eman ta zabal zazu
UPV EHU
ZIENTZIA ETA TEKNOLOGIA FAKULTATEA
FACULTAD DE CIENCIA Y TECNOLOGÍA
50 URTE AÑOS 1968 - 2018
Biba Zientzia! Ciencia Viva

Denbora-menpeko aldagaien modelizazioa eta haien diskriminazio-gaitasunaren ebaluazioa biziraupen ereduetan

Gradu Amaierako Lana
Matematikako Gradua

Antia Enriquez Yurrebaso

Irantzu Barrio Beraza
Irakasleak zuzendutako lana

Leioa, 2023ko ekainaren 22a

Aurkibidea

Sarrera

Gaur egun, eredu aurresaleen erabilera gero eta indar handiagoa hartzen ari da erabakiak hartzeko momentuan. Eredu hauek garatzeko erabili beharreko metodologia, jasotako datuen ezaugarrien eta ikerketaren helburuaren araberakoa da. Ikerketaren interesa gertaera jakin bat (makina baten akatsa, paziente bat hiltzea, gaixotasun baten gertakari berri bat izatea, etab.) eman arte igarotako denbora iragartzean oinarritzen denean, biziraupen ereduak erabiltzen dira. Hauen artean Cox-en arrisku proportzionalen eredua dago [1, 2]. Hori horrela, eredu honen helburua biziraupena aurresatea da denbora-menpekoak diren edo ez diren aldagai aske ezberdinez baliatuz.

Bestalde, eredu aurresaleak erabakiak hartzeko erabiltzen direnez, oso garrantzitsua da euren diskriminazio edo aurresateko gaitasuna altua izatea. Hau da, edozein denbora unetan gertaera izan duten eta izan ez duten indibiduoen arteko bereizketak egiteko gai izan behar dira. Hori dela eta, behar-beharrezkoa da ereduen diskriminazio-gaitasuna neurtzeko egokiak diren estimatzaileak izatea.

Biziraupen eredu batean denbora-menpeko aldagaiak ez daudenean, C-indexa erabiltzen da, biziraupen ereduentzako *Receiver Operating Characteristic* (ROC) kurbaren azpiko azaleraren (*area under the ROC curve*) (AUC)-aren orokorpena dena [3]. Hala ere, parametro orokorra da, denbora-menpekoa ez dena. Eredu batek denbora-menpeko aldagaiak dituenean indibiduo bakoitzaren estimazioa denboraren menpekoa izango da eta beraz, ereduaren diskriminazio-gaitasuna ere denborarekin aldatzen joango da. Hori dela eta, literaturan denboraren menpekotasuna kontuan hartzen duten diskriminazio parametro ezberdinak proposatu izan dira.

Lan honek bi helburu nagusi ditu. Alde batetik, biziraupen ereduetan denbora-menpeko aldagaiak erabiltzeko metodo ezberdinak aztertu dira eta bestetik, biziraupen ereduetako diskriminazio-gaitasunaren inguruko literaturaren berrikusketa sakona egin da. Denbora-menpeko aldagaiak dituen biziraupen eredu batean diskriminazio-gaitasuna neurtzeko existitzen diren parametro eta hauen estimatzaile ezberdinen abantailak eta desabantailak aztertu ahal izateko, parametro hori denbora-menpekoa edo ez izanik.

Horretarako, simulazio bidezko ikerketa bat egin da. Egoera ezberdinak planteatu dira eta hauetako bakoitzean, diskriminazio-gaitasuna neurtzeko proposatu izan diren parametroak estimatu dira estimatzaile ezberdinak erabiliz. Horren ostean, lortutako emaitzak balio teorikoarekin konparatu dira.

Horrez gain, lanean zehar landutako guztia datu-base erreal bati aplikatu diogu, zehazki, Galdakao-Usansoloko Unibertsitate Ospitalean burutu den biriketako gaixotasun buxatzaile kronikoaren (BGBK) ikerketa bateko datuak dira.

Lanaren zati praktikoa R *software* estatistikoarekin burutu da.

Lan honek hurrengo egitura dauka: Lehenengo kapituluan biziraupen ereduen informazio orokorra zehazten da. Bigarrenean, Cox-en arrisku proportzionalen eredua aurkezten da, ereduaren ezaugarri eta propietateak azalduz. Hirugarren kapituluan, biziraupen ereduetan eta bereziki Cox-en arrisku proportzionalen ereduetan diskriminazio-gaitasuna neurtzeko erabiltzen diren parametro eta hauek estimatzeko estimatzaile ezberdinak deskribatzen dira. Hurrengo kapituluan, simulazio bidezko azterketa bat egiten da eta lortutako emaitzak azaltzen dira. 5. kapituluan, landutakoa praktikara eramaten da, datu-base erreal batean aurreko kapituluetan landutako kontzeptuak aplikatuz. Azken kapituluan bildutako informazio guztia erabiliz ateratako ondorioak irakur daitezke. Horren ondoren 3 eranskin daude: A. eranskinean BGBK-ri buruzko informazioa aurkitu ahal da, B. eranskinean biziraupen-denborak simulatzeko erabilitako formula azaltzen da eta azken eranskinean lana burutzeko garatu den R-ko kodea dago. Amaitzeko, lana egiteko erabilitako bibliografia zehaztuta dago.

1. Kapitulua

Biziraupen ereduak

1.1 Sarrera

Biziraupen ereduek biziraupen-denboren eta horietan eragina duten faktoreen azterketa egiten dute. Eredu mota hauek hainbat kasuistika ezberdinetan aplikatzen dira: saiakuntza klinikoetan, animaliekin egiten diren esperimentuetan, ingeniaritza arloan porrot-denborak aztertzeko, besteak beste. Biziraupen-denboren adibide dira jaiotzetik heriotzara arteko denbora, saiakuntza kliniko batean pazientea ikerketan sartzen denetik heriotza edo gaixotasunaren progresioa eman den arteko denbora. Aipagarria da analisiaren amaiera gertaera positibo batek ere sorraraz dezakela. Adibidez, saiakuntza kliniko batean sartzen denetik gaixotasunari erantzuten zaion arte [4].

Horrez gain, biziraupen ereduetako erantzun aldagaia zorizko aldagai diskretu edo jarraitu ez-negatiboa da eta azterketa hasi den unetik gertaera eman den arteko denbora adierazten du. Eredu hauen beste ezaugarri garrantzitsu bat zentsura da, zehazki, eskumako zentsura; indibiduo baten gertaera behatu ez denean. Formalki, T biziraupen-denboraren (gertaera eman den arteko denboraren) zorizko aldagaia bada eta U zentsura eman deneko denboraren zorizko aldagaia izanik, $T^* = \min(T,\ U)$ behatutako denbora izango da. Gertaera-adierazle bezala $\delta = I(T < U)$ definituko dugu non $I(\cdot)$ funtzio adierazlea den.

