Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Karmel Teder

Teise tüübi diabeedi riski prognoosimine ja tagasiside algoritmi väljatöötamine TÜ Eesti geenivaramu andmetel

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: prof. Krista Fischer, PhD

Natalia Pervjakova, PhD

Tartu 2023

TEISE TÜÜBI DIABEEDI RISKI PROGNOOSIMINE JA TAGASISIDE ALGORITMI VÄLJATÖÖTAMINE TÜ EESTI GEENIVARAMU ANDMETEL

Bakalaureusetöö

Karmel Teder

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on koostada mudel, mille abil on võimalik anda Tartu Ülikooli Eesti geenivaramu geenidoonoritele tagasisidet nende teist tüüpi diabeedi saamise 10 aasta riski kohta. Geenidoonorite andmete põhjal valitakse välja parim teise tüübi diabeedi polügeenne riskiskoor ning hinnatakse viis Weibulli parameetrilist mudelit. Selgub, et geenidoonorite teist tüüpi diabeeti haigestumise geneetilist riski kirjeldab kõige paremini riskiskoor PGS002771 ning teist tüüpi diabeeti haigestumist mõjutavad tugevalt inimese vanus, kehamassiindeks, võõümbermõõt ning geneetilise riskiskoori väärtus.

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

Märksõnad: elukestusanalüüs, geenivaramu, polügeenne riskiskoor, teist tüüpi diabeet.

PREDICTION OF THE RISK OF TYPE TWO DIABETES AND DEVELOPMENT OF FEEDBACK FOR GENE DONORS BASED ON DATA FROM UT ESTONIAN BIOBANK

Bachelor's thesis

Karmel Teder

Abstract

The aim of this bachelor's thesis is to create a model that would enable giving the gene donors of UT Estonian Biobank feedback about their 10-year risk of getting type two diabetes. Based on the data from gene donors, the best type two diabetes polygenic risk score is chosen and five Weibull parametric models are fitted. As a result it is seen that the genetic risk of getting type two diabetes is best described by the polygenic risk score PGS002771, and getting type two diabetes is affected by person's age, body mass index, waist circumference, and polygenic risk score value.

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

Key Words: Survival analysis, biobank, polygenic risk score, type two diabetes.

Sisukord

Sissejuhatus

Käesoleva bakalaureusetöö eesmärk on koostada mudel, mille abil on võimalik anda Tartu Ülikooli Eesti geenivaramu geenidoonoritele tagasisidet nende teist tüüpi diabeedi saamise 10 aasta riski kohta. Mudel hinnatakse kasutades geenivaramust saadud andmeid. Teist tüüpi diabeedi riski kirjeldamiseks kaasatakse mudelisse haiguse geneetilise riskiskoori ning teiste oluliste riskifaktorite mõjud.

Teise tüübi diabeet (T2D) on krooniline energia ainevahetuse häire, mida põhjustavad insuliiniresistentsus ning insuliini tootmise häire. See on on erinevatest diabeedi avaldumisvormidest kõige levinum. Eestis on teist tüüpi diabeet diagnoositud umbkaudu 85 protsendil 70 000 diabeedi diagnoosi saanud inimesest (*Mis on diabeet?* 2023). Arvatakse, et tegelik haigete arv on diagnoosi saanud inimeste arvuga võrreldes kahekordne(Association *et al.*, 2008). Diabeeti haigestumise oluline riskitegur on haiguse pärilikkus, kuid haigestumist soodustavad ka ülekaal, eriti vöökohale kogunenud rasvaga, ning vähene liikumine (*Mis on diabeet?* 2023).

Maailmas põdes aastal 2021 diabeeti hinnanguliselt umbes 536,6 miljonit täiskasvanut, mis on ligikaudu 11 protsenti maailma täiskasvanud elanikkonnast. Aastaks 2045 hinnatakse selle arvu kasvu 783,3 miljonini (12%). Diabeedi põhjustatud surmade arvuks täiskasvanute seas aastal 2021 on hinnatud 6,7 miljonit (*IDF Diabetes Atlas* 2021). Diabeediga kaasneb suurenenud risk haigestuda südame ja veresoonkonna haigustesse. Samuti võib muutuda diabeedihaige närvisüsteem või kahjustuda silma võrkkest või neerud. Seetõttu on oluline järgida tervislikku eluviisi ning vähendada nii riski haigestuda diabeeti (*Mis on diabeet?* 2023). Tervislik eluviis on olulisel kohal ka diabeedi ravimisel. Diabeeti ei saa täielikult välja ravida, kuid söömise kontrolli all hoidmisel, kehakaalu vähendamisel ning liikumisosakaalu suurendamisel võivad isegi diabeedi diagnoosiga patsiendid hakkama saada ilma ravimiteta või lükata märgatavalt edasi insuliini ravi alustamist. (Roosimaa *et al.*, 2021)

Teist tüüpi diabeedi päritavuseks on hinnatud 20–80% (Almgren *et al.*, 2011; Ali, 2013). Seejuures on inimestel, kelle õel või vennal on teist tüüpi diabeet, võrreldes nendega, kelle õel ega vennal seda ei ole, ligikaudu kahekordne risk haigusesse haigestuda (Hemminki *et al.*, 2010; Meigs, Cupples ja Wilson, 2000). Inimesel, kelle vanemal või lapsel on diagnoositud teist tüüpi diabeet, on aga kolmekordne risk teise tüübi diabeeti haigestuda, võrreldes nende inimestega, kelle vanemal ega lapsel diabeeti ei ole (Lyssenko *et al.*, 2005).

Vaatamata sellele, et teist tüüpi diabeedi avaldumisel mängib oma rolli geneetika, on võimalik seda haigust ennetada. Seepärast on oluline teada kõrgest teise tüübi diabeeti haigestumise riskist enne haiguse avaldumist. Riski teadvustamise eesmärgil on välja töötatud erinevaid riskihindamise algoritme – näiteks Soome teadlaste poolt pakutud FINDRISK algoritm (Lindström *et al.*, 2010). Need algoritmid ei kasuta aga geeniandmeid. Seega on oluline küsimus, kui palju täpsemalt suudetakse riski prognoosida geneetilise komponendi lisamisel. Ülegenoomsete seoseuuringutega (GWAS) on leitud sadu teist tüüpi diabeediga geenivariante ehk ühenukleotiidseid polümorfisme (SNP) (Flannick ja Florez, 2016; Mahajan *et al.*, 2022), millest tulenevalt on teist tüüpi diabeedile arvutatud sadu polügeenseid riskiskoore (PRS) (*PGS Catalog* 2023).

PRS on geneetilist riski väljendav arvuline näitaja. See arvutatakse summana, mille iga liidetav vastab ühele SNPle, mis haigust mõjutab. Riskiskoori iga liidetav on vastava SNP kaalutud mõju $\hat{\beta}_j$ ja seda SNPd mõjutavate samas kromosoomis paiknevate geenivariatsioonide ehk efektialleelide arvu X_j korrutis. Seejuures on SNP mõju kaal seda väiksem, mida väiksema täpsusega on mõju hinnatud. Seega on k mõju avaldava SNPiga polügenne riskiskoor arvutatav järgnevalt:

$$PRS = \sum_{j=1}^{k} \hat{\beta}_j X_j.$$

(Esko *et al.*, 2019). Teist tüüpi diabeedi korral on optimaalsel geneetilisel riskis-

kooril väga tugev efekt haiguse prognoosimisele ja seega on põhjendatud selle kasutamine haiguseriski prognoosimise algoritmides (Läll *et al.*, 2017).

Töö on kirjutatud kahes peatükis. Neist esimeses on kirjeldatud kasutatud metoodikat ning see jaguneb kaheks alapeatükiks, millest teine omakorda viieks jaguneb. Bakalaureusetöö teises peatükis on kolmes alapeatükis kirjeldatud töös kasutatud andmeid, nende töötlemist ning nende põhjal hinnatud mudeleid.

Autor tänab bakalaureusetöö juhendajaid abistavate nõuannete ning meeldiva koostöö eest.

1 Statistilised meetodid

Käesolev peatükk on jagatud kaheks alapeatükiks. Esimeses alapeatükis on kirjeldatud Soomes välja töötatud teist tüüpi diabeedi riski prognoosimise algoritmi ning teises alapeatükis elukestusanalüüsi meetodeid. Teine ehk elukestusanalüüsi meetodeid kirjeldav alapeatükk on omakorda jagatud viieks. Alapeatüki esimeses osas on kirjeldatud põhilisi elulemusanalüüsis kasutatavaid funktsioone ning seejärel Kaplan-Meieri hinnangut ühele põhifunktsioonidest. Alapeatüki kolmandas ning neljandas alapeatükis on kirjeldatud vastavalt Coxi võrdeliste riskide mudelit ning Weibulli parameetrilist mudelit. Elukestusanalüüsi alapeatüki viimases osas on kirjeldatud, kuidas hinnata neid mudeleid rakendustarkvaras R.

1.1 Riski prognoosimise algoritm FINDRISC

Käesolev peatükk on kirjutatud Lindström ja Tuomilehto (2003) ning Lindström *et al.* (2010) põhjal.

