmetel loodud soolevähi ennustusmudel avaldub kujul

$$Logit(\pi) = -3,29 + 0,05 \cdot VANUS + 0,11 \cdot KMI + 0,57 \cdot MIL \tag{6.2}$$

Tabelist 7 on näha, et mudelis on oluline vaid mikrobioomi komponenti kirjeldav tunnus. Hiina mudelis kasutati MIL väärtuse arvutamiseks lugejas baktereid *Parvimonas micra*, klassifitseerimata *Bacteroidaceae* ja nimetajas klassifitseerimata bakteriliiki 3 perekonnast *Clostridiales* ning klassifitseerimata liiki 1 perekonnast *Faecalibacterium*.

**Prantsusmaa**

Tabelis 8 on toodud *Selbal* meetodi abil saadud logistilise regressiooni mudeli väljund Prantsusmaa andmetel.

Tabel 8: Prantsusmaa andmelt saadud Selbal mudeli väljund

| Koefitsent | Hinnang | Std | $e^{\text{Hinnang}}$ | p-väärtus |
|---|---|---|---|---|
| VABALIIGE | $-3,73$ | $2,43$ | $0,02$ | $0,1255$ |
| VANUS | $0,08$ | $0,03$ | $1,08$ | $0,0058$ |
| KMI | $0,03$ | $0,07$ | $1,03$ | $0,5982$ |
| MIL | $0,72$ | $0,16$ | $2,05$ | $< 7,3 \cdot 10^{-7}$ |

Prantsusmaa treeningandmetel loodud soolevähi ennustusmudel avaldub kujul

$$Logit(\pi) = -3,73 + 0,08 \cdot VANUS + 0,03 \cdot KMI + 0,72 \cdot MIL \tag{6.3}$$

Tabeli 8 põhjal on näha, et mudelis osutus oluliseks lisaks MIL argumendile ka inimese vanus. Vanuse olulisuse põhjuseks võib olla soolevähiga patsiendidele sarnase kehamassiindeksiga kontrollgrupi inimese sobitamine. Tuginedes tabelile 5 on näha, et kontrollgrupi ja soolevähiga patsientide kehamassiindeksid on suhteliselt sarnased ning vanused on erinevad. Prantsusmaa mudeli loomisel kasutatati MIL arvutamiseks lugejas klassifitseerimata bakterit *Dialister* ning nimetajas kasutati klassifitseerimata liiki 3 perekonnast *Clostridiales*, *Bacteroides intestinalis*'t ja klassifitseerimata liiki 4 perekonnast *Clostridiales*.

**Saksamaa**

Tabelis 9 on toodud *Selbal* meetodi abil saadud logistilise regressiooni mudeli väljund Saksamaa andmetel.

Tabel 9: Saksamaa andmelt saadud Selbal mudeli väljund

| Koefitsent | Hinnang | Std | $e^{\text{Hinnang}}$ | p-väärtus |
|---|---|---|---|---|
| VABALIIGE | $-6,10$ | $2,48$ | $0,002$ | $0,0139$ |
| VANUS | $0,05$ | $0,02$ | $1,05$ | $0,0157$ |
| KMI | $0,12$ | $0,08$ | $1,13$ | $0,1167$ |
| MIL | $0,48$ | $0,09$ | $1,62$ | $< 2,15 \cdot 10^{-8}$ |

Saksamaa treeningandmetel loodud soolevähi ennustusmudel avaldub kujul

$$Logit(\pi) = -6,10 + 0,05 \cdot VANUS + 0,12 \cdot KMI + 0,48 \cdot MIL \tag{6.4}$$

Sarnaselt eelmainitud treeningandmestikega on ka Saksamaa andmetel loodud mudelis oluline mikrobioomi kirjeldav tunnus ning sarnaselt Prantsusmaa mudelile, on

oluline inimese vanus. Arvestades tabeli 5 andmetega, võib samuti oletada, et soolevähi patsienditele on vastavusse seatud sarnase kehamassiindeksiga kontrollgrupi inimene, mistõttu ei ole KMI mudelis oluline. Saksamaa mudeli loomisel kasutatati lugejas bakterit *Parvimonas micra* ning nimetajas klassifitseerimata bakterit liik 1 perekonnast *Clostridium*, klassifitseerimata liiki 2 perekonnast *Clostridium*.

**USA**

Tabelis 10 on toodud *Selbal* meetodi abil saadud logistilise regressiooni mudeli väljund USA andmetel.

Tabel 10: USA andmelt saadud Selbal mudeli väljund

| Koefitsent | Hinnang | Std | $e^{\text{Hinnang}}$ | p-väärtus |
|---|---|---|---|---|
| VABALIIGE | $6,04$ | $2,35$ | $419,9$ | $0,0102$ |
| VANUS | $0,01$ | $0,02$ | $1,01$ | $0,6317$ |
| KMI | $-0,14$ | $0,07$ | $0,87$ | $0,0369$ |
| MIL | $1,52$ | $0,33$ | $4,57$ | $< 3,9 \cdot 10^{-7}$ |

USA treeningandmetel loodud soolevähi ennustusmudel avaldub kujul

$$Logit(\pi) = 6,04 + 0,01 \cdot VANUS - 0,14 \cdot KMI + 1,52 \cdot MIL \qquad (6.5)$$

Tabeli 10 põhjal on samuti statistiliselt oluline mikrobioomi komponenti sisaldav tunnus ning erinevalt teiste populatsioonide mudelitest on oluline ka kehamassiindeks. USA mudeli loomisel kasutatati lugejas klassifitseerimata bakterit perekonnast *Dialister* ja nimetajas kasutati klassifitseerimata liiki 3 perekonnast *Clostridium*, klassifitseerimata liiki 5 perekonnast *Clostridiales*, *Clostridium innocuum*'i, *Mycoplasma sp.* ja *Alistipes sp.*

Kõigi populatsiooni andmestikel loodud soolevähi ennustusmudelites osutus statistiliselt oluliseks mikrobioomi kirjeldav tunnus. Inimese vanus osutus oluliseks vaid Saksamaa ja Prantsusmaa mudelites ning USA mudelis inimese kehamassiindeks.