Notazioa.
$\delta_i = 0 \Leftrightarrow i$. indibiduoaren gertaera T^* unean zentsuratua izan da $\Rightarrow T > T^*$
$\delta_i = 1 \Leftrightarrow i$. indibiduoaren gertaera T^* unean eman da $\Rightarrow T = T^*$

Lan honetan zehar eskumako zentsura, zehazki, ausazko eskumako zentsura duten datuekin lan egingo da. Gehienetan, zentsura agertzen den bakoitzean emaitza alboratuak lortzen dira, izan ere, askotan zentsura ez da ausaz ematen. Esate baterako, medikuntza arloan ausazko zentsuraren arrazoietako

bat pazientea saiakuntza klinikotik ateratzea da. Saiakuntzaren uztea benetan ausaz gertatzen bada, eta gaixotasunaren prozesuarekin zerikusirik ez badu, zentsura horrek ez du arazorik sortuko analisian. Baina, heriotzatik hurbil dauden gaixoek beste gaixoek baino aukera gehiago badute saiakuntza uzteko, ereduaren estimazioetan alborapen handiak sor daitezke [4]. Hau jakinda, zentsura zergatik eman den aztertu beharko da eta deskribatuko diren estimatzaile guztiek zentsura kontuan hartu beharko dute, lortutako emaitzak egokiak direla ziurtatzeko. Hemendik aurrera, lanean zehar ausazko zentsura kontsideratuko da.

1.2 Oinarrizko definizioak

Jarraian, biziraupen eredu bat doitu ahal izateko, ezinbestekoak diren kontzeptu batzuk definituko dira, [4]:

1.2.1. definizioa. *Biziraupen funtzioa*k t puntura arte bizirauteko probabilitatea zehazten du eta hurrengo eran definitzen da:

$$S(t) = P(T > t), \quad 0 \leq t < \infty$$

Probabilitate bat denez, argi dago funtzio honen balio maximoa 1 balioa izango dela, $t = 0$ unean lortzen dena, eta inoiz ez duela 0 baino baxuagoa den balioa izango. Gainera, eskuinetik jarraitua den funtzioa da.

1.2.2. definizioa. *Arrisku funtzioa*, une zehatz bateko arrisku-tasa dena, hurrengo ideian oinarritzen da: suposatuta indibiduo batek t unera arte biziraun duela indibiduo horrek zein probabilitate duen hurrengo denbora tarte txikian gertaera jazotzeko, tarte horren luzeraz zatitua. Matematikoki honela idatz daiteke:

$$h(t) = \lim_{\delta \to 0} \frac{P(t < T < t + \delta \mid T > t)}{\delta}$$

1.2.3. definizioa. *Metatutako banaketa funtzioa* ondorengo eran adierazi ahal da:

$$F(t) = P(T \leq t), \quad 0 \leq t < \infty$$

Ohartu funtzio hau biziraupen funtzioaren osagarria dela eta ondorioz, $F(t) + S(t) = 1, \ \forall t \in [0, \infty)$. Kasu honetan ere, eskuinetik jarraitua da.

1.2.4. definizioa. *Probabilitate dentsitate funtzioa* ondokoa da:

$$f(t) = -\frac{d}{dt} S(t) = \frac{d}{dt} F(t)$$

Ohartu t uneko arrisku funtzioa t denboraren inguruan gertaera bat emateko probabilitatea dela, indibiduoa t unean bizirik egoteko probabilitatearekin

zatituta. Formalki:

$$h(t) = \frac{f(t)}{S(t)}$$

1.2.5. definizioa. *Metatutako arrisku funtzioa*, 0tik t puntura arte arrisku funtzioaren azpiko azalera da:

$$H(t) = \int_0^t h(u)du$$

Aurreko guztia kontuan hartuz biziraupen funtzioa eta arrisku funtzioa erlazionatuta daudela ondoriozta dezakegu, izan ere, $S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}$ da.

1.3 Biziraupen banaketak

Biziraupen ereduak doitzeko erabiltzen diren metodoak biziraupen banaketaren araberakoak dira. Biziraupen banaketa zehazteko gehienetan arrisku funtzioa edo biziraupen funtzioa erabiltzen den arren, metatutako banaketa funtzioa ere erabili ahal da. Banaketak bi motatakoak izan daitezke: parametrikoak eta ez-parametrikoak.

Biziraupen banaketa parametrikoen artean gehien erabiltzen direnak banaketa esponentziala, gamma banaketa edo Weibull banaketa dira. Sinpleena banaketa esponentziala da eta bertan arrisku funtzioa konstante bat da, zehazki, $h(t) = \lambda$. Aurretik zehaztuta baldin badago zein izango den probabilitate banaketa eta zeintzuk diren parametroen balioak oso erreza da indibiduo bakoitzaren biziraupen probabilitatea kalkulatzea. Dena dela, nahiz eta zehaztuta izan zein izango den probabilitate banaketa, gehienetan ez da jakingo zeintzuk diren parametroen balioak eta beraz, estimatu beharko dira *egiantz handieneko metodoa* erabiliz.

Hala ere, gizakien edo animalien biziraupena modelizatzeko, normalean zaila da jakitea zein familia parametriko aukeratu behar den eta, askotan, ezaguna den banaketa batek ere ez du behar bezalako malgutasunik datuen benetako forma modelizatzeko. Hori dela eta, medikuntza arloan gehienetan metodo ez-parametrikoak erabiltzen dira, malgutasun handiagoa ematen baitute.

Biziraupen funtzioa estimatzeko metodo ez-parametriko ohikoena *Kaplan-Meier*-en (KM) metodoa da. Bertan biziraupen-denborak erabiltzen dira biziraupen funtzioa estimatzeko [4].

Izan bitez n tamainako lagin bat eta $t_1 \leq t_2 \leq \ldots \leq t_n$ indibiduo bakoitzaren behatutako denborak. Orduan, t_i $(i = 1, \ldots, n)$ bakoitzerako d_i eta n_i

definitzen dira non d_i t_i unean gertaera izan duten indibiduo kopurua den eta n_i t_i unean arriskuan dauden indibiduo kopurua den, hau da, une horretara arte intereseko gertaera izan ez dutenak edo zentsuratuak izan ez direnak. Hori horrela, Kaplan-Meier-en estimatzailea hurrengoa da:

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i} \right)$$

Ohartu errekurtsiboki ere definitu ahal dela. Izan ere:

$$\widehat{S}(t_i) = \widehat{S}(t_{i-1}) \left(1 - \frac{d_i}{n_i} \right), \forall i = 2, \ldots, n$$

Funtzio hau urrats-funtzio ez-gorakorra da, gertaera bat ematen den bakoitzean balioz aldatzen dena eta eskuinetik jarraitua da. Hori dela eta, Kaplan-Meier-en metodoak aukera ematen du biziraupen funtzioa grafikoki irudikatzeko. Praktikan, datu-basean dauden aldagai azaltzaileek ematen duten informaziotik abiatuz, indibiduoen talde ezberdinak sortzen dira eta talde bakoitzerako biziraupen funtzioa estimatzen da.

Aurretik azaldu den Kaplan-Meier-en metodoak aukera ematen du, modu sinple batean talde ezberdinen arteko biziraupen funtzioak konparatzeko. Esate baterako, tratamendu ezberdinak jaso dituzten pazienteen arteko konparaketa egiteko. Baina praktikan, helburua gehienetan aldagai azaltzaile bat baino gehiagoren menpe dagoen biziraupen funtzio bat estimatzea da, aldagai hauek kategorikoak zein jarraituak izanik. Hori estimatzeko maizen erabiltzen den eredua *Cox-en arrisku proportzionalen* erregresio eredua da.