FINDRISC on Soomes välja töötatud teise tüübi diabeedi (T2D) riski prognoosimise algoritm. Riski prognoosimise algoritmi eesmärk on lisaks T2D riski hindamisele tõsta ka inimeste teadlikkust näitajatest, mida nad saavad muuta, et enda teist tüüpi diabeedi riski vähendada. Nendeks näitajateks on kehamassiindeks (KMI) ja vööümbermõõt. KMI ei pruugi alati peegeldada, kas inimene on tervislikus kehakaalus, vööümbermõõt on teist tüüpi diabeedi riskitegur isegi siis, kui KMI mahub normi vahemikku. Nende näitajate hulka kuuluvad ka kehaline aktiivsus ja puu- ning juurviljade söömine.

Lisaks mainitud näitajatele võetakse FINDRISC riskiskoori arvutamisel arvesse neid tegureid, mida inimene muuta ei saa, aga mis diabeedi riski mõjutavad. Sinna kuuluvad inimesel mõõdetud kõrge veresuhkru tase, kas ta on kunagi tarvitanud regulaarselt ravimeid kõrge vererõhu alandamiseks ning kas tema lähisugulastel on diagnoositud esimest või teist tüüpi diabeet.

FINDRISC riskiskoor on tehtud inimestele arusaadavaks, teisendades teist tüüpi diabeedile hinnatud logistilises regressioonimudelis hinnatud argumenttunnuste mõjud skaalasse 0–5 punkti. Enamasti vastab 0 punkti riskiteguri kõige madalamale tasemele ja 5 punkti kõige kõrgemale. Seejuures on eranditeks kehalise aktiivsuse ning puu- ja juurviljade söömise sagedus, kus 0 punkti saavad kõige kõrgemad kategooriad ehk igapäevane puu- ja juurviljade söömine ning enam kui neli tundi liigutamist nädala jooksul. Mudeli kordajad saavad nullilähedase väärtuse puhul ühe, ning arvust 2,2 suurema väärtuse korral viis punkti. Diabeedi riskiskoor arvutatakse indiviidi näitajate põhjal saadud punktide summana, kusjuures ravi vajava diabeedi riskiks loetakse summat, mis on suurem kui 9 või sellega võrdne.

Teist tüüpi diabeedi riski hindamiseks tuleb FINDRISC riskiskoori puhul täita lihtne, kaheksast küsimusest koosnev küsimustik, mida saab täita ka veebis (Lindström ja Tuomilehto, 2023). Selle küsimustiku täitmisel tuleb inimesel enamasti valida kahe kuni kolme vastusevariandi vahel. Vanus on aga riskiskoori hindamiseks jagatud nelja kategooriasse – vähem kui 45 aastat, 45–54 aastat, 55–64 aastat ning enam kui 64 aastat. Need kategooriad lisavad riskiskoorile vastavalt 0, 2, 3 ja 4 punkti. Kõige rohkem võib inimene punkte saada vastates küsimusele, kas mõni tema pereliige on diagnoositud esimest või teist tüüpi diabeediga. Inimesed, kellel diabeedi diagnoosi on saanud nii otsene pereliige, kui ka näiteks tädi, vanaema või nõbu, saavad selle näitaja puhul 8 punkti. Nendest punktides 5 tulevad vastusest, et diagnoosi on saanud otsene pereliige. Seejuures ongi suurim võimalik punktisumma, mida küsimusele vastates võib saada, 8.

Kehamassiindeks arvutatakse kui kilogrammides mõõdetud kaalu ning meetrites mõõdetud pikkuse ruudu jagatis. Ühe punkti saavad siin need inimesed, kelle kehamassiindeks jääb 25 ja 30 vahele ning 3 punkti need, kelle kehamassiindeks 30 ületab. Roietest allpool mõõdetud vööümbermõõtu hinnatakse meestel ja naistel erinevalt. Naised saavad kolm punkti, kui nende vööümbermõõt jääb 80 ja 88 sentimeetri vahele ning mehed, kui näitaja jääb 94 ning 102 sentimeetri vahele. Seejuures

on vahemiku otspunktid kaasa arvatud. Inimesed, kelle vööümbermõõt märgitud ülemist piiri ületab, saavad 4 punkti ja need, kelle vööümbermõõt on väiksem kui alumine piir, 0 punkti.

Kehaliselt väheaktiivsete inimeste riskiskoorile lisandub 2 punkti. Samas neile, kes igapäevaselt puu- või juurvilju ei söö, vaid üks punkt. Inimestel, kellel on kunagi mõõdetud kõrge veresuhkru tase, lisandub teist tüüpi diabeedi riskiskoorile 5 punkti. Kaks punkti lisandub nende inimeste riskiskoorile, kes on kunagi regulaarselt tarvitanud kõrge vererõhu ravimeid.

Punktide summaga hinnatakse inimese riski haigestuda teist tüüpi diabeeti 10 küsimustikule vastamisele järgneva aasta jooksul. Neile, kes saavad vähem kui 7 punkti, hinnatakse diabeediriskiks umbes üks protsent. Küsimustikust enam kui 20 punkti saanud inimeste risk hinnatakse ligikaudu 50 protsendile. Arusaadavuse mõttes väljendatakse neid riske öeldes vastavalt, et üks inimene sajast või üks inimene kahest jääb kümne aasta jooksul haigeks. Näiteks 12—14 punkti saanud inimeste riski haigestuda tesise tüübi diabeeti hinnatakse keskmiseks, kusjuures haigestub hinnanguliselt üks inimene kuuest. 7–11 punkti saanud inimeste risk on ligikaudu 25 ning 15–20 punkti saanud inimeste risk ligikaudu 33 protsenti.

1.2 Elukestusanalüüsi meetodid

Peatükk on kirjutatud Cox ja Oakes (2018) ning Collett (2015) põhjal, kui ei ole märgitud teisiti.

Elukestusanalüüsi kasutatakse peamiselt selleks, et uurida ajavahemiku pikkust kindlast alghetkest huvipakkuva sündmuse toimumiseni. Huvi võib pakkuda näiteks kestuste jaotus või nende võrdlus erinevates gruppides. Seejuures võib huvipakkuvaks sündmuseks olla surm, haigestumine, asja purunemine, tööle saamine või muu selgelt defineeritud sündmus, mis üldjuhul saab katseobjektil toimuda uuringu jooksul kuni ühe korra.

Elukestusanalüüsi on vaja, kuna kestusandmed on tavaliselt ebasümmeetrilise jaotusega, kuid paljude statistiliste testide kasutamise eelduseks on andmete normaaljaotus. Lisaks ei saa tihti elukestusanalüüsis uuritatavatele andmetele tavapäraseid statistilisi teste rakendada, kuna alati ei realiseeru olukord, kus kõigil katsealustel on katseperioodi lõpuks huvi pakkuv sündmus toimunud. Sel juhul on osad andmed tsenseeritud.

Tsenseerituks nimetatakse nende katseobjektide kestusandmeid, kellel huvipakkuvat sündmust enne katse lõppu toimunud ei ole. Tsenseeritud võivad andmed olla ka seetõttu, et katseobjekti ei saa enam katses vaadelda, kuigi huvipakkuvat sündmust pole veel toimunud ning katse kestab veel. Seda põhjustab näiteks inimese teise riiki kolimine. Tsenseeritud andmete puhul on katseobjekti ajavahemiku lõpuks viimane aeg, mil temast midagi teatakse. Kestusandmetes eristatakse tsenseeritud andmed tsenseerimata andmetest.

1.2.1 Põhifunktsioonid

Katsealuse elukestust ehk aega kindlast alghetkest huvipakkuva sündmuse toimumiseni kirjeldab juhuslik suurus, mis ei ole negatiivne. Olgu selle suuruse tähistus T ning katsealuse vaadeldud elulemuse ehk elukestuse tähistus t. Olgu elukestuse tihedusfunktsioon $f(t)$. Siis elulemuse jaotusfunktsioon on

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{1}$$

mida võib nimetada ka kumulatiivseks esinemusfunktsiooniks.

Elulemusfunktiooniga $S(t)$ mõõdetakse tõenäosust, et huvipakkuv sündmus ei toimu enne kindlat ajahetke t. Võrdusest (1) järeldub seega

$$S(t) = P(T \geq t) = 1 - F(t).$$

See on üks põhilistest funktsioonidest, mida elulemusandmete kirjeldamisel kasu-

tatakse. Lisaks sellele on elukestusanalüüsis olulisel kohal riskifunktisoon.

Riskifunktsiooniga $h(t)$ väljendatakse riski, et sündmus toimub kindlal ajahetkel t. See on arvutatud kui piirväärtus tinglikust tõenäosusest ajaühiku kohta. Seejuures tinglik tõenäosus on tõenäosus et elukestust kirjeldav juhuslik suurus T satub ajavahemikku pärast kindlat hetke t eeldusel, et elukestus on pikem kui aeg alghetkest hetkeni t. Niisiis

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

Saab näidata, et riskifunktsioon on võrdne elukestusandmete tihedusfunktsiooni ning elulemusfunktsiooni jagatisega.