Vanuse ja kehamassiindeksi olulisus võib olla seotud vastavates populatsioonides andmete kogumismeetoditega. Tuginedes tabelile 5 on võimalik, et Saksamaa ja Prantsusmaa andmetes on soolevähi patsienditele vastavusse seatud kontrollgrupp KMI põhjal, mis selgitaks vanuse olulisust mudelis ning USA andmete põhjal võib olla kontrollgrupp valitud vanuse põhjal, mis viitaks KMI olulisusele mudelis.

Kõikide populatsioonide mudelites on mikrobioomi tunnuse hinnangud positiivsed ning minimaalne eksponetsiaalne hinnang on 1,62 (vt Tabel 9). Positiivne hinnang viitab samasuunalisele seosele ehk MIL kasvades suureneb haigestumise šanss ning MIL vähenedes šanss muutub väiksemaks. Siinkohal on märkimisväärne, et *Selbal* meetod kasutab mudelite loomisel bakterite tunnuseid vaid siis, kui selle abil on võimalik soolevähki paremini prognoosida. Seega, et mudel valis igas populatsioonis mikrobioomi liikme mudelisse, sh iga kord rohkem kui kahe bakteriga, näitab, et mikrobioomil on seos soolevähiga, mida saab kasvaja diagnoosimiseks kasutada. Täiendavalt on oluline uurida mudelite üldistusvõimet ning MIL väärtuse arvutamiseks leitud bakterite seoseid soolevähiga erinevates populatsioonides.

## 6.3 Mudelite rakendamine teistele populatsioonidele

Mudelite üldistatavuse hindamiseks on oluline uurida mudelite ennustusvõimet ka teistes populatsioonides. Selleks rakendatakse peatükis 6.2 leitud mudeleid teiste populatsioonide andmetele ning uuritakse mudeli headuse näitaja AUC põhjal erinevate mudelite ennustusvõimet. Joonis 1 kirjeldab saadud tulemusi. ROC-kõverad, mille põhjal AUC on arvutatud, on toodud Lisas 2.


Joonis 1: Ennustusmudelite AUC väärtus erinevates populatsioonides

*Populatsioonide tähistused: GER-Saksamaa, FRA- Prantsusmaa, CHI-Hiina, AUS-Austria. Vertikaalteljel on kujutatud populatsioonide treeningandmestikud ning horisontaalteljel kujutatakse vastaval treeningandmestikul sobitatud mudeli ennustusvõimet testandmetele. Juhul, kui treeningandmestik on sama, mis testandmestik, raporteeritakse ristvalideerimisel saadud AUC keskmist väärtust.*

Joonise 1 põhjal on erinevate populatsioonide andmetel treenitud mudelite üldistusvõime üsna varieeruv, kuid ootuspärane, sest mudelite loomisel kasutatud riikide andmed on kogutud erinevate meetoditega, mis mõjutab oluliselt mudeli üldistusvõimet erinevates populatsioonides. Samuti võivad mikrobioomi puhul mängida rolli populatsioonide erinevad kultuurilised taustad ning harrastatav eluviis.

Ühe populatsiooni siseselt mudelit testides ning võrreldes tulemusi teistel popu-

latsioonidel näidatuga, osutusid AUC väärtused keskmisest kõrgemaks nelja riigi andmetel - Austria, Prantsusmaa, Saksamaa ja Hiina. USA andmetel treenitud mudeli populatsioonisisene ennustusvõime osutus madalamaks kui sama mudeli rakendamisel teiste populatsioonide testandmetele (*v.a* Austria).

Parima tulemuse saavutas Saksamaa mudel Hiina testandmetel, AUC väärtusega 0,843, mis on täpsem tulemus kui Saksamaa enda andmetel. Jooniselt 1 on näha, et Hiina ja Saksamaa kõrge üldistusvõime on mõlemasuunaline - nii Saksamaa kui Hiina treeningandmetel loodud mudelid suudavad üksteise populatsioonide andmetel paremini soolevähki prognoosida kui riigi enda andmetel. Kõrge AUC väärtus võib olla tingitud sellest, et MIL väärtuse arvutamisel kasutati nii Saksamaa kui ka Hiina andmetel lugejas bakterit *Parvimonas micra*.

Veendumaks, et leitud mikrobioomi tunnus MIL kirjeldab erinevusi ennustusvõimes, uuriti täiendavalt t-testi abil, kas mudelite leitud MIL väärtusted eristavad erinevates populatsioonides soolevähiga indiviide kontrollidest. Selgub, et Hiina ja Saksamaa puhul võib mudelite ülekantavust kirjeldada asjaolu, et mõlema populatsiooni põhjal leitud MIL väärtused suudavad mõlemas kohordis selgelt eristada kontrolle soolevähi patsientidest (vt Lisa 3). Hiina andmetelt leitud mudeli rakendamisel on jooniselt 1 näha Prantsusmaa ja Saksamaaga sarnast tulemust - mudel ennustab hästi Prantsusmaa, Saksamaa ja Hiina andmetel, kuid mitte Austria ega USA populatsioonile. Hea üldistatavus Hiina ja Prantsusmaa vahel võib olla samuti seotud leitud mikrobioomi tunnusega - mõlemal juhul kasutati MIL väärtuse arvutamisel nimetajas klassifitseerimata liiki 3 perekonnast *Clostridiales*.

Erandlikult on Prantsusmaa treeningandmetel loodud mudeli AUC väärtus Austria andmetele rakendades 0,423, mis viitab sellele, et Prantsusmaa andmetel treenitud mudel ei sobi kindlasti Austria patsiendidele soolevähi prognoosimiseks.

Treeningandmete põhjal loodud ennustusmudelites kasutatud sagedaseimad bakteriliigid on klassifitseerimata liigid 1-5 perekonnast *Clostridiales*, mis esinesid Austria, Hiina, Prantsusmaa ja USA mudelites. Kuigi Prantsusmaa ja USA mudelites

oli ühine bakteriliik klassifitseerimata liik perekonnast *Dialister*, ei õnnestunud populatsioonide vahel head üldistusvõimet saavutada.

