Peatükk 6

Diagnostilised testid II.

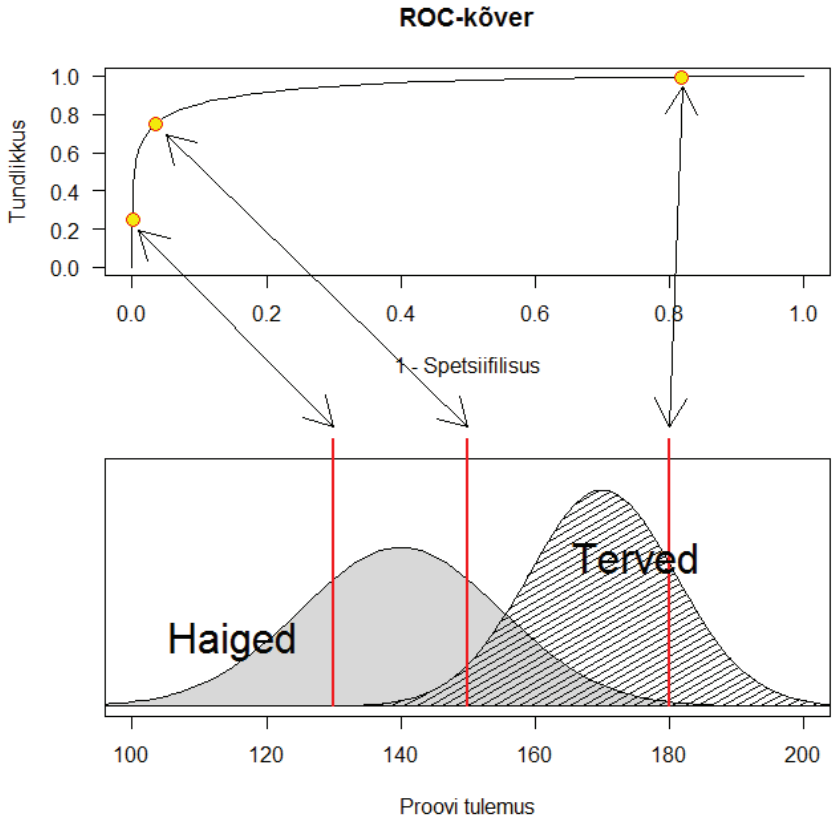
Kuhu tõmmata piir, millest alates hakkame inimest haigeks pidama? Referentsväärtust saab ju siia ja sinnapoole nihutada. Näiteks kui kõrge peaks olema inimese kehatemperatuur, selleks et me inimest haigeks peaksime? Millisest kehatemperatuurist alates ütleme, et tal on palavik? Näiteks ameeriklaste Centers of Disease Control and Prevention (CDC) defineerib, et inimesel on palavik siis, kui tema kehatemeperatuur on üle 38°C, (CDC, 2017). Eestis arvatakse aga sageli, et inimesel on palavik, kui tema kehatemperatuur on üle 37°C, (Virtuaalkliinik, 2016). Wikipeedia aga näiteks arvab, et palavik algab 37,5°C kraadist, (Wikipedia, 2020).

Erinevate referentsväärtuste kasutamine viib aga erinevate spetsiifilisuse ja tundlikkuse väärtusteni. Kuna referents- või normväärtust võidakse mõnikord erinevatel põhjustel muuta, siis soovitakse vahel teada, milliseid erinevaid spetsiifilisuse ja tundlikkuse väärtuseid on võimalik referentsväärtust muutes saavutada, vaata ka probleemi illustreerivat joonist 6.1 . Kõverat, mis kirjeldab võimalikke spetsiifilisuse ja tundlikkuse väärtuseid (mida võime saavutada erinevaid referntsväärtuseid kasutades) tuntakse ROC-kõvera (*ROC-curve*) nime all.