Risk, et huvipakkuv sündmus toimub kümne aasta jooksul alghetkest t_0 on arvutatav kui

$$\begin{aligned} P(T \leq t_0 + 10 \mid T > t_0) &= 1 - P(T > t_0 + 10 \mid T > t_0) = \\ &= 1 - \frac{P(\{T > t_0 + 10\} \cap \{T > t_0\})}{P(T > t_0)} = \\ &= \frac{P(T > t_0 + 10)}{P(T > t_0)} = \\ &= 1 - \frac{S(t_0 + 10)}{S(t_0)}, \end{aligned}$$

kus viimane võrdus kehtib, kuna aeg on pidev tunnus. Juhul, kui elukestusena vaadeldakse möödunud aega alghetkest, mitte katsealuse vanust, on $t_0 = 0$ ning $S(t_0) = S(0) = 1$. Kümne aasta risk avaldub seega kui 1-S(10).

1.2.2 Kaplan-Meieri hinnang elulemusfunktsioonile

Elukestusanalüüsis kasutatakse elulemusfunktsiooni hindamiseks tavaliselt enne muude hinnangute leidmist Kaplan-Meieri hinnangut. Selleks järjestatakse vaadeldud elukestused kasvavas järjekorras. Kaplan-Meieri hinnang elulemusfunktsiooni

väärtusele k-nda ning $k+1$-se ajahetke vahel on

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right),$$

kus n_j on nende inimeste arv, kellel enne j-ndat ajahetke vaadeldav sündmus toimunud ei olnud ning d_j on nende inimeste arv, kellel toimus vaadeldav sündmus j-ndal ajahetkel.

1.2.3 Coxi võrdeliste riskide mudel

Elukestusanalüüsis on tihti vaja hinnata erinevate kirjeldavate tunnuste mõju elulemusele. Need tunnused võivad olla näiteks katsealuse ravi, omadused või välised mõjurid. Kirjeldavate tunnuste väärtused võivad olla konstantsed või sõltuda ajast. Olgu $z = (z_1, z_2, \ldots, z_q)^T$ vektor tunnuste väärtustega, mis kirjeldavad katsealust. Tunnuste analüüsil on laialdaselt kasutusel riskifunktsioon. Konstantsete väärtustega vektori z lihtne võrdeliste riskide mudel on

$$h(t; z) = \psi(z) h_0(t), \tag{2}$$

kus $\psi(z)$ on funktsioon, mille suuremad väärtused väljendavad suuremat riski katsealuse väiksemale elulemusele. Funktsioon $\psi(z)$ peab rahuldama tingimust $\psi(0) = 1$. Funktsioon $h_0(t)$ mudelis (2) on sellise katsealuse riskifunktsioon, kelle kõik kirjeldavate tunnuste väärtused on võrdsed nulliga.

Võrdeliste riskide mudelis (2) kasutatud funktsiooni $\psi(z)$ saab vajadusel viia ka parameetrilisele kujule. Olgu $\beta = (\beta_1, \beta_2, \ldots, \beta_q)$ vektor kirjeldavate tunnuste kordajatega. Funktsiooni $\psi(z)$ väärtused on alati mittenegatiivsed ning kohal 0 peab funktsiooni väärtus võrduma ühega. Seetõttu on funktsiooni $\psi(z)$ enimlevinud parameetriline kuju

$$\psi(z; \beta) = e^{\beta^T z},$$

millest Coxi võrdeliste riskide mudel on

$$h(t;z) = e^{\beta^T z} h_0(t).$$

Mudeli hindamiseks kasutatakse osalise tõepära funktsioon. Tähistagu $\tau_1 < \tau_2 < \cdots < \tau_d$ d indiviidi järjestatud vaadeldud elukestusi. $R(\tau_j)$ on j-nda aja riskigrupp suurusega r_j, kuhu kuuluvad need inimesed, kelle elukestus on vähemalt τ_j. Olgu z_i siinkohal nende indiviidide vaadeldud argumenttunnuste vektor kellel toimus huvipakkuv sündmus hetkel t_i. Siis osalise tõepära funktsioon Coxi võrdeliste riskide mudelile on

$$L = \prod_{i=1}^{d} \frac{h_0(t_i) exp(\beta^T z_i)}{\sum\limits_{k \in R(\tau_i)} h_0(t_i) exp(\beta^T z_k)} = \prod_{i=1}^{d} \frac{exp(\beta^T z_i)}{\sum\limits_{k \in R(\tau_i)} exp(\beta^T z_k)}.$$

Parameetrite hinnangud saadakse selle maksimeerimisel.

Coxi võrdeliste riskide mudeli puhul on elulemusfunktsioon

$$S(t) = [S_0(t)]^{e^{\beta^T z}}.$$

Seega on Coxi mudeli põhjal riski prognoosimiseks vaja hinnangut ka funktsioonile $S_0(t)$. Järgnevas alapeatükis on kirjeldatud muelit, mille puhul elulemusfunktsiooni hindamiseks lisafunktsioone hinnata ei ole vaja.

1.2.4 Weibulli parameetriline mudel

Weibulli jaotus on üks mitmetest pidevatest parameetrilistest jaotustest, millest elulemusandmed olla võivad. Jaotusel on kaks parameetrit. Need parameetrid on skaalaparameeter λ ning kujuparameeter α. Weibulli jaotusega andmete puhul on elulemusfunktsioon kujul

$$S(t) = \exp\left(-(\lambda t)^{\alpha}\right)$$

ning riskifunktsioon kujul

$$h(t) = \alpha\lambda(\lambda t)^{\alpha-1}.$$

Olgu elulemusandmed Weibulli jaotusest, parameetritega λ ning α. Siis on konstantsete kirjeldavate tunnuste väärtuste puhul Weibulli jaotusest ka elulemusandmed määratud kirjeldavate tunnuste korral vektoris z. Seejuures on jaotuse parameetriteks $\lambda\psi(z)$ ning α. Võrdeliste riskide mudel on siis kujul

$$h(t, z) = \alpha\lambda^{\alpha}\psi(z)t^{\alpha-1},$$

kus $\psi(z) = \exp(\beta^T z)$.

1.2.5 Elulemusandmete mudeldamine rakendustarkvaras R

Alapeatükk on kirjutatud Thernau *et al.* (2023) põhjal.

Rakendustarkvaras R saab elukestusandmeid analüüsida paketiga *survival*. Elukestused tuleb selleks viia funktsiooni *Surv* abil kujule, kus tsenseeritud ajad on eristatavad tsenseerimata aegadest. Funktsiooni põhiargumentideks on vektor elukestustega ning sama pikkusega vektor indikaatoritega, mis näitavad, kas eelmises vektoris vastaval kohal olnud aeg oli tsenseeritud. Seejuures indikaatori väärtus 0 märgib, et aeg oli tsenseeritud ning väärtus 1, et sündmus toimus.

Coxi võrdeliste riskide mudeli hindamiseks on paketis funktsioon *coxph*, mille peamine argument on valem kujul $objekt \sim z_1 + z_2 + \cdots + z_p$. Tunnus *objekt* selles valemis on funktsiooni *Surv* abil loodud funktsioontunnus. Tildest paremale jäävad mudelisse kaasatavate kirjeldavate tunnuste nimetused andmestikus. Lisaks on funktsiooni väljakutses vaja enamasti ära märkida ka kasutatav andmestik.

Weibulli jaotusega andmete parameetrilise võrdeliste riskide mudeli hindamiseks, saab kasutada funktsiooni *survreg*. Funktsiooni argumendid on sarnased funktsiooni *coxph* argumentidega. Weibulli mudeli saamiseks tuleb funktsiooni väljakutses

märkida *dist="weibull"*. Samuti võib funktsiooni väljakutses märkida skaalaparameetri λ väärtuse. Selle tegemata jätmisel või mittepositiivse skaalaparameetri märkimisel hinnatakse skaalaparameetri väärtus andmetest.

2 Teist tüüpi diabeedi riski tagasiside geenidoonoritele

Käesolev peatükk on jagatud kolmeks alapeatükiks. Esimeses alapeatükis on kirjeldatud andmeid, mida töös kasutatud on. Teises alapeatükis on kirjeldatud, kuidas neid andmeid analüüsitud on. Kolmandas alapeatükis on kokku võetud tähtsamad tulemused, mis töö käigus saadud on.