Võrreldes teiste populatsioonides väljavalitud bakteritega oli USA mudelis kõige rohkem unikaalseid baktereid, mis kindlasti võis mõjutada ennustusmudeli üldistuvust Prantsusmaa andmetele. Lisaks võib ennustustäpsust mõjutada asjaolu, et Prantsusmaa mikrobioomi andmed koguti enne kolonoskoopia uuringut ning USA andmed on kogutud uuringu järgselt.

USA mudeli madal üldistusvõime teistes populatsioonides võib olla mõjutatud ka proovide pikaajalisest kuivkülmutamisest ja sooleuuringu järgsest andmete kogumisest, mis võis mõjutada sekveneerimise tulemusel saadud bakterite mitmekesisust ja suhtelist arvukust. Samuti tasub mõelda asjaolule, et kuna inimese mikrobioom on pidevalt mõjutatud ümbritsevast keskkonnast, siis 25-aasta tagune soolebakterite kooslus ja mitmekesisus võib toitumisharjumuste ja elustiili muutusest tulenevalt erineda.

Tulemustele tuginedes ei ole mõistlik ennustusmudeli valikul lähtuda AUC keskmisest väärtusest üle testandmestike, sest populatsioone võrreldes on mudeli prognoositäpsused väga erinevad. Võib osutuda, et ühe populatsiooni andmetel loodud mudel sobib soolevähki ennustama paremini kui selle sama populatsiooni andmetel nagu selgus Saksamaa ja Hiina mudelite põhjal. Seetõttu tuleks võrrelda konkreetsete populatsioonide sobivust arvestades andmete kogumisest tulenevaid erinevusi ja mikrobioomi omadusi. Seda kinnitab asjaolu, et sama populatsiooni siseselt töötasid mudelid märksa paremini, sest ühesuguse uuringudisaini ja sekveneerimise meetodiga andmed on sarnasemad ja see võimaldab vähendada andmete kogumise protseduuride varieeruvusest tingitud mõjusid. Ennustusmudelite puhul loetakse AUC väärtust 0,7 mõõdukaks täpsuseks, kuid reaalsete andmete korral on vajalik oluliselt kõrgem täpsus.

## 6.4 Tulemuste võrdlus originaalartikliga

Wirbel jt artiklis „Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer“ kasutati samadel eesmärkidel masinõppealgoritmi *LASSO* regressiooni bakterite suhtelisel sagedusel [22]. Artiklis raporteeritud keskmine treeningmudelite üldistusvõime näitaja AUC oli *LASSO* meetodi puhul vahemikus 0,65−0,88, *Selbal* meetodil jääb näitaja vahemikku 0,59−0,69 (vt Joonis 1). Seega mudelite keskmised tulemused on madalamad võrreldes *LASSO* meetodiga. Kuigi *Selbali* mudelite täpsuse hinnangud on madalamad, on võrreldes artikliga näha siiski samu trende.

Kõige madalama AUC väärtustega on Austria treeningandmetel loodud mudeli tulemused, kusjuures ennustustäpsus Austria enda testandmetel on oluliselt kõrgem selle täpsusest teistes populatsioonides. Kõige parema keskmise täpsuse saavutas Saksamaa mudel, kusjuures sarnaselt *Selbali* meetodiga, oli ka artiklis Saksamaa mudel Hiina andmetel kõige kõrgema AUC väärtusega. Samuti osutus paremuselt teiseks mudeliks Hiina andmetel treenitud mudel. Suurimaks erinevuseks mudelite täpsuse võrdlemisel oli USA mudeli kõrge ennustusvõime populatsioonisiseselt, väärtusega 0,76. Vastav *Selbali* mudeli täpsus oli 0,64, kuid mõlemal juhul oli USA mudeli keskmine ennustustäpsus teistes populatsioonides kõrgem kui USA enda andmetel. *Selbal* metoodikaga loodud mudelite madalama täpsuse põhjuseks võib olla väheste argumentide kaasamine mudelisse. Artiklis kasutati lisaks ka elustiili ja haigustega seotud tunnuseid, näiteks suitsetamise staatus, diabeedi olemasolu ja taimetoitluse tunnust.

Lisaks *LASSO* mudeli kasutamisele uuriti Wirbel jt artiklis meta-analüüsi kasutades, milliste liikide arvukus on soolevähiga seotud. Tabel 11 kirjeldab *Selbal* mudelitesse valitud liikide kattuvust meta-analüüsis esinenud bakteritega.

Tabel 11: *Selbal* meetodil saadud mudelite poolt kasutatud bakteriliigid

| Populatsioon | Bakter lugejas | Bakter nimetajas |
|---|---|---|
| Austria | ***Gemella morbillorum*** | ***Clostridiales (2)*** |
| | *Prevotella Copri* | |
| | ***Clostridiales (1)*** | |
| Hiina | ***Parvimonas micra*** | ***Clostridiales (3)*** |
| | *Bacteroidaceae* | *Faecalibacterium* |
| Prantsusmaa | ***Dialister*** | ***Clostridiales (3)*** |
| | | *Bacteroides intestinalis* |
| | | ***Clostridiales (4)*** |
| Saksamaa | ***Parvimonas micra*** | *Clostridium (1)* |
| | | *Clostridium (2)* |
| USA | ***Dialister*** | *Clostridium (3)* |
| | | ***Clostridiales (5)*** |
| | | *Clostridium innocuum* |
| | | *Mycoplasma sp.* |
| | | *Alistipes sp.* |

***Paksus kirjas*** *on märgitud töös leitud bakteriliigid ning nende kattuvus Wirbel jt artiklis raporteeritud liikidega, sulgudes olev number viitab klassifitseerimata bakteriliigile.*

Meta-analüüsi tulemusel leiti artiklis 29 erinevat bakteriliiki, mille arvukus oli soolevähiga inimestel oluliselt erinev kontrollidest. *Selbali* metoodikaga õnnestus erinevates mudelites mikrobioomi komponenti kaasata 18 unikaalset bakterit, kusjuures 8 (44,4%) bakterit kattus artiklis leituga (Tabel 11).