6.1 ROC-kõver

ROC-kõvera nimi tuleneb väljendist ***R**eceiver **O**perating **C**haracteristic curve*. See väljund pärineb ajast, kui termin leiutati sidetööstuse jaoks — üle mürarikaste sidekanalite kanti üle informatsiooni ja signaali vastuvõtja seisis raske otsuse ees: kas see informatsioon, mida talle saadeti, oli nüüd 1 või 0. Kui aga sama graafiku tüüp meditsiinis kasutusele võeti diagnoosimeetodite täpsuse kirjeldamiseks (kui hästi suudame määrata, kas patsiendil on

Joonis 6.1: ROC-kõver. Referentsväärtuste muutumise korral muutuvad ka testi tundlikkus ja spetsiifilisus. Milliseid võimalikke tundlikkuse-spetsiifilisuse väärtuseid võime referentsväärtust muutes näha, seda iseloomustab ROC-kõver.

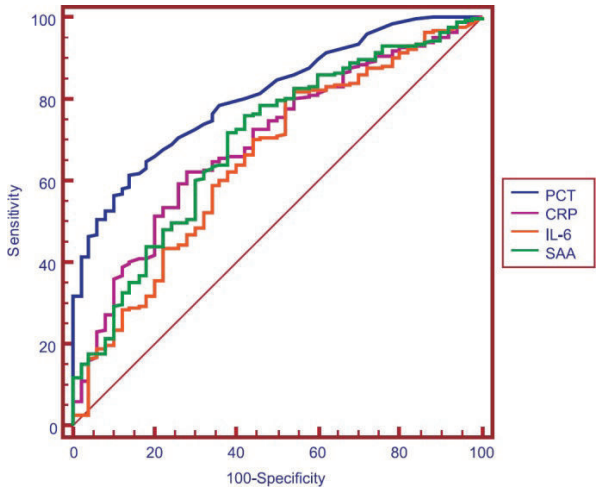


haigus või mitte), siis võeti üle küll graafiku lühendatud nimi (ROC-kõver), kuid heas seltskonnas välditakse pigem selle lühendi päritolu ja tähenduse meenutamist.

Sageli võime kohata ka võrdlevaid ROC-kõveraid samal joonisel. Näiteks joonisel 6.2 on kujutatud erinevaid meetodeid, mida saab kasutada otsustamiseks, kas palavikuga patsiendil on bakteriaalne infektsioon (sellisel juhul on patsiendi ravimisel kasu antibiootikumidest) või mitte (palavik on tingitud näiteks viirusinfektsioonist). Eestis kasutavad perearstid bakteriaalse in-

fektsiooni tuvastamiseks vereproovist määratud C-reaktiivset valku (CRV), haiglasse sattunud patsientide kohta tehakse otsus aga prokaltsitoniini (joonisel PCT) põhjal. Jooniselt näeme, et mistahes fikseeritud spetsiifilisuse korral on prokaltsitoniini test tundlikkum — ehk suudab üles leida rohkem bakteriaalse infektsiooni all kannatajaid. Joonis on võetud artiklist (Qu *et al.*, 2015).

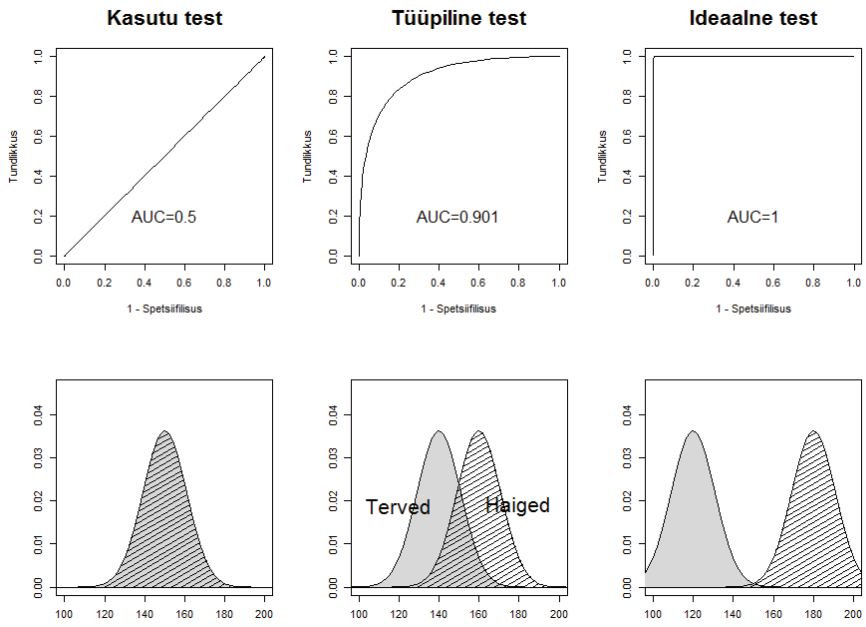
Joonis 6.2: ROC-kõverate võrdlus. Kui hästi on võimalik erinevate verenäitajate abil tuvastada bakteriaalse infektsiooni olemasolu.