2.1 Ülevaade andmetest

Tartu Ülikooli Eesti geenivaramu loodud riiklikkusse biopanka on andmeid kogutud alates aastast 2002. Praeguseks on biopangas enam kui 200 000 geenidoonori andmed. Seejuures on ligikaudu 75% geenidoonoritest liitunud geenivaramuga aastatel 2018 ning 2019. Lisaks geeniandmetele kogub geenivaramu ka doonorite terviseandmeid (*Tartu Ülikooli Eesti geenivaramu* 2023). Geenidoonorite andmed saab geenivaramu nii küsimustikest, mille doonorid varamuga liitudes täidavad, kui ka riiklikest andmebaasidest ning registritest. Registritest ning andmebaasidest saadud andmeid värskendatakse regulaarselt. (Leitsalu *et al.*, 2015)

Töö käigus on kasutatud 17 andmetabelit. Nendest andmetabelitest peamine sisaldab 211 590 Tartu Ülikooli Eesti geenivaramu geenidoonori andmeid. Tabelis on igal geenidoonoril unikaalne kood ning doonori ühe geeniproovi unikaalne kood. Selles andmestikus on tunnused, mis näitavad kõrgvererõhktõve, teist tüüpi diabeedi ning südame isheemiatõve kohta, kas inimesel on vastav haigus diagnoositud, seda on kahtlustatud või tal pole seda haigust. Töö käigus on analüüsist välja jäetud 14 336 inimest, kellel on teist tüüpi diabeeti kahtlustatud, kuid pole kindlalt teada, kas haigus esineb või mitte. Lisaks on andmetabelis kuupäev, mil geenidoonor geenivaramuga liitus, tema vanus sel kuupäeval, sugu, teist tüüpi diabeedi diagnoosi saamise kuupäev, geenidoonori kehalised näitajad ja mõned elustiili kirjeldavad näitajad. Elustiili kirjeldavate näitajate hulka kuuluvad geenidoonori söömisharju-

mused, suitsetamisstaatus ning nädalane kehaline aktiivsus tundides.

Kaheteistkümnes andmetabelis on 212 955 geeniproovile arvutatud geneetilised riskiskoorid. Andmestikesse on jagatud 80 erinevat riskiskoori. Riskiskooride taustinfo on ühes andmetabelis. Riskiskoore kirjeldavas andmestikus on välja toodud näiteks, mis aastal avaldatud ning kelle artiklist riskiskoor pärineb, kas selle väljatöötamisel on kasutatud Eestist saadud andmeid ning mitme ühenukleotiidse polümorfismi põhjal riskiskoori arvutatakse. Käesoleva töö raames ei ole autor ise geneetilisi riskiskoore arvutanud, kuid riskiskooride arvutamisel ning nendega seonduvate andmetabelite loomisel on kasutatud polügeensete riskiskooride kataloogi (*PGS Catalog* 2023), mis on loodud, et koondada kõik olulisemad riskiskooride hindamise algoritmid ning nende kohta käiv info ühte kohta. PGS kataloogi alusel arvutatud riskiskoorid olid arvutatud Eesti geenivaramu töötajate poolt.

Riskiskooride võrdlemisel on oluline, et analüüsis ei kasutataks andmeid, mida kasutati riskiskooride väljatöötamisel. Andmete kattumise korral väheneb seose tugevus arvutatud riskiskoori ja sellega mõõdetava haiguse olemasolu vahel sõltumatus andmestikus, kuna kattuvate andmetega andmestikus on seose tugevus üle hinnatud. (Choi, Mak ja O'Reilly, 2020)

Tagamaks, et andmeid analüüsides ei kasutata andmeid, mida on kasutatud geneetiliste riskiskooride väljatöötamisel, on töös kasutatud ka kolme andmetabelit geeniproovide unikaalsete koodidega. Andmetabelites on nende geeniproovide koodid, mida on mõne polügeense riskiskooride väljatöötamisel kasutatud. Nende andmetabelite nimed on välja toodud riskiskoore kirjeldava andmestiku selles veerus, kus on kirjeldatud, kas polügeenset riskiskoori välja töötades kasutati Eestist saadud andmeid.

2.2 Töö käik

Teist tüüpi diabeedi saamise 10 aasta riski hindamiseks on esmalt võrreldud riskiskoore, mis geeniproovidele arvutatud on. Selleks on hinnatud logistilise regressiooni mudelid tunnusele, mis näitab, kas inimene sai teist tüüpi diabeedi diagnoosi enne geenivaramuga liitumist või ta pole teist tüüpi diabeeti haigestunud. Hinnatud on 80 mudelit, kus igas mudelis on olnud kirjeldavaks tunnuseks skaleeritud riskiskoor, inimese sugu ning vanus geenivaramuga liitumisel. Riskiskoorid on skaleeritud, et tagada skooride võrreldavus. Samuti riskiskooride võrreldavuse tagamiseks, on mudeleid hinnates jäetud andmestikust välja need geenidoonorid, kelle proovi on kasutatud mõne riskiskoori väljatöötamisel. Lisaks on analüüsist välja jäetud 14 336 geenidoonorit, kellel on olnud teist tüüpi diabeeti kahtlustatud, kuid selle olemasolu kindel pole. Nii on mudelite hindamisel kasutatud 183 012 geenidoonori andmeid.

Kõige parem riskiskoor on valitud selle põhjal, millisel polügeensel riskiskooril oli sellega hinnatud mudelis kõige suurem Z-statistiku väärtuse absoluutväärtus ning seega ka tugevaim seos haiguse diagnoosiga. Tabelis 1 on välja toodud kümme kõige suurema Z-statistiku absoluutväärtusega riskiskoori, mis on järjestatud väärtuse järgi kahanevalt. Iga riskiskoori kohta on tabelis selle kood, mudelis hinnatud kordaja väärtus $\hat{\beta}$, selle standardviga $\mathrm{se}(\hat{\beta})$ ning Z-statistiku $\frac{\hat{\beta}}{\mathrm{se}(\hat{\beta})}$ absoluutväärtus. Sellise valiku põhjal on töös edaspidi kasutatud Mars jt 2022 riskiskoori PGS002771, mille väljatöötamisel ei olnud kasutatud Eestist saadud andmeid. Selle riskiskoori arvutamisel kasutatakse 1 091 608 SNPi.

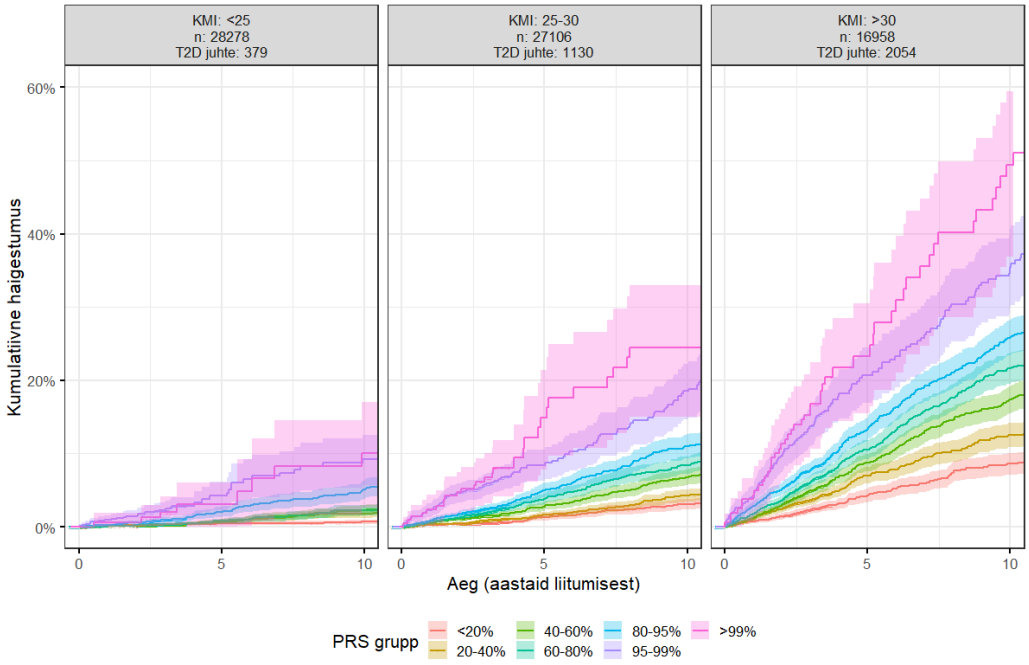
Tabel 1: 10 parimat riskiskoori

Riskiskoori ID	Kordaja mudelis	Standardviga	Z-statistik
PGS002771	0,77	0,01	61,77
PGS002308	0,76	0,01	60,63
PGS003118	0,69	0,01	56,16
PGS003103	0,69	0,01	55,86
PGS000330	0,66	0,01	54,38
PGS003099	0,64	0,01	52,27
PGS003117	0,61	0,01	49,49
PGS003102	0,61	0,01	49,49
PGS000729	0,59	0,01	49,39
PGS002354	0,59	0,01	48,64

Pärast kõige parema riskiskoori välja valimist on selle skaleeritud väärtused lisatud geenidoonorite andmestikku. Seejärel on leitud riskiskoori kvantiilid ning geenidoonorid nende põhjal gruppidesse jagatud. Igale grupile on funktsiooni *survfit* abil hinnatud elulemusfunktsiooni väärtused. Seejuures on mõõdetud aega geenivaramuga liitumisest teist tüüpi diagnoosi saamiseni. Inimesed, kes on teist tüüpi diabeedi diagnoosi saanud enne geenivaramuga liitumist, on selles osas analüüsist välja jäetud. Kokku jäid analüüsi 184 892 inimese andmed, kellest 4700-l oli diagnoositud teise tüübi diabeet. Tsenseerimiskuupäevana on kasutatud 2021. aasta 31. detsembrit, kuna viimane teist tüüpi diabeedi diagnoosi saamise kuupäev andmestikus on 30.12.2021.