Üldiselt on *Selbal* ja *LASSO* metoodika abil loodud mudelid hästi võrreldavad, mõlemal juhul on märgata ühesuguseid trende. Olenemata sellest, et *Selbal* metoodika mudelite ennustustäpsused on madalamad, on seosed erinevate populatsioonide mudelite sobivuses sarnased. Erisused mudelite täpsusel võivad olla tingitud *Selbal* mudelites vähemate argumentide kaasamisest ja kindlasti mõjutab ennustustäpsust oluliselt vähemate bakteriliikide kaasamine.

# Kokkuvõte

Bakalaureusetöö eesmärk oli uurida mikrobioomi kasutamise võimalusi soolevähi diagnoosimiseks. Lisaks leida soolevähi ennustusmudelid, kasutades lisaks vähkkasvaja riskiteguritele mikrobioomi sisaldavat tunnust. Veel sooviti uurida, kui täpselt suudavad mudelid erinevates populatsioonides soolevähi esinemist prognoosida. Soolevähi mudelite loomiseks kasutati spetsiaalselt mikrobioomi andmete analüüsimiseks väljatöötatud metoodikat *Selbal*.

Tulemused näitasid, et *Selbal* metoodika võimaldab mikrobioomi põhjal soolevähki ennustada, kuid saadud mudelite üldistusvõime erinevate populatsioonide andmetel on madal. Mudeli täpsuse näitaja AUC jäi erinevate populatsioonide andmetel vahemikku $0,59-0,69$, mis ei ole piisav täpsus, et mudeleid soolevähi diagnoosimiseks praktikas rakendada. Kehv üldistusvõime võib olla seotud erinevate populatsioonide mikrobioomi proovide kogumisest tingitud erinevustest. Seda kinnitas ka tulemus, et ühe populatsiooni siseselt oli mudelite prognoosivõime kõrgem kui keskmine täpsus teistes populatsioonides.

Populatsioonides, kus kasutati mikrobioomi tunnuse arvutamisel ühesuguseid baktereid, sobitusid ennustusmudelid üksteise populatsiooni andmetele paremini võrreldes populatsioonidega, kus ühiseid bakteriliike ei leidunud. Seega võib arvata, et lisaks mikrobioomi kogumisest tulenevatele erisustele, võivad tulemusi mõjutada ka populatsioonidevahelised erinevused elustiilis või toitumises. Täiendavalt võrreldi *Selbal* metoodikaga saadud tulemusi masinõppemudeliga Wirbel jt tööst. Selgus, et masinõppemeetodil leitud mudelid on parema ennustusvõimega, kuid tulemustes oli näha ühesuguseid trende.

Üldiselt võib *Selbal* metoodikat hinnata sobivaks soolevähi ennustusmudelite loomisel, sest see suudab üsna väheste bakterite abil hinnata inimese haigusriski. Lisaks on loodud mudelid väga lihtsasti interpreteeritavad erinevalt mitmetest masinõppel põhinevatest meetoditest. Täpsema hinnangu andmiseks meetodi headusele, on vajalik täiendavalt uurida mudeli prognoosivõimet rohkemate populatsioonide

andmetel, kaasates mudelitesse rohkem soolevähi riskitegureid ning mikrobioomi mõjutavaid faktoreid. Siiski on selge, et mikrobioomi saab kasutada lisaks klassikalistele riskifaktoritele täiendava tegurina soolevähi prognoosimisel.

# Kasutatud allikad


[1] J. R. Marchesi ja J. Ravel, "The vocabulary of microbiome research: a proposal," *Microbiome*, köide 3, nr 31, lk. 1–3, 2015. DOI: 10.1186/s40168-015-0094-5.

[2] S. Cooper, R. Mathews, L. Bushar, B. Paddock, J. Wood ja R. Tammara, "The Human Microbiome: Composition and Change Reflecting Health and Disease," *HAPS Educator*, köide 23, nr 4, lk. 432–435, 2019. DOI: 10.21692/haps.2019.020.

[3] Y. Fan ja O. Pedersen, "Gut microbiota in human metabolic health and disease," *Nature Reviews Microbiology*, köide 19, lk. 55–71, 2021. DOI: 10.1038/s41579-020-0433-9.

[4] J. A. Gilbert, M. J. Blaser, G. J. Caporaso, J. K. Jansson, S. V. Lynch ja R. Knight, "Current understanding of the human microbiome," *Nature Medicine*, köide 24, nr 4, lk. 392–398, 2018. DOI: 10.1038/nm.4517.

[5] Eesti Vähiliit, *Vähi teke ja areng*, 2020. aadress: http://cancer.ee/info-vahist (vaadatud 29.03.2021).

[6] Eesti Haigekassa, *Soolevähi sõeluuring*. aadress: http://www.haigekassa.ee/inimesele/haiguste-ennetus/jamesoolevahi-soeluuring (vaadatud 29.03.2021).

[7] Tartu Ülikooli kliinikum, *Käär- ja pärasoolevähk (jämesoolevähk)*. aadress: https://www.kliinikum.ee/ho/info-haiguste-kohta/2-uncategorised/89-kaeaer-ja-paerasoole-vaehk-jaemesoole-vaehk (vaadatud 29.03.2021).

[8] G. Mori ja M. R. Pasca, "Gut Microbial Signatures in Sporadic and Hereditary Colorectal Cancer," *International Journal of Molecular Sciences*, köide 3, nr 22, 2021. DOI: 10.3390/ijms22031312.

[9] European Cancer Information System (ECIS), *Colorectal cancer burden in EU-27*. aadress: https://ecis.jrc.ec.europa.eu/pdf/Colorectal_cancer_factsheet-Mar_2021.pdf (vaadatud 26.02.2021).

[10] Eesti Tervise Arengu Instituut, *Jämesoolevähi elulemus Eestis paraneb, ent kaugmetastaasidega juhtude osakaal endiselt suur*. aadress: https://www.terviseinfo.ee/et/uudised/4984-jamesoolevahi-elulemus-eestis-paraneb-ent-kaugmetastaasidega-juhtude-osakaal-endiselt-suur (vaadatud 30.03.2021).