Millise kujuga on halva testi ROC-kõver (halb test - test mis põhineb näitajal, mis ei sisalda tegelikult informatsiooni haiguse olemasolu kohta), millise kujuga on hea testi ROC-kõver? Ka halva testiga saame saavutada tundlikkuse 1 (ehk 100%). Lihtsalt loeme kõik testitavad haigeteks — sellisel juhul suudame haiguse üles leida kõigil, kes ka tegelikult on haiged. Kui loeme kõik testitavad haigeteks, siis on muidugi meie spetsiifilisus väga madal, 0. Ehk 1-Spetsiifilisus (mis näitab, kui paljudele tervetele me ekslikult paneme haiguse diagnoosi) väärtus oleks 1 või 100%. Samas on aga alati võimalik Spetsiifilisus tõsta 100% peale — näiteks lugedes kõik testitavad terveteks. Tundlikkus kukub sellise lähenemise puhul muidugi nulliks. Aga iga testi puhul (mistahes näitaja alusel diagnoose pannes) läbib ROC-kõver punkte (0;0) ja (1;1). Kui paneksime diagnoose juhuslike arvude generaatori poolt tekitatud juhuslike suuruste põhjal, näiteks kasutaksime diagnoosipanekuks suurust $U \sim U(0;1)$ (need juhuslikud suurused ei sisalda tegelikult mittemingit informatsiooni selle kohta, kas inimesel on otsitav

haigus või mitte), siis võiksime saavutada testi tundlikkuse 80% (kui loeme kõik inimesed, kelle puhul juhuslik suurus U oli väiksem kui 0,8 haigeteks). Paraku sellisel juhul paneksime ka 80% tervetele ekslikult haigusdiagnoosi. Seega kui kasutaksime haiguse testimisel näitajat, mis ei sisalda mittemingit informatsiooni haiguse olemasolu kohta, näeks ROC-kõver välja kui ruudu diagonaal mis ühendab punkte (0;0) ja (1;1). Kui kasutut näitajat kasutades sooviksime saavutada tundlikkust x%, siis peame paraku ka x%-le tervetest ekslikult panema haigusdiagnoosi (ja sellisel juhul 1-spetsiifilisus = x). Kui aga tegemist on perfektse testiga, mille abil on võimalik (vähemalt mingi referentsväärtuse korral) täiuslikult haiged tervetest eristada, siis on võimalik saavutada samaaegselt 100% spetsiifilisust ja 100% tundlikkust. Sellise testi puhul ROC-kõver läbib puntkti (0;1). Vaata ka joonist 6.3

Joonis 6.3: ROC-kõverad kasutu testi, tüüpilise testi ja ideaalse testi korral.