Joonisel 1 on gruppide elulemuskõverad liitumishetkel vähemalt 40 aastat vanad olnud geenidoonorite seas. Joonise x-teljel on ajavahemiku pikkus geenivaramuga liitumisest teist tüüpi diabeedi diagnoosi saamise või tsenseerimiseni. Joonisel märgitud tõenäosus y-teljel on tõenäosus, et teist tüüpi diabeedi diagnoosi ei saada enne vaadeldavat aega. On näha, et kehamassiindeksi kõigis kategooriates on grupis, milles polügeense riskiskoori väärtused on suuremad 99-protsentiilist, võrreldes 20-protsentiilist väiksemate väärtustega riskiskooridega grupiga mitmekordne risk haigestuda teist tüüpi diabeeti. Seetõttu on väga oluline, et suure riskiskoori väär-

tusega inimesed saaksid tagasisidet oma diabeedi riski kohta.



Joonis 1: Elulemuskõverad riskiskoori ja KMI gruppides

Nägemaks, millised tunnused teist tüüpi diabeeti haigestumise riski mõjutavad ning kuidas need seda teevad, on teist tüüpi diabeedi diagnoosimise aegadele hinnatud Weibulli parameetriline mudelid. Seejuures on analüüsist välja jäetud inimesed, kes olid Eesti geenivaramuga liitudes enam kui 90 aasta vanused või vähem kui 18-aastased. Mudelid on hinnatud aastates väljendatud ajavahemiku pikkusele, mis oli geenidoonori jaoks teist tüübi diabeedi diagnoosi saamise hetkeks geenivaramuga liitumisest möödunud. Teist tüübi diabeedi saamise kümne aasta riski paremaks hindamiseks on tsenseerituks loetud need andmed, kus diagnoos oli saadud enam kui kümme aastat pärast liitumist. Seejuures on ajaks loetud kümme aastat, et ajaskaala ei oleks liiga pikk ning parameetrilise jaotuse eeldused Weibulli mudeli hindamiseks oleksid paremini täidetud.

Mudelid on hinnatud eraldi kolmes vanusgrupis. Alla 40-aastaste seas on hinnatud mudel vaid neile inimestele, kes on kehamassiindeksi poolest ülekaalulised

(KMI≥25), kuna selles vanusgrupis väiksema kehamassiindeksiga noored tavaliselt diabeeti ei haigestu. Seetõttu pole ka kindel, kas selles grupis andmestikus leiduvad diagnoosid on korrektsed. Enam kui 40-aastaste kuid vähem kui 70-aastaste inimeste jaoks töö käigus hinnatud kolm mudelit. Neist mudelitest esimene on hinnatud neile geenidoonoritele, kes kehamassiindeksi poolest rasvunud ei ole (KMI<30), teine esimese taseme rasvunutele (30≤KMI<35) ning kolmas neile, kelle kehamassiindeks ületab 35 ehk kellel esineb haiguslik ülekaal. Vähemalt 70-aastaste inimeste teise tüübi diabeeti haigestumisele on hinnatud üks mudel.

Enne mudelite hindamist on väärandmete vähendamiseks jäetud andmestikust välja need inimesed, kelle kehamassiindeks on märgitud väiksemaks kui 15 või suuremaks kui 50, samuti need inimesed, kelle vööümbermõõt on väiksem kui 50 või suurem kui 150 sentimeetrit. Tagamaks, et kõiki inimesi oleks mõnda aega jälgitud, on andmestikust enne mudelite hindamist välja jäetud ka nende inimeste andmed, kes on liitunud Tartu Ülikooli Eesti geenivaramuga 2021. aastal või hiljem.

Mudelite hindamiseks on andmestikku loodud FINDRISC riski prognoosimise algoritmi eeskujul binaarsed tunnused geenidoonori elustiili kohta. Need tunnused on, kas inimene on söönud igapäevaselt värskeid puuvilju ja marju, kas ta on igapäevaselt keedetud juurvilju söönud, kas ta on igapäevaselt värskeid juurvilju söönud ning kas ta on olnud nädalas kokku vähemalt kolme tunni vältel kehaliselt aktiivne. Lisaks on mudeleid hinnates loodud andmestikku binaarne tunnus, mis näitab, kas inimene suitsetab või mitte, kuna mudelite hindamisel oli näha olnud, et statistiliselt olulist erinevust ei ole, kas inimene on suitsetanud ja selle siis maha jätnud või ei ole mitte kunagi suitsetanud.

Mudeleid hinnates on kasutatud argumenttunnustena lisaks eelnevalt kirjeldatud binaarsetele tunnustele ka binaarseid indikaatoreid, kas inimene on naine ning kas tal on diagnoositud südame isheemiatõbi, kõrgvererõhktõbi või müokardiinfarkt. Samuti on argumenttunnustena kasutatud inimese vanust, kehamassiindeksit, vööüm-

bermõõtu ning välja valitud polügeense riskiskoori väärtust. Mudelite hindamisel on kontrollitud ka, et hinnatud mudelite argumenttunnuste kordajad ei oleks ajast sõltuvad ning erinevate argumenttunnuste väärtustega inimeste riskide erinevus oleks alati sama. Selleks on kasutatud funktsiooni *cox.zph* (Thernau *et al.*, 2023, lk 32–33).

Uurimaks, kas mudeli prognoositud teist tüüpi diabeeti haigestumise 10 aasta risk vastab tegelikult riskile, on iga mudelit kalibreeritud järgneva algoritmiga.

1. Andmestik on jagatud testandmestikuks ning treeningandmestikuks.
2. Treeningandmestikus on leitud hinnangud Weibulli mudeli argumenttunnuste z kordajatele $\beta_0, \beta_1, \ldots, \beta_q$ ning jaotuse kujuparameetrile α.
3. Testandmestikus on hinnatud 10 aasta risk igale inimesele valemiga

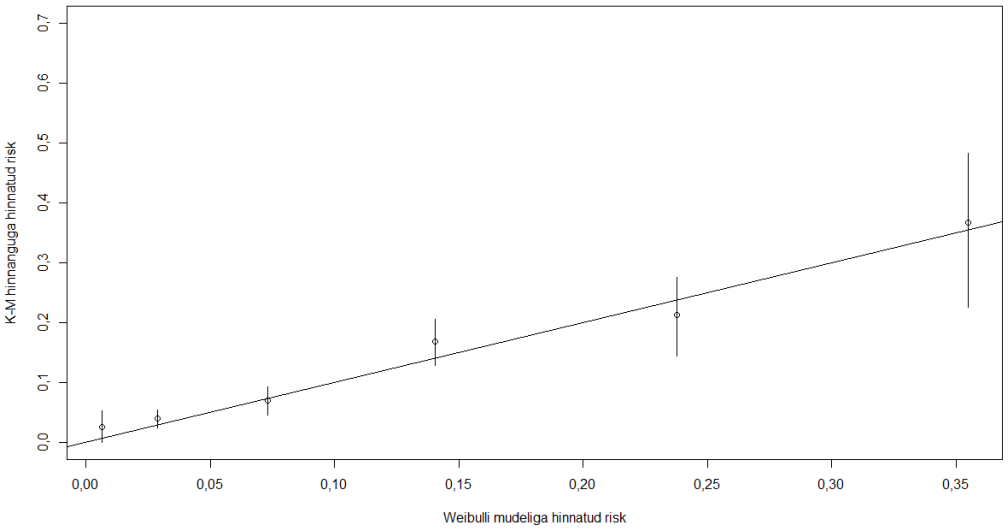
$$1 - \hat{S}_W(10) = 1 - e^{-(10\hat{\lambda})^{\alpha}},$$

   kus

$$\hat{\lambda} = e^{-(\hat{\beta}_0 + \hat{\beta}_1 z_1 + \ldots + \hat{\beta}_q z_q)} = e^{-\hat{\beta}^T (1,z)}$$

4. Testandmestikus on jagatud inimesed gruppidesse riski suuruse põhjal.
5. Igas riskikategoorias on leitud Kaplan-Meieri hinnang 10 aasta riskile kui 1-$\hat{S}_{KM}(10)$ ning võrreldud seda grupi keskmise mudeli prognoositud riskihinnanguga $1 - \overline{\hat{S}_W(10)}$.

Riske on võrreldud joonise abil, kus x-teljel on Weibulli mudeli abil hinnatud 10 aasta risk ning y-teljel Kaplan-Meierihinnang 10 aasta riskile. Joonisel 2 on toodud näide võrdlusjoonisest enam kui 70-aastastele inimestele hinnatud Weibulli mudeli kalibreerimisel. Riskid on jagatud gruppidesse <1%, 1–5%, 5–10%, 10–20%, 20–30% ning >30%. Võrdluseks on joonisel sirge tõusuga 1 ning näidatud on ka Kaplan-Meieri hinnngu abil hinnatud riskide usaldusvahemik.