[11] Eesti Tervise Arengu Instituut, *VSR27: Jämesoolevähi sõeluuringul avastatud vähijuhud*. aadress: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__07Soeluuringud/VSR27.px/table/tableViewLayout2/ (vaadatud 11.05.2021).

[12] Eesti Tervise Arengu Instituut, *VSR25: Jämesoolevähi sõeluuringule kutsutavate meeste ja naiste hõlmatus kutse ja uuringuga ravikindlustuse olemasolu, maakonna ja soo järgi*. aadress: https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__07Soeluuringud/VSR25.px/table/tableViewLayout2/ (vaadatud 11.05.2021).

[13] K. L. Krigul, O. Aasmets, K. Lüll, T. Org ja E. Org, "Using fecal immunochemical tubes for the analysis of the gut microbiome has the potential to improve colorectal cancer screening," 2021. DOI: 10.1101/2021.03.15.435399.

[14] J. Rivera-Pinto, J. J. Eqozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian ja M. L. Calle, "Balances: a New Perspective for Mic-

robiole Analysis," *MSystems*, köide 3, nr 4, lk. 1–12, 2018. DOI: 10.1128/mSystems.00053-18.

[15] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn ja J. J. Egozcue, "Microbiome Datasets Are Compositional: And This Is Not Optional," *Frontiers in Microbiology*, köide 8, nr 2224, lk. 1–5, 2017. DOI: 10.3389/fmicb.2017.02224.

[16] V. Pawlowsky-Glahn, J. J. Egozcue ja R. Tolosana-Delgado, *Lecture Notes on Compositional Data Analysis*, 2007, lk. 5–10. aadress: http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDa.pdf (vaadatud 11.04.2021).

[17] J. Rivera Pinto, *Statistical methods for the analysis of microbiome compositional data in HIV studies. Compositional data and microbiome analysis. Universitat Central De Catalunya. Doktoritöö.* 2018. (vaadatud 12.05.2021).

[18] G. James, T. Hastie, R. Tibshirani ja D. Witten, *An Introduction to Statistical Learning*. Springer, New York, NY, 2013, ISBN: 978-1-4614-7137-0.

[19] T. Kaart, *Binaarsete tunnuste analüüsimeetodid, Binaarse tunnuse seos pideva arvtunnusega*, 2012. aadress: http://www.eau.ee/~ktanel/bin_tunnuste_analyys/bin_tunnuste_analyys.pdf (vaadatud 29.01.2021).

[20] J. Egozcue ja V. Pawlowsky-Glahn, "Groups of Parts and Their Balances in Compositional Data Analysis," *Mathematical Geosciences*, köide 37, lk. 795–828, 2005. DOI: 10.1007/s11004-005-7381-9.

[21] J. Aitchison, "The Statistical Analysis of Compositional Data," *Journal of the Royal Statistical Society Series B*, köide 44, nr 2, lk. 139–177, 1982. aadress: https://www.jstor.org/stable/2345821.

[22] J. Wirbel, P. T. Pyl, K. E., K. Zych, A. Kashani, A. Milanese, J. S. Fleck, A. Y. Voigt, A. Palleja, R. Ponnudurai, S. Sunagawa, L. P. Coelho, P. Schrotz-King, E. Vogtmann, N. Habermann, E. Niméus, E. Thomas, P. Manghi, S. Gandini, D. Serrano, S. Mizutani, H. Shiroma, S. Shiba, T. Shibata, S. Yachida, T. Yamada, L. Waldron, A. Naccarati, N. Segata, R. Sinha, C. M. Ulrich, H. Brenner, M. Arumugam, P. Bork ja G. Zeller, "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer," *Nature Medicine*, köide 25, lk. 679–689, 2019. DOI: 10.1038/s41591-019-0406-6.

# Lisa 1. Kirjeldav statistika

Tabel 12: Uuringus osalenute jaotus andmestikus

| Grupp/Sugu | Naine | Mees | $\sum$ |
|---|---|---|---|
| CRC | 107 | 178 | 285 (49, 6%) |
| CTR | 123 | 167 | 290 (50, 4%) |
| $\sum$ | 230 (40%) | 345 (60%) | 575 (100%) |

Tabel 13: Kirjeldav statistika uuringus osalenute võrdluses

| Tunnus | Min. | Keskmine (std) | Mediaan | Max. |
|---|---|---|---|---|
| VANUS | 25 | 63 (11, 01) | 64 | 90 |
| KMI | 13, 3 | 25, 3 (3, 91) | 24, 7 | 40 |

Tabel 14: Kirjeldav statistika soolevähi diagnoosiga (CRC) patsientide grupis

| Tunnus | Min. | Keskmine (std) | Mediaan | Max. |
|---|---|---|---|---|
| VANUS | 31 | 65 (11, 82) | 65 | 90 |
| KMI | 13, 3 | 25, 3 (4, 10) | 25 | 40 |

Tabel 15: Kirjeldav statistika kontrollgrupi (CTR) kohta

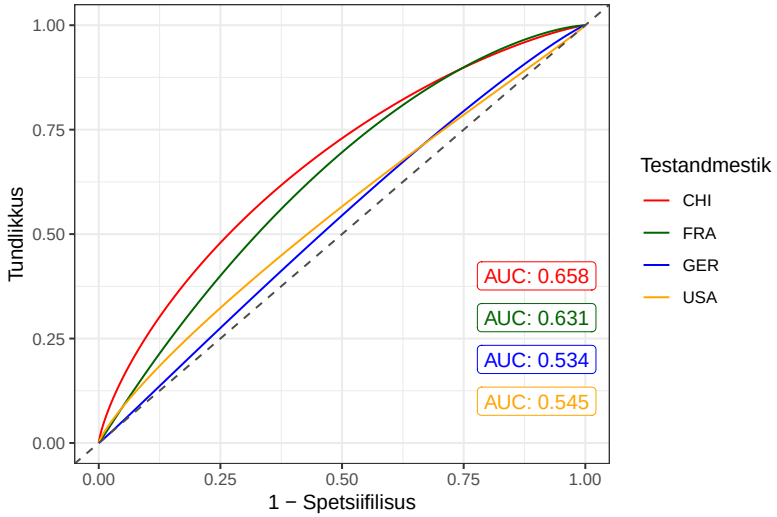
| Tunnus | Min. | Keskmine (std) | Mediaan | Max. |
|---|---|---|---|---|
| VANUS | 25 | 62 (9, 90) | 64 | 84 |
| KMI | 16, 8 | 24, 5 (3, 72) | 25, 3 | 38, 4 |