6.2 ROC-kõvera alune pindala (AUC)

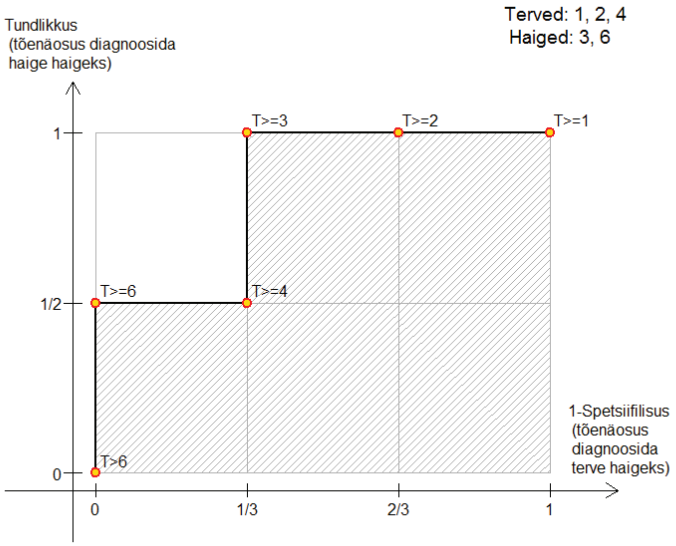
Mida kõrgemal on ROC-kõver, seda parema testiga on tegemist - seda suurema tundlikkuse me võime mistahes spetsiifilisuse korral saavutada. Mida

kõrgemal on aga ROC-kõver, seda suurem on ROC-kõvera alune pindala ehk AUC (***A**rea **U**nder the **C**urve*). Maksimaalne võimalik ROC-kõvera alune pindala on 1 (saavutatav sellise testi puhul, mis suudab perfektselt haigeid tervetest eristada ehk sellise testi puhul saavutatakse mingit referentsväärtust kasutades tundlikkus 100% ja spetsiifilisus 100%. Kasutu testi puhul (haigeks diagnoositud haigete osakaal= haigeks diagnoositud tervete osakaal) on AUC väärtus 1/2. AUC väärtus on väiksem 0,5-st siis, kui meie poolt kasutatav näitaja sisaldab informatsiooni haiguse olemasolu kohta, aga me kasutame seda näitajat valesti — näiteks kui me lahterdaksime kõrge palavikuga inimesi terveteks ja ilma palavaikuta inimesi haigeteks, võiks meie testi AUC-väärtus tulla väiksem kui 0,5.

Üks võimalus AUC-väärtust interpreteerida on muidugi ROC-kõvera alune pindala. Aga sellele näitajale on võimalik anda ka teistsugune interpretatsioon. Selle teistsuguse interpretatsiooni avastamiseks vatame esmalt ühte lihtsat näidet ROC-kõvera joonistamise ja AUC-väärtuse arvutamise kohta. Oletame, et oleme mõõtnud mingi näitaja (T) väärtuse kahel tegelikult haigel (mõõtmistulemused: 3; 6) ja kolmel tegelikult tervel inimesel (mõõtmistulemused: 1; 2; 4). Kui hästi antud näitaja suudab haigeid tervetest eristada? Kui kasutaksime näiteks normväärtust või referentsväärtust 3 (kui $T >= 3$ siis loeme inimese haigeks) siis saavutame tundlikkuse 1 (2/2; kõik meie valimis olevad haiged saaksid ju positiivse testitulemuse ehk oleksid testi arvates haiged), aga paraku oleme ka ühele tegelikult tervele inimesele öelnud, et ta on haige — 1-spetsiifilisus seega on 1/3. Kui aga tõstaksime normväärtuse 6 peale (kui $T >= 6$ siis on inimene haige), siis saaksime spetsiifilisuse üheks (1-spetsiifilisus=0/3), aga testi tundlikkus langeks 1/2 peale (sest ühele tegelikule haigele me enam testi abil haigusdiagnoosi ei pane), vaata ka joonist 6.4.