Joonis 2: Kaplan-Meieri elulemusfunktsiooni hinnanguga ning Weibulli mudeliga hinnatud riskide võrdlus

2.3 Tulemused

Töös on riski hindamisel kasutatud Weibulli mudelit, mis eeldab, et vaadeldud ajad on Weibulli jaotusega. Weibulli jaotuse korral

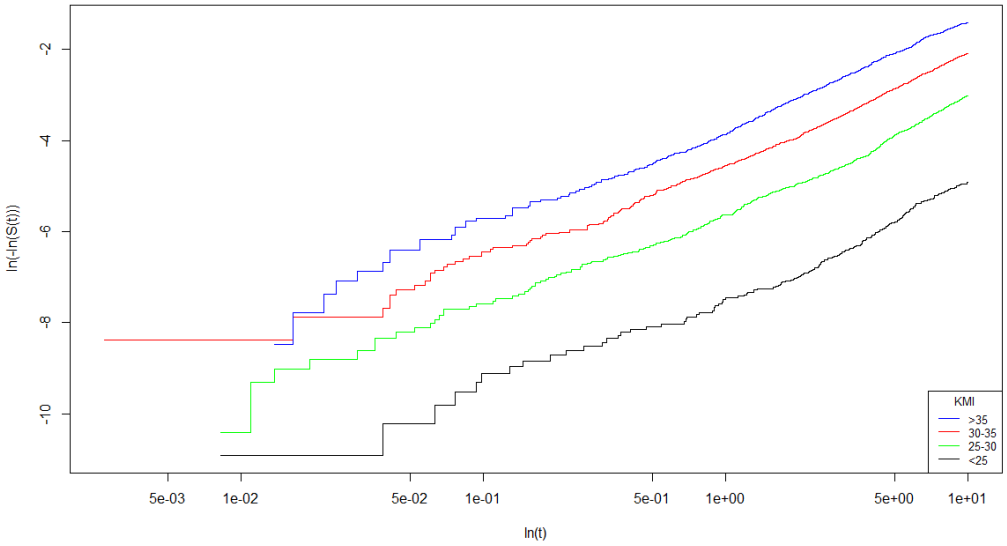
$$\ln(-\ln(S(t))) = \ln\left(-\ln\left(\exp\left(-(\lambda t)^{\alpha}\right)\right)\right) = \ln\left(-\left(-(\lambda t)^{\alpha}\right)\right) = \alpha \ln(\lambda t) =$$
$$= \alpha(\ln(\lambda) + \ln(t)),$$

kust

$$\ln(t) = -\ln(\lambda) - \frac{1}{\alpha}\ln(-\ln(S(t))).$$

Seetõttu peaks olema joonisel, kus x-teljel on logaritmitud väärtused aegadest t ning y-teljel $\ln(-\ln(S(t)))$, sirge tõusuga $\frac{1}{\alpha}$ ning vabaliikmega $-\ln(\lambda)$. Töö käigus on kontrollitud Weibulli jaotuse kehtivust neljas kehamassiindeksi põhjal moodustatud grupis, millest esimeses on inimesed, kelle kehamassiindeks on väiksem kui 25, teises need, kelle kehamasiindeks jääb 25 ja 30 vahele, kolmandas need, kelle

kehamasiindeks jääb 30 ja 35 vahele ning neljandas need, kelle kehamasiiindeks on suurem kui 35. Jooniselt 3 on näha, et kõigis gruppides on vaadeldud ajad geenivaramuga liitumisest kooskõlas Weibulli jaotusega.



Joonis 3: Weibulli jaotuse eelduste kontroll KMI gruppides

Lisaks on Weibulli mudeli korral vajalik võrdeliste riskide eelduse kehtimine. Selle kontrollimiseks on andmetele hinnatud esmalt Coxi võrdeliste riskide mudel ning testitud eelduste kehtimist funktsiooni *cox.zph* abil. Sellega on selgunud, et mitme olulise argumenttunnuse puhul ei ole riskide suhe ajas konstantne ning seega pole eeldused täidetud. Seetõttu on andmestik jagatud gruppidesse KMI ja vanuse alusel ning veendutud, et igas grupis eraldi hinnatud mudelis on eeldused paremini täidetud.

Töö käigus on hinnatud erinevad mudelid inimestele, kes on vähem kui 70 aastat vanad ning neile, kes on vähemalt 70-aastased, kuna pärast 70. eluaaastat ei mõjuta vanus enam teist tüüpi diabeeti haigestumise riski. Samuti on hinnatud eraldi mudel alla 40-aastastele ülekaalulistele (KMI$\geq$ 25) inimestele, kuna teist tüüpi diabeeti soovitatakse tavaliselt kontrollida vähemalt 40- või 45-aastastel inimestel

(Roosimaa *et al.*, 2021) ning vähem kui 40 aastat vanadel KMI poolest normaalkaalulistel inimestel diabeeti üldjuhul ei diagnoosita. Vanusgrupis 40–70 aastat on hinnatud kolm eraldi mudelit kehamassiindeksi põhjal moodustatud gruppides.

Mudelis, mis on hinnatud 18–40-aastastele ülekaalulistele inimestele, osutusid olulisteks tunnusteks inimese vanus, vööümbermõõt ehk tunnus vöö, teist tüüpi diabeedi polügeenne riskiskoor, inimese kehamassiindeks ning kehamassiindeksi ja vööümbermõõdu koosmõju. Tabelis 2.3 on iga tunnuse kohta välja toodud selle kordaja väärtus mudelis, kordaja standardviga ning tunnuse olulisuse tõenäosus mudelis. Lisaks on tabelis samad näitajad mudeli vabaliikme kohta. Mudeli hindamisel on kasutatud 16 132 inimese andmeid, kellest teist tüüpi diabeet oli diagnoositud 187 inimesel. Weibulli jaotuse kujuparameetri väärtuseks on hinnatud $\alpha = 1{,}322$. Jaotuse skaalaparameeter on igale inimesele hinnatud kui $\lambda = \exp(-(\hat{\beta}^T z))$

Tabel 2: Mudel alla 40-aastastele inimestele, kelle KMI on suurem kui 25

Tunnus	$\hat{\beta}$	Standardviga	p-väärtus
Vabaliige	19,628	2,854	<0,001
Vanus	-0,037	0,010	<0,001
Vöö	-0,092	0,026	<0,001
PRS	-0,565	0,066	<0,001
KMI	-0,362	0,082	<0,001
Vöö*KMI	0,002	0,001	0,002

Seega saab geenivaramu andmetel hinnatud mudeli põhjal teist tüüpi diabeedi diagnoosi saamise kümne aasta riski 18–40-aastaste ülekaaluliste täiskasvanute seas hinnata kui

$$1 - \hat{S}(10) = 1 - \exp(-(10\hat{\lambda})^{1{,}322}),$$

kus

$$\hat{\lambda} = e^{-19{,}628+0{,}037*vanus+0{,}092*vöö+0{,}565*PRS+0{,}362*KMI-0{,}002*vöö*KMI}.$$

Hinnates mudelit nimestele vanuses 40–70 eluaastat ja kehamassiindeksiga alla 30, on kasutatud 40 113 inimese andmeid. Neist teise tüübi diabeedi diagnoosi oli saanud 657 inimest. Mudelis osutusid teist tüüpi diabeeti haigestumist oluliselt mõjutavateks tunnusteks inimese vanus, vööümbermõõt, kehamassiindeks, polügeenne riskiskoor ning binaarsed tunnused, mis näitavad, kas inimesel on diagnoositud kõrgvererõhktõbi (tunnus HT) või müokardiinfarkt (tunnus MI) ning kas ta suitsetab (tunnus suits). Olulisi koosmõjusid tunnuste vahel mudeli hindamisel tuvastatud ei ole. Tabelis 2.3 on välja toodud mudeli vabaliige ning kõigi mudeli argumenttunnuste kordajad koos standardveaga. Lisaks on tabelis tunnuste ja vabaliikme olulisuse tõenäosused.

Tabel 3: Mudel 40–70-aastastele inimestele, kelle KMI on väiksem kui 30

Tunnus	$\hat{\beta}$	Standardviga	p-väärtus
Vabaliige	13,810	0,629	<0,001
Vanus	-0,038	0,005	<0,001
Vöö	-0,015	0,004	<0,001
PRS	-0,560	0,039	<0,001
KMI	-0,180	0,019	<0,001
HT	-0,639	0,081	<0,001
MI	-0,298	0,093	<0,001
Suits	-0,352	0,078	<0,001

Mudeli põhjal on Weibulli jaotuse kujuparameetriks hinnatud $\alpha = 1{,}144$. Seega saab selles vanuses inimesel, kelle kehamassiindeks on väiksem kui 30, teist tüüpi diabeeti haigestumise 10 aasta riski hinnata kui

$$1 - \hat{S}(10) = 1 - \exp(-(10\hat{\lambda})^{1{,}144}),$$

kus

$$\hat{\lambda} = e^{-13{,}81+0{,}038*vanus+0{,}015*vöö+0{,}56*PRS+0{,}18*KMI+0{,}639*HT+0{,}298*MI+0{,}352*suits}.$$

Mudeli põhjal, mis on hinnatud 40–70-aastastele inimestele, kelle kehamassiindeks on 30 ja 35 vahel, mõjutab teist tüüpi diabeeti haigestumise riski ka inimese sugu. Mudelis osutusid oluliseks inimese vanus, vööümbermõõt, polügeense riskiskoori väärtus, kehamassiindeks ning indikaatortunnused, mis näitavad, kas inimesel on diagnoositud kõrgvererõhktõbi, kas ta suitsetab ning kas tegemist on naisega (sugu). Indikkaatortunnuse *sugu* väärtus on 0, kui inimene on mees. Ka selles mudelis ei olnud tunnustel olulisi koosmõjusid ning mudeli tunnuste kordajate ning vabaliikme väärtuste hinnangud on standardvigade ja olulisuse tõenäosustega välja toodud tabelis 2.3. Mudeli hindamisel on kasutatud 8 639 inimese andmeid. Neist teise tüübi diabeedi diagnoosi oli saanud 627 inimest.