2. Kapitulua

Cox-en arrisku proportzionalen eredua

Cox-en arrisku proportzionalen (CPH) eredua ikerketa ezberdinetan erabili ohi den biziraupen eredua da non indibiduoen biziraupen-denboren eta aldagai aske bat edo gehiagoren arteko lotura ikertzen den. Eredu honen helburua aldagai askeek biziraupenean duten eragina aldi berean ebaluatzea da.

2.1 Ereduaren doikuntza

Normalean, CPH eredua arrisku funtzioaren bidez adierazten da, hau da, $h(t)$ bidez eta indibiduo bakoitzaren intereseko gertaera t une zehatzean izateko arriskua zehazten du. Ereduaren itxura honako hau da, [5]:

$$h(t|\mathbf{X}) = h_0(t)e^{\sum_{k=1}^{p}\beta_k X_k} = h_0(t)e^{\boldsymbol{\beta}^t\mathbf{X}} \tag{2.1}$$

non $\mathbf{X} = (X_1,\ X_2, \ldots, X_p)$ aldagai askeen bektorea, $h_0(t)$ *oinarrizko arrisku funtzioa* eta $\boldsymbol{\beta} = (\beta_1,\ \beta_2, \ldots, \beta_p)$ koefizienteen bektorea diren.

Eredu honen ezaugarrietako bat oinarrizko arrisku funtzioa, $h_0(t)$, funtzio zehaztugabea dela da. Hori dela eta, Cox-en arrisku proportzionalen eredua eredu erdiparametrikoa da.

Eredu honek duen propietate garrantzitsuenetariko bat, arriskuak proportzionalak direlaren hipotesia kontuan hartzen duela da. Hau da, bi indibiduoen arteko arriskua denboran zehar konstante mantentzen dela.

Horrez gain, formula honen ezaugarri garrantzitsu bat, arrisku proportzionalen hipotesiari dagokiona, oinarrizko arrisku funtzioa t-ren menpekoa dela da, baina ez duela $\mathbf{X}$ kontuan hartzen. Aldiz, 2.1 ekuazioan agertzen den

adierazpen esponentziala **X**-ren araberakoa da baina ez da denboraren menpekoa (t-ren menpekoa). Hala ere, posiblea da denbora-menpeko aldagai askeen bektore bat kontsideratzea. Kasu horretan ere 2.1 formula erabili ahalko da eredua deskribatzeko baina, eredu hori ez da arrisku proportzionalen eredua izango *Cox-en eredu hedatua* baizik. Bestalde, ziurtatu behar da arrisku proportzionalen hipotesia betetzen dela. Ikusi 2.4 eta 2.5 atalak.

2.2 Ereduaren estimazioa

Ereduaren koefizienteen ($\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$) estimazioa lortzeko, *egiantz handieneko* metodoa erabiltzen da eta estimatutako koefiziente bakoitza $\hat{\beta}_k,\ k \in \{1,\ 2, \ldots, p\}$ bidez denotatuko da [5].

Erregresio logistikoan, egiantz handieneko metodoak parametroen estimazioak lortzen ditu egiantz-funtzioa erabiliz. Cox-en arrisku proportzionalen ereduan, aldiz, egiantz-funtzio partziala erabiltzen da.

Izan bitez n tamainako lagin bat eta $t_1, \ldots, t_n$ indibiduo bakoitzaren behatutako denborak. CPH ereduarekin lan egiten gabiltzanez, j. indibiduoaren arrisku funtzioa t_i denboran $h_j(t_i) = h_0(t_i)e^{\sum_{k=1}^{p} \beta_k x_{k,j}}$ izango da non $x_{k,j}$-k k. aldagai azaltzailean j. indibiduoak hartzen duen balioa adierazten duen. $R(t_i)$ t_i denboran arriskuan dauden indibiduoen multzoa eta δ_i i. indibiduoaren gertaera-adierazlea izanik, egiantz-funtzio partziala hurrengoa da:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{h_0(t_i)e^{\sum_{k=1}^{p} \beta_k x_{k,i}}}{\sum_{j \in R(t_i)} h_0(t_i)e^{\sum_{k=1}^{p} \beta_k x_{k,j}}} \right)^{\delta_i} =$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\sum_{k=1}^{p} \beta_k x_{k,i}}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \beta_k x_{k,j}}} \right)^{\delta_i} = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}^t \mathbf{x}_i}}{\sum_{j \in R(t_i)} e^{\boldsymbol{\beta}^t \mathbf{x}_j}} \right)^{\delta_i}$$

Beraz, egiantz-funtzio partzialak gertaera izan duten indibiduoen behatutako denborak aintzat hartzen ditu baina ez ditu esplizituki zentsuratutako indibiduoen behatutako denborak kontsideratzen. Funtsean, gertaera eman den une bakoitzean gertaera izateko probabilitateen biderkadura da.

Euskarri-funtzio partziala $L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})]$ bezala definituko da. $L(\boldsymbol{\beta})$ maximizatzen duen $\boldsymbol{\beta}$ lortzeko, euskarri-funtzio partziala $\beta_k,\ k = 1, 2, ..., p$ bakoitzeko deribatuko da eta bertatik estimatutako koefizienteak ($\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)$) lortuko dira. Ekuazio sistema ebazteko, *Newton–Raphson* edo *Expectation-Maximitazion* algoritmoak erabiltzen dira [6].

Behin koefizienteen estimazioak lortuta, test estatistiko ezberdinetaz baliatuz haien adierazgarritasuna aztertu behar da. Test erabilienak Wald-en testa eta egiantz arrazoiaren testa dira.

2.2.1 Wald-en testa

Test honek $\hat{\beta}_k$, $\forall k = 1, \ldots, p$ bakoitzari ondorengo hipotesi kontrastea aplikatzen dio:

$$\begin{cases} H_0: & \beta_k = 0 \\ H_1: & \beta_k \neq 0 \end{cases}$$

Test estatistikoa $W = \dfrac{\hat{\beta}_k - \beta_k}{\mathrm{sd}(\hat{\beta}_k)}$ da, sd desbideratze estandarra izanik eta banaketa normal estandarizatuari darraio. H_0 egia dela suposatuz, estatistikoaren balioa $w = \dfrac{\hat{\beta}_k}{\mathrm{sd}(\hat{\beta}_k)}$ da eta hipotesi kontrasteari dagokion p-balioa $p = 2 \cdot P(Z > |w|)$ da [4].

$(1-\alpha) \cdot 100\%$-eko konfiantza-tartea ere eraiki ahal da:

$$I_{\beta_k}^{1-\alpha} = (\hat{\beta}_k - z_{\alpha/2} \cdot \hat{\mathrm{sd}}(\hat{\beta}_k),\ \hat{\beta}_k + z_{\alpha/2} \cdot \hat{\mathrm{sd}}(\hat{\beta}_k))$$

2.2.2 Egiantz arrazoiaren testa

Test hau erabiltzen da q aldagai askeen adierazgarritasuna aztertzeko, $q \leq p$ izanik. Aplikatzen den hipotesi kontrastea hurrengoa da:

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \ldots = \beta_q = 0 & (\omega \text{ eredua}) \\ H_1: & \exists k \in \{1, 2, \ldots, q\} : \beta_k \neq 0 & (\Omega \text{ eredua}) \end{cases}$$

Estatistikoa $G = -2 \ln \dfrac{\omega \text{ ereduaren egiantz partziala}}{\Omega \text{ ereduaren egiantz partziala}}$ da χ_q^2 banaketari darraiona. Hipotesi kontrasteari dagokion p-balioa $p = P(\chi_q^2 > G)$ da [6].