Paneme tähele, et mistahes andmestiku korral on meil võimalik saavutada tervete arv+1 erinevat spetsiifilisuse väärtust ja haigete arv +1 erinevat tundlikkuse väärtust. Kokku saab siis joonisele tekitada tervete arv $\cdot$ haigete arv ristkülikukest (sellist ristkülikut, nagu kujutatud joonisel 6.4). Kui paljud neist ristkülikutest jäävad ROC-kõverast allapoole? Hakkame seda vaatama veergude kaupa. Kui kõrgel on ROC-kõver esimeses veerus? Selle kõrguse saame aga leida vastates küsimusele: kui suurel osal haigetest on uuritava näitaja väärtus suurem kui kõige suuremat näitaja T väärtust omaval tervel? ROC kõvera kõrguse teises veerus saaksime leida, kui võtaksime suuruselt 2. terve ja vaataksime, kui suurel osal haigetest on uuritava näitaja väärtus suurem kui 2. tervel jne. Kui palju ristkülikuid kokku ROC-kõvera alla jääb? Selle teadasaamiseks peaksime moodustama kõikvõimalikud haige-terve paarid (antud näite puhul siis $2 \cdot 3 = 6$ paari) ja vaatama,

Joonis 6.4: ROC-kõver ja AUC lihtsa näite korral.



kui paljud neist paaridest on sellised, kus haigel on uuritava tunnuse väärtus suurem kui tervel. Iga ristküliku pindala aga on $1/\#terveid \cdot 1/\#haigeid$ (kus #haigeid on haigete arv ja #terveid tähistab tegelikult tervete uuritute arvu), seega saame kokku arvutusvalemi AUC väärtuse leidmiseks:

$$\hat{AUC} = \frac{\#(T_{haige} > T_{terve})}{\#terveid \cdot \#haigeid},$$

kus $\#(T_{haige} > T_{terve})$ tähistab nende paaride arvu, kus näitaja T väärtus on haigel suurem kui tervel. Saadud AUC hindamiseks kasutatav valem võimaldab aga AUC väärtust interpreteerida järgmisel viisil: kui valime juhuslikult ühe tegelikult terve inimese ja haigete seast nopime juhuslikult välja ühe tegelikult haige, siis millise tõenäosusega on haigel mõõdetud näitaja T väärtus suurem kui tervel,

$$P(AUC) = P(T_{haige} > T_{terve}).$$

Kui mingis teises olukorras lahterdaksime haigeteks inimesi kellel tunnuse T väärtus on väiksem mingist referntsnivoost siis ka antud AUC-väärtuse

interpretatsioonis muutuks võrratuse märk: AUC näitaks siis tõenäosust, et haigel on uuritava näitaja väärtus väiksem kui tervel.

Märkus. Üks võimalus AUC-väärtuse leidmiseks praktikas. Mitteparameetrilise Mann-Whitney (Wilcoxoni) testi korral kasutatakse teststatistikuna nende paaride arvu, mille puhul ühest valimist pärit väärtus on suurem kui teisest valimist pärit väärtus. See võimaldab Mann-Whitney (Wilcoxoni) teststatistikut kasutades kergesti leida ka testi AUC-väärtuse. Peame lihtsalt teststatistiku väärtuse jagama kõikmõeldavate haige-terve paaride arvuga:

```
> T=c(1,2,4,3,6)
> terve=c(1,1,1,0,0)
> wilcox.test(T~terve)

        Wilcoxon rank sum test

data:  T by terve
W = 5, p-value = 0.4
alternative hypothesis: true location shift is not
                  equal to 0

> AUC=wilcox.test(T~terve)$statistic /
              (sum(terve==1)*sum(terve==0))
> AUC
        W
0.8333333
```

Siin tuleb muidugi jägida, et kergesti oleksime võinud saada Wilcoxoni teststatistiku väärtuseks ka 1 (näiteks siis, kui oleksime kasutanud indikaatortunnuse terve asemel indikaatortunnsut haige mis tähistaks 1-ga seda, et inimene on tegelikult haige — sellisel juhul oleks kokku loetud paare, kus haigel on uuritava tunnuse väärtus väiksem kui tervel). Sellisel juhul peaksime saadud teststatistiku väärtuse ennem jagamistehet kõigi võimalike haigete-tervete paaride arvust (6-st) lahutama.