Tabel 4: Mudel inimestele vanuses 40–70 aastat, kelle KMI on 30–35

Tunnus	$\hat{\beta}$	Standardviga	p-väärtus
Vabaliige	11,097	0,911	<0,001
Vanus	-0,029	0,005	<0,001
Vöö	-0,017	0,004	<0,001
PRS	-0,504	0,040	<0,001
KMI	-0,103	0,027	<0,001
HT	-0,712	0,091	<0,001
Suits	-0,240	0,087	0,006
Sugu	0,217	0,086	0,012

Weibulli jaotuse kujuparameetri väärtuseks on selle mudeli juures hinnatud $\alpha = 1{,}124$. Seega saab 40–70-aastaste inimeste, kelle kehamassiindeks on vahemikus 30–35, 10 aasta riski haigestuda teist tüüpi diabeeti hinnata kui

$$1 - \hat{S}(10) = 1 - \exp(-(10\hat{\lambda})^{1{,}124}),$$

kus

$$\hat{\lambda} = e^{-11{,}097+0{,}029*vanus+0{,}017*vöö+0{,}504*PRS+0{,}103*KMI+0{,}712*HT+0{,}240*suits-0{,}217*sugu}.$$

Vähemalt teise taseme rasvunutele 40–70-aastastele inimestele hinnatud mudelis osutusid teist tüüpi diabeetei haigestumisele olulist mõju avaldavateks tunnusteks inimese vanus, vööümbermõõt, kehamassiindeks, teist tüüpi diabeedi polügeenne riskiskoor ning indikaatortunnused, kas inimesel on kõrgvererõhktõbi ning kas ta suitsetab. Mainitud tunnuste kordajate ja vabaliikme väärtuste hinnangud on välja toodud tabelis 2.3. Tunnuste vahel ei ole koosmõjusid ka selles mudelis ning tase madalama KMI grupi mudeli tunnustest on erinev vaid see, et inimestel, kelle kehamassiindeks on vähemalt 35 ei mõjuta sugu teist tüüpi diabeeti haigestumise riski oluliselt. Mudeli hindamisel on kasutatud 3 224 inimese andmeid, kusjuures teise tüübi diabeedi diagnoosi oli saanud 451 inimest.

Tabel 5: Mudel inimestele vanuses 40–70 aastat, kelle KMI on suurem kui 35

Tunnus	$\hat{\beta}$	Standardviga	p-väärtus
Vabaliige	9,161	0,738	<0,001
Vanus	-0,017	0,006	0,004
Vöö	-0,020	0,004	<0,001
PRS	-0,442	0,052	<0,001
KMI	-0,049	0,015	<0,001
HT	-0,439	0,116	<0,001
Suits	-0,293	0,113	0,010

Weibulli jaotuse kujuparameetri väärtuseks on selle mudeli juures hinnatud $\alpha = 1{,}045$. Seega saab 40–70-aastaste haiguslikus ülekaalus inimeste, 10 aasta riski haigestuda teist tüüpi diabeeti hinnata kui

$$1 - \hat{S}(10) = 1 - \exp(-(10\hat{\lambda})^{1{,}045}),$$

kus

$$\hat{\lambda} = e^{-9{,}161+0{,}017*vanus+0{,}02*vöö+0{,}442*PRS+0{,}049*KMI+0{,}439*HT+0{,}293*suits}.$$

Pärast 70-ndat eluaastat inimese vanus enam mudelis oluliseks ei osutunud. Mu-

deli hindamisel on kasutatud 5 744 inimese andmeid. Seejuures teist tüüpi diabeedi diagnoosi oli saanud 361 inimest. Tabelis 2.3 on välja toodud vanusgrupis 70–90 eluaastat hinnatud mudeli olulised tunnused kordajate hinnangute, nende standardvigade ja olulisuse tõenäosustega. Mudeli põhjal mõjutavad vähemalt 70-aastaste inimeste puhul teist tüüpi diabeeti haigestumise riski inimese vööümbermõõt, polügeenne riskiskoor, kehamassiindeks, see, kas inimesel on diagnoositud kõrgvererõhktõbi ning vööümbermõõdu ja polügeense riskiskoori koosmõjud kehamassiindeksiga.

Tabel 6: Mudel üle 70-aastastele inimestele

Tunnus	$\hat{\beta}$	Standardviga	p-väärtus
Vabaliige	16,268	2,574	<0,001
Vöö	-0,071	0,025	0,004
PRS	-1,882	0,405	<0,001
KMI	-0,307	0,084	<0,001
HT	-0,665	0,204	<0,001
Vöö*KMI	0,002	0,001	0,020
PRS*KMI	0,042	0,013	<0,001

Mudeli põhjal on Weibulli jaotuse kujuparameetriks vähemalt 70-aastaste inimeste seas hinnatud $\alpha = 0{,}838$. Seega saab 70–90-aastaste teist tüüpi diabeeti haigestumise 10 aasta riski hinnata kui

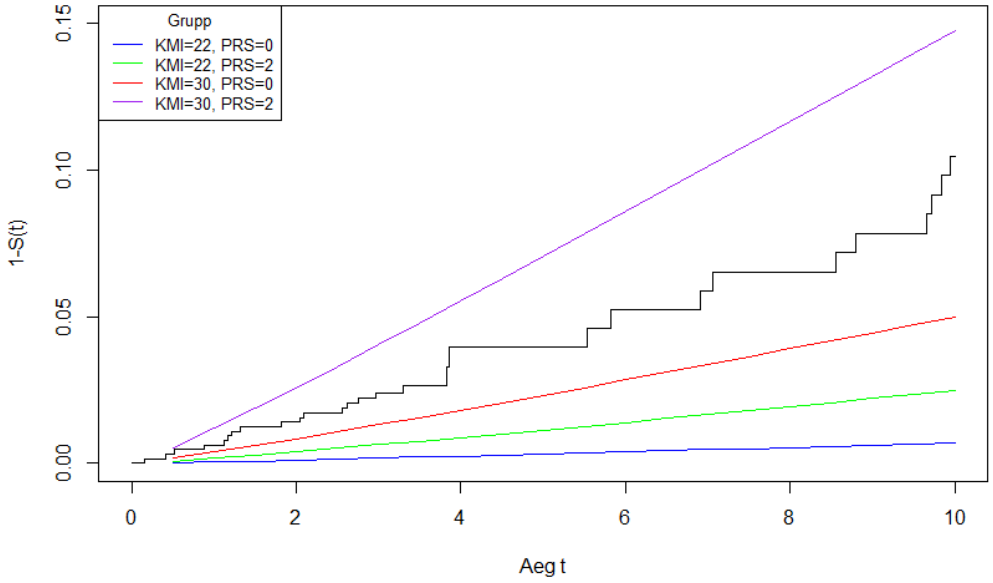
$$1 - \hat{S}(10) = 1 - \exp(-(10\hat{\lambda})^{0{,}838}),$$

kus

$$\hat{\lambda} = e^{-16{,}268+0{,}071*vöö+1{,}882*PRS+0{,}037*KMI+0{,}665*HT-0{,}002*vöö*KMI-0{,}042*PRS*KMI}.$$

Joonisel 4 on võrreldud 50-aastaste meeste, kes ei suitseta ning kellel ei ole diagnoositud kõrgvererõhktõbe ega müokardiinfarkti, mudeli põhjal hinnatud teist tüüpi diabeeti haigestumise riske erinevate kehamassiindeksite ning polügeensete ris-

kiskooride korral. Võrdluseks on joonisel välja toodud ka Kaplan-Meieri hinnang 50-aastaste meeste teist tüüpi diabeeti haigestumise riskile. Jooniselt on näha, et suurema kehamassiindeksiga meeste risk on suurem. Samuti on näha, et kui kahel inimesel on kõik näitajad peale polügeense riskiskoori samas, on risk haigestuda teist tüüpi diabeeti kõrgem sellel, kellel on kõrgem riskiskoori väärtus.