Tabel 16: Uuringus osalenute sooline jaotus soolevähi diagnoosi põhjal

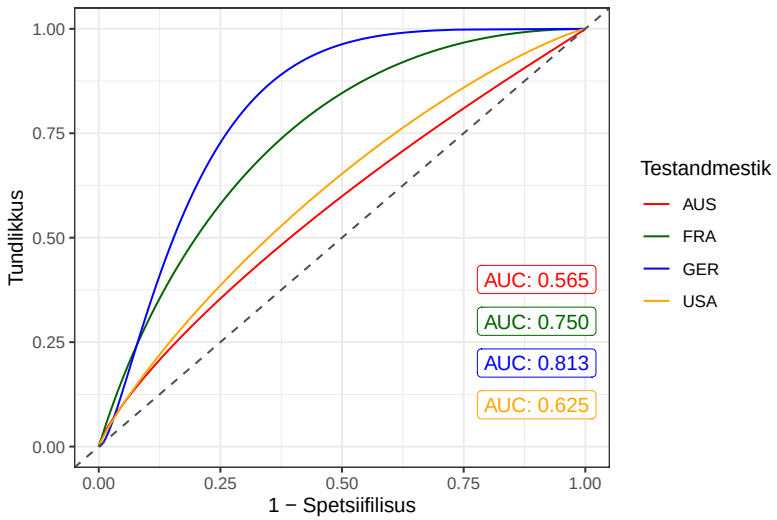
| Grupp/Sugu | Naine (%) | Mees (%) | $\sum$ |
|---|---|---|---|
| CRC | 107 (37, 5) | 178 (62, 5) | 285 (100%) |
| CTR | 123 (42, 4) | 167 (57, 6) | 290 (100%) |

# Lisa 2. ROC-kõverad

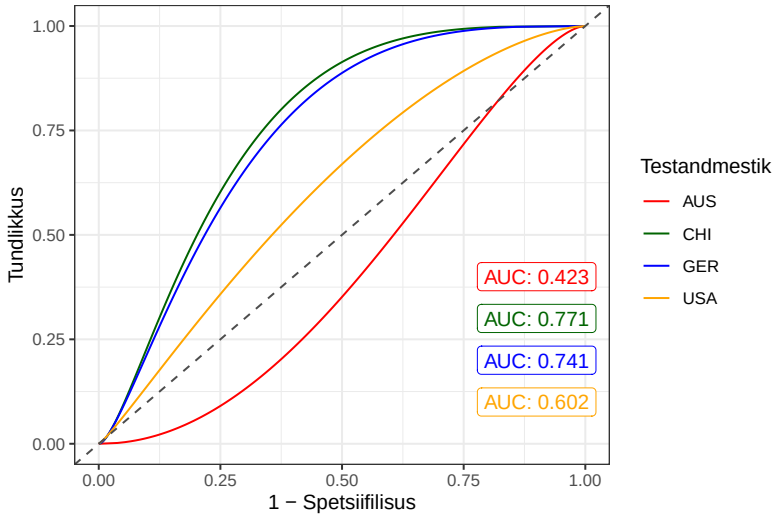
Järgnevalt on graafiliselt kujutatud *Selbal* metoodika abil loodud mudelite üldistusvõime hindamiseks ROC-kõverad. ROC-kõverate teoreetiline kirjeldus on alapeatükis 4.2. Testandmestikud on tähistatud järgnevalt: AUS - Austria, FRA - Prantsusmaa, GER - Saksamaa, CHI - Hiina.



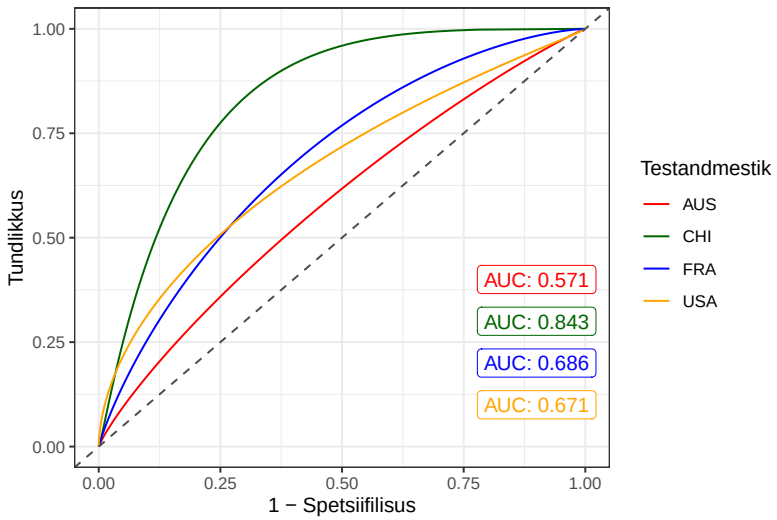

Joonis 2: Austria andmetel loodud mudeli ROC-kõverad teiste populatsioonide andmetel



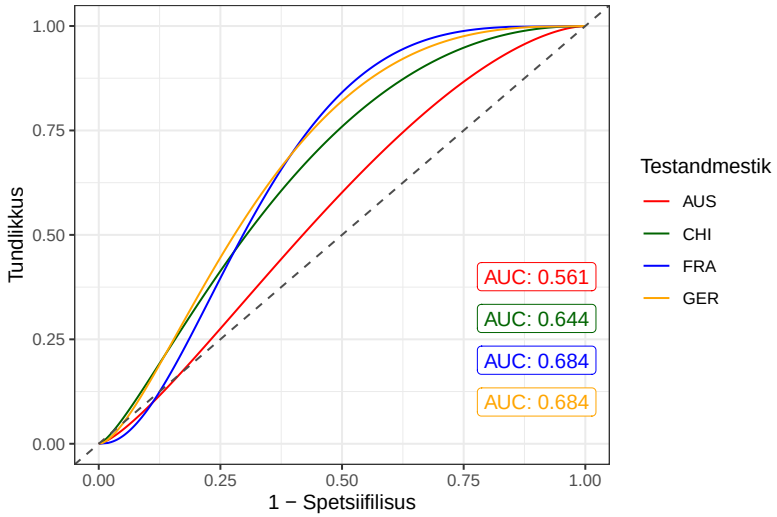

Joonis 3: Hiina andmetel loodud mudeli ROC-kõverad teiste populatsioonide andmetel