Oharra. Koefiziente bakar baten adierazgarritasuna neurtu nahi bada, bi testekin aplikatzen den hipotesi kontrastea berdina izan arren, lortuko diren emaitzak ez dira beti berdinak izango, estatistikoa aldatzen baita. Egiantz arrazoiaren testak ahalmen handiagoa duenez, test horretako emaitzak hobetsiko dira [5].

2.3 Hazard ratioa

CPH ereduaren helburu nagusietako bat aldagai askeen balio ezberdinak dituzten indibiduoen arrisku tasak alderatzea da. Biziraupen absolutua estimatu beharrean taldeen arteko konparaketak egin nahi badira, *hazard ratioa* (HR) erabiltzen da.

Hazard ratioa bi indibiduoen arriskuen arteko zatiketa da non indibiduo bakoitzak aldagai askeen multzo ezberdina duen. HR hurrengo eran estimatu ahal da, [6]:

$$\widehat{\mathrm{HR}} = \frac{h_i(t|\mathbf{x}_i)}{h_j(t|\mathbf{x}_j)} = \frac{h_0(t)e^{\sum_{k=1}^{p}\hat{\beta}_k x_{k,i}}}{h_0(t)e^{\sum_{k=1}^{p}\hat{\beta}_k x_{k,j}}} = e^{\sum_{k=1}^{p}\hat{\beta}_k(x_{k,i}-x_{k,j})} = e^{\hat{\boldsymbol{\beta}}^t(\mathbf{x}_i-\mathbf{x}_j)}$$

non $\mathbf{x}_i$ eta $\mathbf{x}_j$ $i.$ eta $j.$ indibiduoen aldagai askeen multzoak diren, hurrenez hurren.

$(1-\alpha)\cdot 100\%$-eko konfiantza-tartea ere eraiki ahal da:

$$I_{\mathrm{HR}}^{1-\alpha} = (e^{\hat{\boldsymbol{\beta}}^t(\mathbf{x}_i-\mathbf{x}_j)} - z_{\alpha/2}\cdot\hat{\mathrm{sd}}(\hat{\boldsymbol{\beta}}^t(\mathbf{x}_i-\mathbf{x}_j)),\ e^{\hat{\boldsymbol{\beta}}^t(\mathbf{x}_i-\mathbf{x}_j)} + z_{\alpha/2}\cdot\hat{\mathrm{sd}}(\hat{\boldsymbol{\beta}}^t(\mathbf{x}_i-\mathbf{x}_j)))$$

Oharra. Aldagai azaltzailea dikotomikoa baldin bada, aldagaiaren balioak 1 edo 0 izanik, HR ondorengoa da:

$$\widehat{\mathrm{HR}} = e^{\hat{\beta}(x_i-x_j)} = e^{\hat{\beta}(1-0)} = e^{\hat{\beta}}$$

HR-ak arriskua neurtzen du eta bere balioa 1 baino handiagoa bada, arriskuaren gehikuntza adieraziko du. Aldiz, $\mathrm{HR} < 1$ denean, arriskuaren murrizketa dagoela esan ahalko da eta $\mathrm{HR} = 1$ bada, ez dagoela efekturik.

Amaitzeko, esan beharra dago arrisku proportzionalen baldintzak HR denboran zehar mantenduko dela inplikatzen duela, hau da, indibiduo baten arriskua beste edozein indibiduoren arriskuaren proportzionala da eta proportzionaltasuna ez da denboraren menpe aldatzen.

2.4 Denbora-menpeko aldagaiak

CPH ereduak aukera ematen du indibiduoen biziraupen-denborak estimatzeko aldagai askeen informazioa erabiliz. Kontuan hartu behar da eredu honetan aldagai askeen balioak $t = 0$ unean zehaztu behar direla, indibiduoa azterketan sartzen denean, eta ordutik aurrera konstante mantendu. Dena dela, biziraupen datuak ditugunean askotan gertatzen da aldagai azaltzaileen balioak denboran zehar aldatzen joaten direla eta beraz, eredu bat doitzerako orduan informazio gehiago erabiltzeko aukera izanda hobe litzateke jasotako informazio guztiaz baliatzea. Kasu hauetan Cox-en arrisku proportzionalen ereduan denbora-menpeko aldagaiak erabiltzen dira aldagai askeen balio ezberdinak kontuan hartu ahal izateko.

Denbora-menpeko aldagaiak erabiltzen direnean, Cox-en eredu hedatu baten aurrean gaude, denbora-menpeko CPH eredua ere deitu ahal zaio. Normalean, eredu mota hau arrisku funtzioaren bidez zehazten da.

Izan bitez $\mathbf{X}(t) = (X_1(t),\ X_2(t), \ldots, X_p(t))$ denbora-menpeko aldagai askeen bektorea, $\boldsymbol{\beta} = (\beta_1,\ \beta_2, \ldots, \beta_p)$ koefizienteen bektorea eta $h_0(t)$ oinarrizko arrisku funtzioa. Ereduaren arrisku funtzioak honako itxura hau du:

$$h(t|\mathbf{X}(t)) = h_0(t)e^{\sum_{k=1}^{p}\beta_k X_k(t)} = h_0(t)e^{\boldsymbol{\beta}^t\mathbf{X}(t)} \tag{2.2}$$

Oharra. $\mathbf{X}(t)$-n denbora-menpekoak ez diren aldagaiak egon ahal dira. Kasu horietan haien balioa konstantea izango da azterketa-denbora osoan zehar.

Koefizienteen estimazioa lortzeko, oraingoan ere egiantz handieneko metodoa erabiltzen da, baina kasu honetan behatutako denbora bakoitzean aldagai askeen balioa aldatzen joango da. Egiantz-funtzio partzialaren itxura hurrengoa da:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{e^{\sum_{k=1}^{p}\beta_k x_{k,i}(t_i)}}{\sum_{j\in R(t_i)} e^{\sum_{k=1}^{p}\beta_k x_{k,j}(t_i)}}\right)^{\delta_i} = \prod_{i=1}^{n}\left(\frac{e^{\boldsymbol{\beta}^t\mathbf{x}_i(t_i)}}{\sum_{j\in R(t_i)} e^{\boldsymbol{\beta}^t\mathbf{x}_j(t_i)}}\right)^{\delta_i}$$

non $R(t_i)$ t_i denboran arriskuan dauden indibiduoen multzoa eta δ_i i. indibiduoaren gertaera-adierazlea diren.

2.5 Arrisku proportzionalen hipotesia

CPH eredu bat doitzen denean arrisku proportzionalen hipotesia betetzen dela asumitzen da, hala ere, behin eredua doitua dagoenean ereduaren hondarren azterketa egin behar da hipotesia betetzen dela ziurtatzeko. Gehienetan Schoenfeld-en hondarrak erabiltzen dira eta aplikatzen den testaren estatistikoak, ρ, koefizienteen eta biziraupen denboraren arteko korrelazioa aztertzen du. Gainera, Khi-karratu banaketari darraion estatistikoa da [7].