Näites toodud Wilcoxoni test on ühtlasi testiks, mis kontrollib hüpoteesi: H0: AUC=0.5. Kui Wilcoxoni testi p-väärtus on väike, siis võime öelda — haiguse diagnoosimiseks kasutatud tunnus sisaldab vähemalt mingitki informatsiooni haiguse olemasolu kohta, tegemist pole täiesti kasutu diagnostilise testiga (sellegipoolest võib test olla kasutu kliinilises praktikas: kui suudame 60% haigetest haiguse üles leida ja samal ajal paneme 45% tervetest ka

ekslikult diagnoosi, siis arstid arvatavasti ei taha sellist testi kasutada).

ROC-kõveraid ja AUC-väärtuseid saab leida ka mitme erineva R-i lisamooduli abil, näiteks võime kasutada lisamoodulit *pROC*:

```
# install.packages("pROC")
library(pROC)
T=c(1,2,4,3,6)
haige=c(0,0,0,1,1)
plot(roc(haige~T), print.auc=TRUE, print.auc.x=0.2,
  print.auc.y=0.2)
```

Mis saab siis, kui soovime haiguse diagnoosimiseks kasutada mitme eri verenäitaja väärtuseid samaaegselt (või lisaks vereproovile arvestada ka anamneesi ehk patsiendi küsitlemise käigus saadud informatsiooni)? Kõige lihtsam võimalus on sellisel juhul kasutada logistilist regressiooni (prognoosime, kas inimene on haige või terve) ja panna logistilise regressioonimudelisse sisse kõik meile teadaolevad ja haiguse olemasolu prognoosivad tunnused. Saadud mudeli prognoosi — nn prognoositud tõenäosust (mis on tõenäosus vaid siis, kui meie valimis on haigete proportsioon sama mis uuritavas populatsioonis) võime siis kasutada selle näitajana, mille põhjal teeme otsuse haiguse olemasolu või puudumise kohta. Ja seda prognoositud tõenäosust kasutades saame soovi korral ka ROC-kõvera joonistada ja AUC-väärtuse leida.

Kui logistilise regressiooni mudelis on palju hinnatavaid parameetreid või kui me lõpliku mudeli valimiseks oleme eelnevalt uurinud ja kõrvale heitnud palju teisi kanditaatmudeleid võime ülesobitamise/mitmese testimise probleemi tõttu saada lõpuks näivalt väga hea AUC-väärtusega mudeli — aga selle mudeli prognoosivõime uute patsientide lahterdamisel haieteks/terveteks ei pruugi olla tegelikult väga hea. Ideaalne lahendus oleks järgmine: jagame oma andmed kahte osasse. Suuremat osa (näiteks 75% vaatlustest) kasutame oma mudeli hindamiseks ja valimiseks (nn treeningandmed). Väiksemat osa (25%), nn testandmestikku, hoiame tagavaraks. Kui oleme kasutatava mudeli lõpuks välja valinud, siis võime oma väljavalitud mudeli prognoosivõimet katsetada testandmestiku peal. ROC-kõvera ja AUC väärtuse leiamegi kasutades testandmestikku.

Kui andmeid on väga vähe, siis me vahel ei raatsi andmestikku test- ja treeningandmeteks jagada. Üks võimalik lahendus sellisel juhul oleks kasutada ristvalideerimist — jätame ühe vaatluse andmestikust välja; hindame ülejäänud vaatluste põhjal logistilise regressioonanalüüsi mudeli; prognoosime saadud mudelit kasutades väljajäänud vaatlust (kas see inimene võiks olla haige või mitte?). Seda protseduuri võime korrata kõigi vaatlustega ja

saadud tulemuste põhjal saame leida ROC-kõvera ja AUC-väärtused. Taoline protseduur leevendab märgatavalt ülesobitamisest tingitud viga, aga ei kaitse parima mudeli valiku käigus tehtava vea eest.