Joonis 4: Prantsusmaa andmetel loodud mudeli ROC-kõverad teiste populatsioonide andmetel




Joonis 5: Saksamaa andmetel loodud mudeli ROC-kõverad teiste populatsioonide andmetel




Joonis 6: USA andmetel loodud mudeli ROC-kõverad teiste populatsioonide andmetel

# Lisa 3. Mikrobioomi tunnuse seos soolevähiga

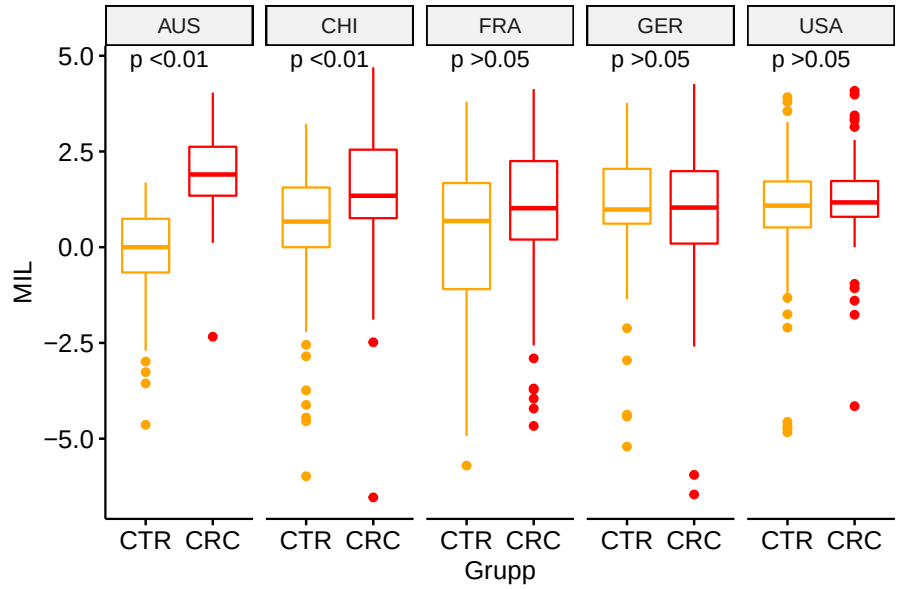
Järgnevalt on graafiliselt kujutatud iga populatsiooni andmetel *Selbal* metoodika põhjal leitud mikrobioomi tunnuse MIL jaotus teistes populatsioonides soolevähi patsientidel ning kontrollidel. Vertikaalteljel on kujutatud MIL väärtus, horisontaalteljel inimese soolevähi diagnoos. Diagnoos CRC tähistab soolevähi esinemist ning CTR tähistab negatiivset diagnoosi. Joonis on jaotatud alamjoonisteks populatsioonide kaupa. Populatsioonide lühendite vastavus on järgnev: AUS-Austria, CHI-Hiina, FRA-Prantsusmaa, GER-Saksamaa.

MIL olulisust erinevates gruppides testitakse t-testi abil, mille hüpoteesid on järgnevad:

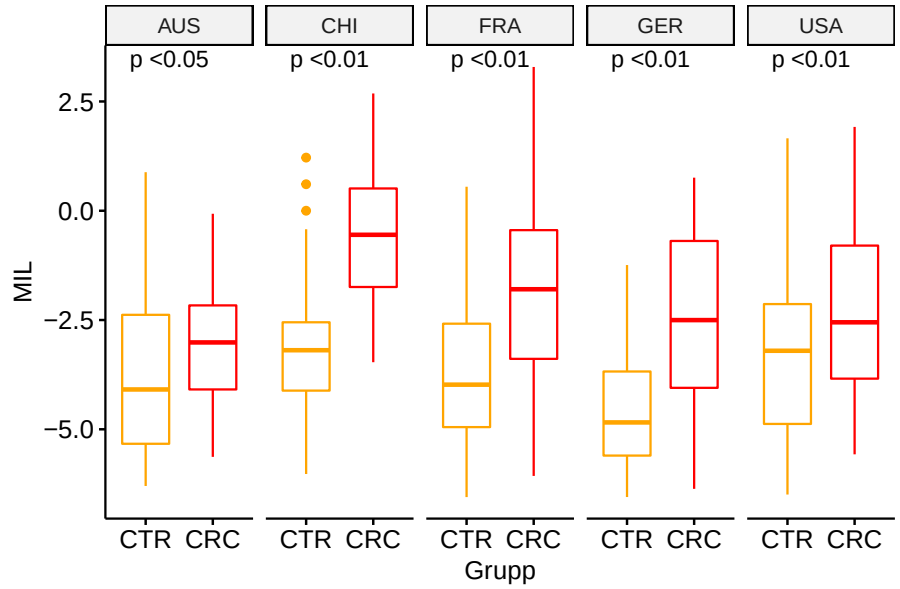
$H_0$ : Keskmine MIL väärtus soolevähiga patsientidel ja kontrollgrupis on ühesugune.

$H_1$ : Keskmine MIL väärtus soolevähiga patsientidel ja kontrollgrupis on erinev.