Hipotesia ez denean betetzen, aldagai askeen efektua denbora-menpekoa dela esaten da. Arazo horri aurre egiteko aukera ezberdinak daude: CPH ereduan denbora-menpeko koefizienteak erabiltzea, arrisku proportzionalen hipotesia betetzen ez duten aldagaiak jarraituak badira aldagaiak kategorizatzea edo erregresio *spline*-ak erabiltzea.

2.5.1 Denbora-menpeko koefizienteak

CPH ereduetan denbora-menpeko koefizienteak erabiltzen direnean, kasu honetan ere, Cox-en eredu hedatu baten aurrean egongo gara.

Eredu honen arrisku funtzioak honako itxura du:

$$h(t|\mathbf{X}) = h_0(t)e^{\sum_{k=1}^{p}\beta_k(t)X_k} = h_0(t)e^{\boldsymbol{\beta}^t(t)\mathbf{X}} \tag{2.3}$$

non $\mathbf{X} = (X_1,\ X_2, \ldots, X_p)$ aldagai askeen bektorea, $h_0(t)$ oinarrizko arrisku funtzioa eta $\boldsymbol{\beta}(t) = (\beta_1(t),\ \beta_2(t), \ldots, \beta_p(t))$ denbora-menpeko koefizienteen bektorea diren.

Denbora-menpeko koefizienteak deskribatzeko metodo erabiliena aurretik zehazturiko funtzio parametriko jarraitu bat erabiltzea da. Hau da, $\beta(t) = g(t)$ izatea, g funtzio parametriko jarraitua izanik. Maizen erabiltzen den funtzioa $g(t) = a + b\log(t)$ da, modu horretan hazard ratioak forma interpretagarria duelako. Hala ere, batzuetan denbora eskala aldatzen da $g(t) = a + b\log(t + k)$, $k > 0$ funtzioa erabiliz ikerketa egin den egun guztien linealtasuna bermatzeko [8]. Behin funtzioa zehaztuta CPH eredua ohiko eran doitzen da.

Koefizienteen estimazioa egiteko kasu honetan ere egiantz handieneko metodoa erabiltzen da. Egiantz-funtzio partzialaren itxura honako hau da:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\sum_{k=1}^{p} \beta_k(t_i) x_{k,i}}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \beta_k(t_i) x_{k,j}}} \right)^{\delta_i}$$

non $R(t_i)$ t_i denboran arriskuan dauden indibiduoen multzoa eta δ_i i. indibiduoaren gertaera-adierazlea diren.

Zehazki $\beta(t) = g(t) = a + b\log(t)$ bada, egiantz-funtzio partziala ondorengoa izango da:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(\frac{e^{\sum_{k=1}^{p} \beta_k(t_i) x_{k,i}}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \beta_k(t_i) x_{k,j}}} \right)^{\delta_i} = \\ &= \prod_{i=1}^{n} \left(\frac{e^{\sum_{k=1}^{p} (a_k + b_k \log(t_i)) x_{k,i}}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} (a_k + b_k \log(t_i)) x_{k,j}}} \right)^{\delta_i} \end{aligned}$$

Egiantz handieneko metodoa aplikatzen da $\widehat{a}_k$ eta $\widehat{b}_k$, $\forall k \in \{1, \ldots, p\}$ estimatzeko.

Oharra. Aukera dago ere eredu berean denbora-menpeko aldagaiak eta koefizienteak egotea [7]. Eredu horiek doitzeko erabili beharreko arrisku funtzioaren itxura hurrengoa da:

$$h(t|\mathbf{X}(t)) = h_0(t) e^{\sum_{k=1}^{p} \beta_k(t) X_k(t)} \tag{2.4}$$

non $\mathbf{X}(t) = (X_1(t),\ X_2(t), \ldots, X_p(t))$ denbora-menpeko aldagai askeen bektorea, $h_0(t)$ oinarrizko arrisku funtzioa eta $\boldsymbol{\beta}(t) = (\beta_1(t),\ \beta_2(t), \ldots, \beta_p(t))$ denbora-menpeko koefizienteen bektorea diren.

2.5.2 Erregresio *spline*-ak

Demagun, X aldagai jarraitu batekin Cox-en arrisku proportzionalen eredu bat doitu dela. Orduan, ereduaren itxura honako hau izango da:

$$h(t|X) = h_0(t)e^{\beta X}$$

Eredu honek esaten digu X aldagaiarekiko log hazard ratioa (LHR), X aldagaiaren funtzio lineala dela, izan ere:

$$\text{LHR}(X) = \log\left(\frac{h(t|X)}{h_0(t)}\right) = \log(e^{\beta X}) = \beta X$$

CPH eredu batean arrisku proportzionalen hipotesia betetzen ez denean, LHR(X)-aren jarrera ez-lineala izan ahal da eta beraz, CPH eredua doitzeko orduan malgutasun gehiago behar da. Hori lortzeko, $h(t|X) = h_0(t)e^{f(X)}$ itxura duen CPH eredua doitu daiteke, modu horretan, $\text{LHR}(X) = f(X)$ izango da.

$f(X)$ funtzioa hurbiltzeko aukera sinpleenetariko bat X-ren menpeko d mailako funtzio polinomiko bat kontsideratzea da. Hau da, $f(X) = \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^d$ erabiltzea. Hala ere, gaur egun funtzio leun orokorragoak existitzen dira. Haien artean erregresio polinomiko haztatu lokala, Kernel bidezko leunketak edo *spline*-ak daude.

Spline-ak zatikako funtzio polinomikoak dira zeinetan nodo deritzen loturapuntuetan murrizketak ezartzen zaizkien. 3 elementuen araberakoak dira: polinomioen maila, nodoen kopurua eta nodoen kokapena. Nodoek mugatutako tarte bakoitzean d mailako eta $\mathcal{C}^{d-1}$ klaseko polinomioak doitzen direnean erregresio *spline*-ak lortzen dira.

Nodoen kopurua zehazteko aukera ohikoena nodo kopurua 3 eta 7 bitartean egotea da. Lagin-tamaina handia denean ($n \geq 100$), gehienetan, 5 nodoekin lan egiten da, lagin-tamaina txikia denean ($n \leq 30$), aldiz, 3 nodo zehazten dira. Nodoen kokapena zehazteko askotan X aldagaiaren pertzentilak erabilitzen dira, hala ere, zehaztu ahal da nodo guztien arteko distantzia berdina izatea. Pertzentiletan kokatzea erabakitzen bada, 3 nodo ezarriz gero, 5, 50 eta 95 pertzentiletan kokatzen dira eta 5 nodo badaude, 5, 25, 50, 75 eta 95 pertzentiletan [9].

Ohikoena $d = 3$ zehaztea da, honela, *spline* kubikoekin lan egiten egongo ginateke.

Demagun, m nodo ditugula $c_1 < \ldots < c_m$ izanik eta u zorizko aldagai bat izanik, $(u)_+ = \max\{u, 0\}$ definitzen dugula. $c_1, \ldots, c_m$ nodoak dituen *spline*

kubiko baten itxura honako hau da:

$$C(u) = \beta_0 + \beta_1 u + \sum_{j=1}^{m-2} \theta_j C_j(u)$$

non $C_j(u) = (u - c_j)_+^3 - \frac{(u-c_{m-1})_+^3 (c_m - c_j)}{(c_m - c_{m-1})} + \frac{(u-c_m)_+^3 (c_{m-1} - c_j)}{(c_m - c_{m-1})}$ den.