Näide: test- ja treeningandmestiku kasutamine

```
andmed=data.frame(
  haige=rep(c(1,0), c(20, 40)),
T1=c(4.1, 4.8, 3.5, 7.0, 4.6, 5.5, 2.1, 5.1, 3.1, 5.3,
     4.8, 4.0, 5.7, 3.7, 3.8, 3.4, 3.7, 5.3, 4.8, 5.4,
     2.6, 3.4, 3.2, 1.6, 4.8, 3.1, 3.8, 4.0, 2.9, 2.7,
     3.9, 2.0, 5.0, 2.6, 4.7, 4.5, 3.1, 3.6, 2.0, 3.3,
     4.0, 3.1, 2.5, 2.3, 3.0, 4.1, 2.5, 2.9, 1.7, 3.5,
     4.3, 4.5, 3.8, 1.1, 3.5, 3.5, 2.6, 3.2, 2.1, 3.7),
T2=c(223, 216, 221, 213, 209, 201, 226, 210, 203, 219,
     221, 229, 204, 206, 206, 206, 206, 207, 224, 203,
     196, 207, 211, 192, 195, 205, 210, 197, 186, 217,
     214, 193, 199, 182, 206, 216, 184, 192, 194, 193,
     180, 205, 185, 200, 206, 198, 209, 200, 194, 206,
     196, 218, 200, 209, 202, 170, 186, 196, 202, 177),
T3=c(54.9, 61.3, 61.1, 65.2, 63.4, 65.9, 63.3, 66.5,
     63.5, 69.8, 62.1, 66.2, 61.5, 65.8, 64.8, 58.3,
     64.4, 69.7, 67.7, 65.6, 65.2, 74.7, 70.3, 68.2,
     76.6, 66.9, 71.2, 75.1, 72.4, 68.8, 68.3, 71.4,
     72.1, 70.1, 75.3, 69.7, 67.2, 72.1, 61.2, 71.6,
     72.0, 69.1, 65.5, 68.4, 76.2, 67.0, 60.7, 73.2,
     68.0, 68.0, 65.1, 69.9, 72.8, 67.7, 67.6, 74.4,
     69.0, 69.4, 67.5, 73.6)
)


set.seed(1)
ind=rbinom(length(haige), 1, 0.75)
treening=andmed[ind==1,]
test=andmed[ind!=1,]
m=glm(haige~T1+T2+T3, data=treening, family=binomial())
summary(m)

library(pROC)
plot(roc(test$haige, predict(m, newdata=test)), print.auc=TRUE)
```

Näide: Ristvalideerimise kasutamine (samad andmed mis eelmises näites):

```
n=dim(andmed)[1]
prognoos=rep(NA, n)
for (i in 1:n){
  m=glm(haige~T1+T2+T3, data=andmed[-i,],
        family=binomial())
  prognoos[i]=predict(m, newdata=andmed[i,],
        type="response")
}

abi=roc(andmed$haige, prognoos)

plot(1-abi$specificities, abi$sensitivities, type="s", lwd=2,
   xlim=c(0,1), ylim=c(0,1),
   xlab="1-Spetsiifilisus", ylab="Tundlikkus")
text(0.75, 0.25, paste("AUC=", abi$auc))
```

Ülesanne

Verenäitaja X väärtused haigetel ja tervetel olid järgmised:

```
Haiged:  12, 18, 22, 31
Terved:  26, 30, 38
```

Milline on tundlikkus ja spetsiifilisus siis, kui loeme haigeteks kõik need, kel verenäitaja X väärtus on väiksem kui 25?

Joonista ROC-kõver ja leia AUC-väärtus.

bibliograafia

[1] CDC. *Definitions of Signs, Symptoms, and Conditions of Ill Travelers*. 2017. URL: https://www.cdc.gov/quarantine/maritime/definitions-signs-symptoms-conditions-ill-travelers.html (vaadatud 30.04.2020).

[2] Junyan Qu *et al.* "Evaluation of procalcitonin, C-reactive protein, interleukin-6 & serum amyloid A as diagnostic biomarkers of bacterial infection in febrile patients". *The Indian journal of medical research* 141.3 (2015), lk. 315.

[3] Virtuaalkliinik. *Palavik*. 2016. URL: https://www.virtuaalkliinik.ee/haigusteave/2016/05/31/palavik (vaadatud 30.04.2020).

[4] Wikipedia. *Fever*. 2020. URL: https://en.wikipedia.org/wiki/Fever (vaadatud 30.04.2020).