Joonis 4: 50-aastaste meeste riskihinnangu võrdlus

Kokkuvõttes võib 50-aastaste meeste 10 aasta diabeedirisk sõltuvalt riskitegurite väärtusest varieeruda vahemikus 1%–15%. Samuti on jooniselt näha, et ka kõrge polügeense riskiskoori korral mõjutab haiguseriski suurel määral inimese kehamassiindeks, mida mõjutab kehakaal. Näiteks üldiselt samade näitajatega meeste puhul, kelle PRS on 2, kuid kellest ühel on KMI 30 ning teisel 22, on kõrgema kehamassiindeksiga mehel teisest enam kui 10% võrra kõrgem risk. Ka väiksema polügeense riskiskoori puhul on jooniselt näha, et kõrgema kehamassiindeksiga mehel on võrreldes madalama kehamassiindeksiga mehega mitmekordne risk

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli koostada mudel, mille abil on võimalik anda Tartu Ülikooli Eesti geenivaramu geenidoonoritele tagasisidet nende teist tüüpi diabeedi saamise 10 aasta riski kohta. Töö käigus valiti välja teist tüüpi diabeediga enim seotud polügeenne riskiskoor (PGS002771) ning hinnati geenivaramu andmeid kasutades viis erinevat Weibulli parameetrilist mudelit, milles olid lisaks polügeensele riskiskoorile muud tunnused, mis grupis teist tüüpi diabeeti haigestumise riski oluliselt mõjutasid. Mudelid hinnati eraldi vähem kui 40-aastastele kehamassiindeksi poolest ülekaalulistele täiskasvanutele, 70–90-aastastele inimestele ning 40–70-aastastele inimestele. Seejuures viimases grupis hinnati eraldi mudelid gruppides, kus inimese kehamassiindeks oli väiksem kui 30, 30–35 või vähemalt 35. Mudelite põhjal hinnati inimeste 10 aasta risk haigestuda teist tüüpi diabeeti.

Selgus, et enam kui 70-aastastel inimestel ei mõjuta vanus diabeeti haigestumise riski, kuid teistes vanusegruppides on vanus oluline riskifaktor. Kõigis gruppides mõjutab teist tüüpi diabeeti haigestumise riski oluliselt inimese vööümbermõõt, tema kehamassiindeks ning teist tüüpi diabeedile arvutatud polügeenne riskiskoor. Vähemalt 40-aastastel inimestel on oluliseks riskifaktoriks ka see, kas neil on diagnoositud kõrgvererõhktõbi. 40–70-aastastel inimestel tõstab teist tüüpi diabeeti haigestumise riski see, kas nad suitsetavad. 18–40-aastaste ning 70–90-aastaste inimeste mudelites on ka tunnuste koosmõjusid. 40–70-aastastel inimestel mõjutab neil, kelle KMI on väiksem kui 30, teist tüüpi diabeeti haigestumist ka see, kas neil on diagnoositud müokardiinfarkt ning neil, kelle KMI on 30–35 nende sugu.

Töös hinnatud mudelid on kalibreeritud, kuid et mõnel juhul esines mudelite või nende eelduste kontrollis kõrvalekaldeid, soovitakse enne geenidoonoritele tagasiside andmist mudelitega edasi töötada. Pärast mudelite kontrollimist ja korrigeerimist on vaja välja töötada sobivad tekstilised ja visuaalsed vahendid riskihinnangute kujutamiseks ning implementeerida algoritm vastava tarkvaralahendusena.

Kasutatud allikad

Ali, Omar (2013). „Genetics of type 2 diabetes“. *World journal of diabetes* 4.4, lk. 114. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3746083/ (vaadatud 07.05.2023).

Almgren, Peter, Maarit Lehtovirta, Boris Isomaa, Leena Sarelin, Marja-Riitta Taskinen, Valeriya Lyssenko, Tiinamaija Tuomi, Leif Groop ja Botnia Study Group (2011). „Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study“. *Diabetologia* 54, lk. 2811–2819. URL: https://pubmed.ncbi.nlm.nih.gov/21826484/ (vaadatud 06.05.2023).

Association, American Diabetes *et al.* (2008). „Standards of medical care in diabetes 2008“. *Diabetes care* 31, lk. 12–54.

Choi, Shing Wan, Timothy Shin Heng Mak ja Paul F. O'Reilly (2020). „A guide to performing Polygenic Risk Score analyses“. *Nature Protocols* 15 (9), 2759—2772. DOI: 10.1038/s41596-020-0353-1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7612115/ (vaadatud 09.04.2023).

Esko, Tõnu, Reedik Mägi, Krista Fischer, Lili Milani, Kristi Läll, Maris Alver, Mart Kals abd Kristi Krebs, Sulev Reisberg ja Tõnis Tasa (2019). „Geneetika- ja genoomikaalased alusuuringud personaalmeditsiini rakendamiseks Eestis“. *Eesti Vabariigi preemiad*, lk. 94–107. URL: https://vana.akadeemia.ee/_repository/file/PUBLIKATSIOONID/2019/EV_preemiad_2019_sisu.pdf (vaadatud 26.03.2023).

Flannick, Jason ja Jose C Florez (2016). „Type 2 diabetes: genetic data sharing to advance complex disease research“. *Nature Reviews Genetics* 17.9, lk. 535–549. URL: https://pubmed.ncbi.nlm.nih.gov/27402621/ (vaadatud 07.05.2023).

Hemminki, Kari, Xinjun Li, Kristina Sundquist ja Jan Sundquist (2010). „Familial risks for type 2 diabetes in Sweden“. *Diabetes care* 33.2, lk. 293–297. URL: https://pubmed.ncbi.nlm.nih.gov/19903751/ (vaadatud 07.05.2023).

IDF Diabetes Atlas (2021). 10. väljaanne. URL: https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf (vaadatud 24.02.2023).

Leitsalu, Liis, Toomas Haller, Tõnu Esko, Mari-Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, Pauline C Ng, Reedik Mägi, Lili Milani *et al.* (2015). „Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu“. *International journal of epidemiology* 44.4, lk. 1137–1147. (Vaadatud 06.05.2023).

Lindström, Jaana, Pilvikki Absetz, Katri Hemiö, Päivi Peltomäki ja Markku Peltonen (2010). „Reducing the risk of type 2 diabetes with nutrition and physical activity – efficacy and implementation of lifestyle interventions in Finland“. URL: https://www.cambridge.org/core/journals/public-health-nutrition/article/reducing-the-risk-of-type-2-diabetes-with-nutrition-and-physical-activity-efficacy-and-implementation-of-lifestyle-interventions-in-finland/38A64837E83871F0CC955B997E8B44A4.

Lindström, Jaana ja Jaakko Tuomilehto (2023). *FINDRISC (Finnish Diabetes Risk Score)*. URL: https://www.mdcalc.com/calc/4000/findrisc-finnish-diabetes-risk-score (vaadatud 08.05.2023).

Lyssenko, Valeriya, Peter Almgren, Dragi Anevski, Roland Perfekt, Kaj Lahti, Michael Nissén, Bo Isomaa, Bjorn Forsen, Nils Homstrom, Carola Saloranta *et al.* (2005). „Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of type 2 diabetes“. *Diabetes* 54.1,

lk. 166–174. URL: https://pubmed.ncbi.nlm.nih.gov/15616025/ (vaadatud 07.05.2023).

Läll, Kristi, Reedik Mägi, Andrew Morris, Andres Metspalu ja Krista Fischer (2017). „Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores“. *Genetics in Medicine* 19 (3), lk. 322–329. URL: https://pubmed.ncbi.nlm.nih.gov/27513194/ (vaadatud 26.03.2023).

Mahajan, Anubha, Cassandra N Spracklen, Weihua Zhang, Maggie CY Ng, Lauren E Petty, Hidetoshi Kitajima, Grace Z Yu, Sina Rüeger, Leo Speidel, Young Jin Kim *et al.* (2022). „Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation“. *Nature genetics* 54.5, lk. 560–572. URL: https://pubmed.ncbi.nlm.nih.gov/35551307/ (vaadatud 07.05.2023).

Meigs, James B, L Adrienne Cupples ja PW Wilson (2000). „Parental transmission of type 2 diabetes: the Framingham Offspring Study.“ *Diabetes* 49.12, lk. 2201–2207. URL: https://pubmed.ncbi.nlm.nih.gov/11118026/ (vaadatud 07.05.2023).

Mis on diabeet? (2023). Eesti Diabeediliit. URL: http://www.diabetes.ee/mis-on-diabeet# (vaadatud 24.02.2023).

PGS Catalog (2023). URL: https://www.pgscatalog.org/ (vaadatud 04.05.2023).

Roosimaa, Mart, Aune Rehema, Evelin Raie, Ulvi Tammer-Jäätes, Kaia Tammiksaar, Maarja Randväli, Marko Tähnas, Anneli Vatsa ja Marelle Maiste (2021). *2. tüüpi diabeedi diagnostika ja ravi*. URL: https://ravijuhend.ee/tervishoiuvarav/juhendid/154/2-tuupi-diabeedi-diagnostika-ja-ravi (vaadatud 04.05.2023).

Tartu Ülikooli Eesti geenivaramu (2023). URL: https://geenidoonor.ee/geenivaramu (vaadatud 05.05.3023).

Thernau, Terry M, Thomas Lumley, Atkinson Elizabeth ja Crowson Cynthia (2023). *Survival Analysis*. URL: https://cran.r-project.org/web/packages/survival/survival.pdf (vaadatud 21.03.2023).

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karmel Teder,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Teise tüübi diabeedi riski prognoosimine ja tagasiside algoritmi väljatöötamine TÜ Eesti geenivaramu andmetel“, mille juhendajad on Krista Fischer ja Natalia Pervjakova, 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 4.0, 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 intellektuaaloman-di ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Karmel Teder
09.05.2023