***Spline* kubikoen propietateak**

(i) Erregresio *spline*-en definiziotik ondoriozta dezakegu $C(u)$-ren lehenengo eta bigarren deribatuak jarraituak direla.

(ii) $C(u)$ funtzio lineala da, m koefiziente dituena: $\beta_0, \beta_1, \theta_1, \ldots, \theta_{m-2}$.

(iii) Orokorrean, badakigu *spline*-ak funtzio leun ezezagun bat hurbiltzeko erabiltzen direla, $f(u)$. Orduan,
$\beta_1 = \theta_1 = \ldots = \theta_{m-2} = 0 \Rightarrow f(u)$ funtzio konstantea.
$\theta_1 = \ldots = \theta_{m-2} = 0 \Rightarrow f(u)$ funtzio lineala.

Aurreko guztia kontuan izanik, CPH eredu batek arrisku proportzionalen hipotesia betetzen ez duenean eta $h(t|X) = h_0(t)e^{f(X)}$ itxura duen CPH eredua doitzea erabakitzen denean, $f(X)$ funtzioa hurbiltzeko *spline* kubikoak erabil daitezke.

Hasteko, kontuan hartu behar da CPH ereduen definizioan oinarrituz, $X = 0$ denean, $h(t|X = 0) = h_0(t)$ izango dela eta ondorioz, $f(0) = 0$ izan behar dela.

Hori horrela, ondoko eran hurbildu ahal da $f(X)$ funtzioa:

$$f(X) \simeq C(X) - C(0) = \beta_1 X + \sum_{j=1}^{m-2} \theta_j (C_j(X) - C_j(0))$$

Oharra. Nodorik txikiena zenbaki ez-negatiboa bada, hau da, $c_1 \geq 0$ bada, $C_j(0) = 0$ izango da $\forall j = 1, \ldots, m-2$ izanik.

Zehaztasun gehiago [9] artikuluan aurki daiteke.

3. Kapitulua

Diskriminazio-gaitasuna

Eredu aurresaleak oso tresna garrantzitsua dira etorkizuneko gertaerak aurreikusi ahal izateko. Eredu hauek erabiltzeko haien diskriminazio-gaitasuna neurtzea ezinbestekoa da. Izan ere, ez du inolako zentsurik eredu bat doitzea bere aurresateko gaitasuna txanpon bat airera botatzea bezain ona bada. Biziraupen ereduetan, zehazki Cox-en arrisku proportzionalen eredu bat doitu denean, aurresateko gaitasuna neurtzeko parametrorik erabiliena *C-index*-a da.

3.1 C-index-a

C-index-a edo *concordance index*-a biziraupen ereduen ebaluazio globala egiteko maizen erabiltzen den parametroa da. AUC-aren hedapena da zentsuradun biziraupen ereduetara [3] eta funtsean behatutako denboren eta iragarritako arriskuen arteko korrelazioa kuantifikatzen du. Zenbat eta balio altuagoak izan, orduan eta diskriminazio-gaitasun handiagoa izango du ereduak [10].

C-index-aren definizio teorikoa eman aurretik kontzeptu batzuen deskribapena eta notazioa zehaztuko da:

Izan bedi (i, j) bi indibiduoz osatutako bikotea.

3.1.1. definizioa. (i, j) bikotea *konparagarria* da indibiduo baten gertaera lehenago jazo dela edo bi gertaerak aldi berean jazo direla esan ahal baldin bada. Hau da, behintzat bat ez da zentsuratua egon behar.

Oharra. Biziraupen eredu bat doitzen denean, denbora zehatz batean indibiduo bakoitzaren intereseko gertaera izateko arriskua zehazten da. Lan osoan zehar i. indibiduoaren arriskua intereseko gertaera izateko $M_i = e^{\boldsymbol{\beta}^t \mathbf{X}_i} = e^{\eta_i}$ izango da, $\eta_i = \boldsymbol{\beta}^t \mathbf{X}_i$ i. indibiduoaren aurresale lineala izanik.

3.1.2. definizioa. (i, j) bikote konparagarria *konkordantea* izango da gertaera lehenago izan duen indibiduoak arrisku altuagoa badu.
Kontrako kasuan, bikote *ez-konkordantea* dela esango dugu.

Aurreko guztia kontuan izanik eta bi indibiduoen arriskuak eta biziraupendenborak zorizko aldagaitzat hartuta, C-index-a hurrengo probabilitate baldintzatu gisa defini dezakegu:

$$C = P(M_j > M_i | T_j < T_i) \tag{3.1}$$

Hau ikusita, esan ahal da C-index-aren arabera eredu on batek beti arrisku handiagoa ematen diela gertaera lehenago izan duten indibiduoei. Gainera, parametro honek hiru dimentsio ezberdin (arriskua, behatutako denbora eta zentsura) zenbaki bakar batean laburtzeko gaitasuna du eta erdietsitako zenbakiaren arabera zehaztu ahal da lortu den eredua ona ala ia ausazkoa den.

C-index-a estimatzeko hainbat estimatzaile ezberdin proposatu dira: Harrell-en C-index-a, Uno-ren C-index-a, Harrell-en C-index-aren bertsio haztatua dena eta Gonen eta Heller-en neurria, besteak beste. Ohikoena Harrell-ena da, interpretazio intuitiboena duelako [11].

3.1.1 Harrell-en C-index-a

Harrell-en estimatzailea bikote konkordante eta konpargarrien arteko proportzioan oinarritzen da C-index-a estimatzeko [12]. Ondorengo formula erabiliz lortzen da:

$$\widehat{C} = \frac{\sum_{i=1}^{n} \delta_i \sum_{j=i+1}^{n} [I(T_i^* < T_j^*) + (1-\delta_j) I(T_i^* = T_j^*)][I(\widehat{M}_i > \widehat{M}_j) + \frac{1}{2} I(\widehat{M}_i = \widehat{M}_j)]}{\sum_{i=1}^{n} \delta_i \sum_{j=i+1}^{n} [I(T_i^* < T_j^*) + (1-\delta_j) I(T_i^* = T_j^*)]} \tag{3.2}$$

non $\widehat{M} = e^{\sum_{k=1}^{p} \hat{\beta}_k x_k}$ estimatutako arriskua eta $I(\cdot)$ funtzio adierazlea diren.

Gainera, $T_i^* = T_j^*$ eta $\widehat{M}_i = \widehat{M}_j$ ezinezkoak direlaren hipotesia kontuan hartzen bada, estimatzailearen bertsio sinpleago bat lortzen da:

$$\widehat{C} = \frac{\sum_{i=1}^{n} \delta_i \sum_{j=i+1}^{n} I(T_i^* < T_j^*) I(\widehat{M}_i > \widehat{M}_j)}{\sum_{i=1}^{n} \delta_i \sum_{j=i+1}^{n} I(T_i^* < T_j^*)} \tag{3.3}$$

Estimatzailearen propietateak

3.3 ekuazioko estimatzailean oinarrituz (3.2 ekuaziora hedatzea tribiala da), Harrell-en estimatzailearen bi propietate aztertuko dira.