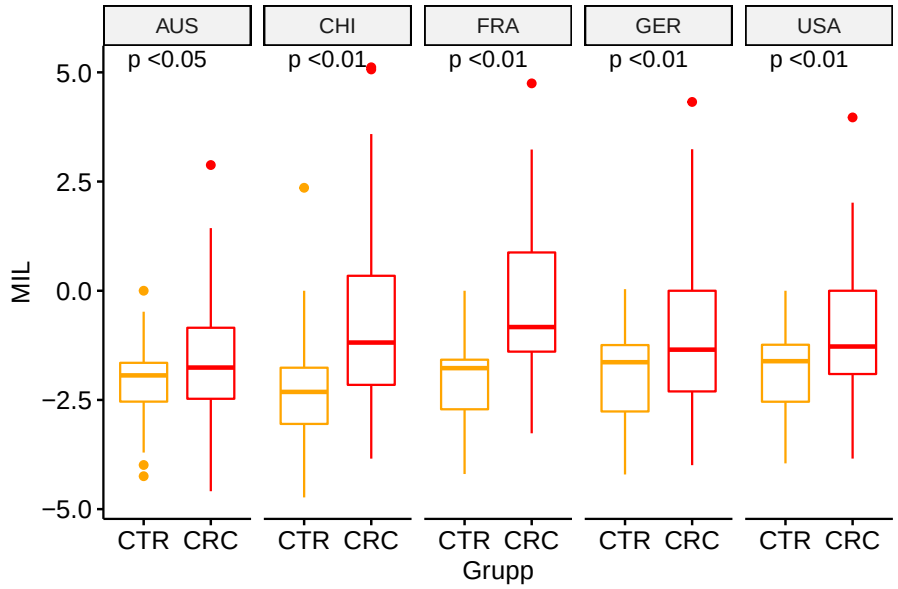
MIL väärtuse olulisuse hindamiseks kasutatakse olulisusnivood $\alpha = 0,05$, joonistel on kujutatud olulisustõenäosus sümboliga $p$. Kui olulisustõenäosus $p \leq \alpha$, siis võetakse vastu hüpotees $H_1$.



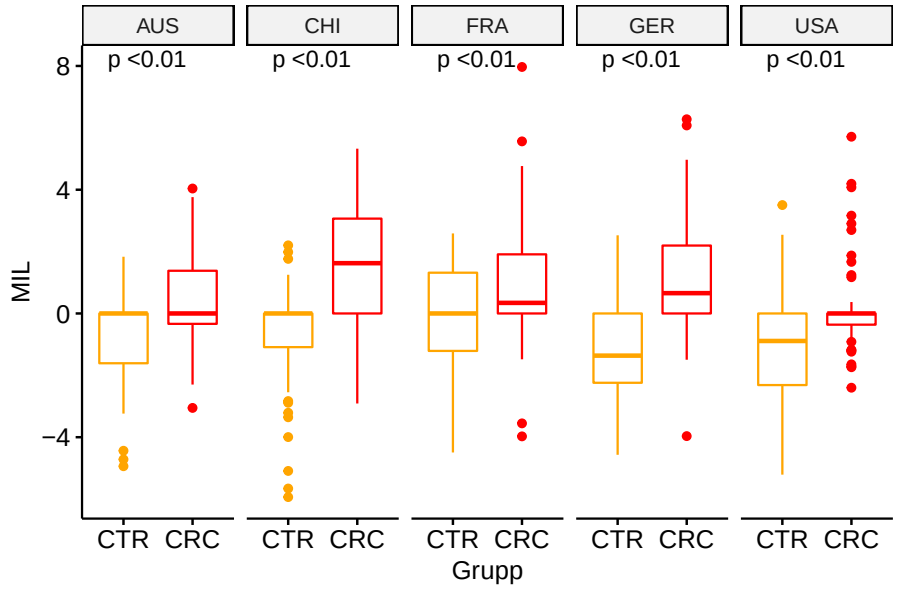

Joonis 7: Austria andmetel leitud MIL tunnuse jaotumine diagnooside võrdluses



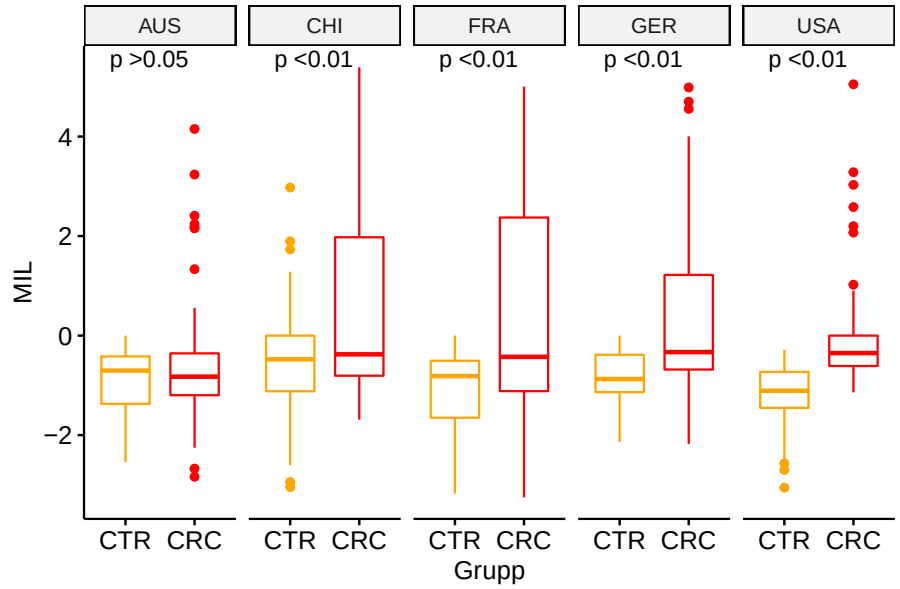

Joonis 8: Hiina andmetel leitud MIL tunnuse jaotumine diagnooside võrdluses




Joonis 9: Prantsusmaa andmetel leitud MIL tunnuse jaotumine diagnooside võrdluses




Joonis 10: Saksamaa andmetel leitud MIL tunnuse jaotumine diagnooside võrdluses




Joonis 11: USA andmetel leitud MIL tunnuse jaotumine diagnooside võrdluses

**Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, Triin Bulõgina,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Mikrobioomil põhinevad ennustusmudelid soolevähi diagnoosimiseks“, mille juhendaja on Oliver Aasmets, 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.

Triin Bulõgina
18.05.2021