Alde batetik, (i, j) bikote konparagarri guztiak konkordanteak badira, hau da, konkordantzia perfektua ematen bada, $\widehat{C} = 1$ izango da. Aldiz, bikote

konparagarri guztiak ez-konkordanteak badira, anti-konkordantzia, $\widehat{C} = 0$ izango da. Honekin frogatu dugu $\widehat{C}$ probabilitate bat estimatzen dagoela $0 \leq \widehat{C} \leq 1$ betetzen delako.

Bestalde, Harrell-en estimatzaileak eredu batek bi indibiduo ongi ordenatu dituelaren probabilitatea kalkulatzen du indibiduo bakoitzaren arriskua eta behatutako denbora kontuan izanik. Horrez gain, esan ahal da estimatzaile hau zentsura banaketagatik baldintzatua dagoela, hau da, δ_i-gatik baldintzatuta dagoela, baita datu-basearen tamainarengatik ere. Izan ere, behatutako gertaera guztiak denbora finitu bat, τ, baino lehenago ematen dira. Ondorio hauetara heldu ahal izan gara [10] artikuluan $\widehat{C} \to P(M_i > M_j | T_i < T_j,\ \delta_i = 1,\ T_i < \tau)$ konbergentzia frogatzen delako.

3.2 Denbora-menpeko diskriminazio-gaitasuna

Aurreko atalean azaldu den bezala, CPH eredu bat dugunean gehien erabiltzen den parametroa ereduaren aurresateko gaitasuna neurtzeko C-index-a da. Hala ere, denbora-menpeko aldagaiak ditugunean C-index-a erabiltzea ez da guztiz egokia, denbora-menpekoa ez den parametroa delako eta beraz, beste parametro batzuk erabiltzen dira denboraren menpeko aurresateko gaitasuna neurtzeko.

Badakigu biziraupen eredu baten helburua analisia egin den une guztietarako gertaerak eta ez-gertaerak ahalik eta zehaztasun handienarekin bereiztea dela. Ideia horretan oinarrituta, erregresio logistikoan erabiltzen diren espezifikotasunaren, sentikortasunaren eta AUC-aren definizioak biziraupen ereduetara heda daitezke.

Mozketa-puntu finko baterako (c), *sentikortasuna* egiazko positiboen arrazoia da (TPF) eta *espezifikotasuna*, ordea, egiazko negatiboen arrazoia. Ondorioz, 1− espezifikotasuna, positibo faltsuen arrazoia da (FPF). Matematikoki:

$$\text{sentikortasuna}(c) = P(M > c|\ \text{gertaera}) \Rightarrow TPF(c) = P(M > c|\ \text{gertaera})$$

$$\text{espezifikotasuna}(c) = P(M \leq c|\ \text{ez-gertaera}) \Rightarrow$$
$$\Rightarrow\ FPF(c) = P(M > c|\ \text{ez-gertaera})$$

ROC kurba 1−espezifikotasuna eta sentikortasuna elkarrekin irudikatzen dituen grafikoa da:

$$\begin{aligned}\text{ROC}(\cdot) &= \{(FPF(c), TPF(c)) :\ c \in \mathbb{R}^+\} = \\ &= \{(1 - \text{espezifikotasuna}(c),\ \text{sentikortasuna}(c)) :\ c \in \mathbb{R}^+\}\end{aligned}$$

ROC kurba definitzeko beste era bat $\text{ROC}(\cdot) = \{(p, \text{ROC}(p)) : \ p \in (0,1)\}$ da non $\text{ROC}(p) = TPF(c)$ den, c mozketa-puntua $p = FPF(c)$ egiten duena izanik.

Mozketa-puntu ezberdinentzako emaitzak, denak batera, kontuan hartzen dituen parametroa AUC-a da, ROC kurbaren azpiko azalera. Honela kalkulatu ahal da:

$$\text{AUC} = \int_0^1 \text{ROC}(p)dp \tag{3.4}$$

AUC-a kalkulatzeko beste aukera bat hurrengo probabilitatea kalkulatzea da, izan ere 3.4 ekuazioaren baliokidea da [13].

$$\begin{aligned}\text{AUC} = P(M_i > M_j | &i. \text{ indibiduoak gertaera izan du,} \\ &j. \text{ indibiduoak ez du gertaera izan)}\end{aligned}$$

Denbora-menpeko aldagaiak daudenean, indibiduo bakoitzaren estimazioak denborarekin alda daitezke. Ondorioz, eredu batek duen gaitasuna gertaerak eta ez-gertaerak bereizteko denboran zehar ere alda daiteke, eta horrek diskriminazio-gaitasunean eragina izan dezake.

Jarraian, denbora-menpeko diskriminazio-gaitasuna neurtzeko literaturan proposatu diren parametro ezberdinak deskribatuko dira. Hauetan denbora-menpeko espezifikotasuna, sentikortasuna eta AUC-a erabiltzen dira.

3.2.1 Metatutako gertaerak / ez-gertaera dinamikoak

Parametro honetan interesekoa den denbora-tarte bat zehazten da, (s,t), eta denbora-tarte horretan gertaera izan duten indibiduoek gertaera multzoa osatzen dute, A bidez denotatuko dena. Denbora-tarte hori baino beranduago gertaera izan duten indibiduoek ez-gertaera multzoa osatuko dute, D bidez denotatuko duguna. Azkenik, gertaera denbora-tarte hori baino lehen izan duten indibiduoak ez dira analisian kontuan izango. Hori horrela, T biziraupen-denbora izanik, $A = \{i \mid T_i \in (s,t)\}$ eta $D = \{i \mid T_i > t\}$ izango dira.

Hau jakinda eta $c \in \mathbb{R}^+$ mozketa-puntu bat zehaztuta denbora-menpeko sentikortasunaren eta espezifikotasunaren definizioak ondorengoak dira:

$$\text{sentikortasuna}^A(c \mid (s,t)) = P(M > c \mid s \leq T \leq t) = P(M_i > c \mid i \in A)$$

$$\text{espezifikotasuna}^D(c \mid (s,t)) = P(M \leq c \mid T > t) = P(M_i \leq c \mid i \in D)$$

$\text{ROC}_{s,t}^{A/D}$ kurba $\text{ROC}_{s,t}^{A/D}(\cdot) = \{(p, \text{ROC}_{s,t}^{A/D}(p)) : \ p \in (0,1)\}$ izango da non $\text{ROC}_{s,t}^{A/D}(p) = TPF(c)$ den, c mozketa-puntua $p = FPF(c)$ egiten duena

izanik. Denbora-tarte bakoitzerako definitzen den AUC-a mozketa-puntu guztientzako definitu den $\mathrm{ROC}_{s,t}^{A/D}$ kurbaren azpiko azalera da:

$$\mathrm{AUC}^{A/D}(s,t) = \int_0^1 \mathrm{ROC}_{s,t}^{A/D}(p)dp$$

Honen baliokidea honako probabilitate hau kalkulatzea da:

$$\begin{aligned}\mathrm{AUC}^{A/D}(s,t) &= P(M_j > M_i|\ s \leq\ T_j \leq t,\ T_i > t) = \\ &= P(M_j > M_i|\ j \in A,\ i \in D)\end{aligned} \tag{3.5}$$

Zentsura ez dagoenean, $\mathrm{AUC}^{A/D}$ estimatu ahalko da erregresio logistikoan egiten den bezala. Hau da, Mann Whitney-ren estimatzailea erabiliz, non 3.5 ekuazioan adierazten den probabilitatea estimatzen den edo ROC kurbaren azpiko azalera estimatuz, sentikortasun/espezifikotasun bikote ezberdinek osatzen dituzten poligonoetatik abiatuta kurbaren azpiko azalera kalkulatuz. Bi metodoak baliokideak dira [14].

Zentsura egonez gero, ordea, $\mathrm{AUC}^{A/D}$ estimatu ahal izateko biziraupen-denborak estimatu beharko dira. Horretarako, literaturan estimatzaile ezberdinak proposatu izan dira. Hala ere, [15] artikuluan frogatzen da estimatzaile guztietatik bakarrik bik ondo funtzionatzen dutela: [16] artikuluan proposatutako estimatzaile bat, NNE bidez denotatuko duguna eta [17] artikuluan proposatutakoa, MB bidez denotatuko duguna. Proposatu diren beste estimatzaile guztiek limitazioren bat dute.

Hala ere, estimatzaile guztien artetik bakarrik 2 R-n inplementatuta daude, [16] artikuluan deskribatzen diren biak: KM bidez denotatuko dugun estimatzailea eta aurreko paragrafoan aipatutako NNE estimatzailea, hain zuzen. Aurreko guztia kontuan hartuz, azken bi estimatzaileekin lan egingo dugu. Lehenengoak Kaplan-Meier-en estimatzailea erabiltzen du biziraupen funtzioa estimatzeko eta hortik abiatuta $\mathrm{AUC}^{A/D}$ estimatzen du. Estimatzaile honen limitazioa lortzen diren balioak ez daudela $[0,1]$ tartean bornatuak da [15]. Bigarrenak, aldiz, auzo hurbilenaren metodoa erabiltzen du, biziraupen funtzioak Kernelean oinarritutako leunketen bidez lortuz. Hori horrela, ondoriozta dezakegu bi metodoak ez direla baliokideak eta gehienetan bigarren metodoa hobesten da [16].

Praktikan, denbora-menpeko diskriminazio-gaitasuna neurtzeko $\mathrm{AUC}^{A/D}$ erabiltzen da zehaztutako denbora-tarte ezberdinetarako eta gehienetan, zehazten diren denbora-tarte guztiek luzera berdina dute.

3.2.2 Gertakariak / ez-gertaera dinamikoak

Parametro honek aurrekoarekin konparatuz duen diferentziarik handiena oraingoan ez dela denbora-tarte bat zehaztu behar da. Kasu honetan, t den-

bora bat zehazten da soilik. Gertaeren multzoa, G, gertaera t unean izan duten indibiduoek osatuko dute, hau da, $G = \{i \mid T_i = t\}$ izango da. Bestalde, ez-gertaeren multzoa, aurreko ataleko berdina izango da, $D = \{i \mid T_i > t\}$, gertaera t baino beranduago izan dutenak. Horrez gain, gertaera t denbora baino lehen izan duten indibiduoak ez dira kontuan izango.

Egoera honetan, c mozketa-puntu bat zehaztuta, denbora-menpeko sentikortasunaren eta espezifikotasunaren definizioak hurrengoak dira:

$$\text{sentikortasuna}^G(c|t) = P(M > c \mid T = t) = P(M_i > c \mid i \in G)$$

$$\text{espezifikotasuna}^D(c|t) = P(M \leq c \mid T > t) = P(M_i \leq c \mid i \in D)$$

$\text{ROC}_t^{G/D}(\cdot) = \{(p, \text{ROC}_t^{G/D}(p)) : \ p \in (0,1)\}$ bidez definitu ahal da $\text{ROC}_t^{G/D}$ kurba non $\text{ROC}_t^{G/D}(p) = TPF(c)$ den, c mozketa-puntua $p = FPF(c)$ egiten duena izanik. Aurretik finkatutako t denbora bakoitzerako definitzen den AUC-a mozketa-puntu guztientzako definitu den $\text{ROC}_t^{G/D}$ kurbaren azpiko azalera da:

$$\text{AUC}^{G/D}(t) = \int_0^1 \text{ROC}_t^{G/D}(p)dp$$

Baliokideki:

$$\text{AUC}^{G/D}(t) = P(M_j > M_i |\ T_j = t,\ T_i > t) = P(M_j > M_i |\ j \in G,\ i \in D)$$

Zentsura egon ezean, $\text{AUC}^{G/D}$ estimatu ahalko da erregresio logistikoan egiten den antzera. Zentsura badago, oraingoan ere biziraupen denborak estimatu beharko dira baina, kasu honetan erabili beharreko estimatzaileak ez dira $\text{AUC}^{A/D}$ estimatzeko erabiltzen diren berdinak. [16] artikuluan bi estimatzaile proposatzen dituzte. Lehenengo aukera egiantz handieneko metodoan oinarritzen den estimatzaile erdi-parametriko bat erabiltzea da. Estimatzaile hau *rank* bidez denotatuko da. Bigarrena, aldiz, estimatzaile ez-parametriko bat erabiltzea da non Kernelean oinarritutako leunketak erabiliz estimazioak lortzen diren, *mean* estimatzailea deituko dugu. Gehienetan estimatzaile ez-parametrikoa erabiltzea nahiago da hipotesi gutxiago kontuan hartu behar direlako [16].

Aplikazio askotan, ez dago intereseko t denbora zehatzik, eta ereduaren denbora-menpeko diskriminazio-gaitasuna balio bakar batean laburtzen duen parametroa nahi da. Honela, eredu ezberdinen arteko konparaketak egin ahalko dira. Parametro honek aukera ematen du hau lortzeko, izan ere, $\text{AUC}^{G/D}$-ren batez-besteko haztatuak C-index-aren definizioarekin, 3.1 ekuazioarekin, bat egiten du [16]. Beraz:

$$C = \int_t \text{AUC}^{G/D}(t)w(t)dt$$

non $w(t) = 2f(t)S(t)$ den, $f(t)$ probabilitate dentsitate funtzioa eta $S(t)$ biziraupen funtzioa izanik.

C-ren balioa estimatzeko, $\text{AUC}^{G/D}$ aurreko paragrafoan azaldu diren metodo ez-parametrikoarekin edo erdiparametrikoarekin estimatzen da eta $w(t)$ Kaplan-Meier-en estimatzailea erabiliz lortzen da.

3.2.3 Bi parametroen arteko konparaketa

Praktikan $\text{AUC}^{G/D}(s)$ eta $\text{AUC}^{A/D}(s,t)$-ren balioak antzekoak izan ohi dira s eta t-ren arteko gakoa txikia denean. Horrez gain, kontextu deskriptibo batean, $\text{AUC}^{G/D}$ hobesten da, grafikoki sinplea den hurbilketa eta C-indexaren laburpena ematen digulako denbora-tarte bat zehaztu gabe. Hala ere, iraupen laburreko biziraupen iragarpena egin behar denean, $\text{AUC}^{A/D}$ erabiltzea komenigarriagoa da. Amaitzeko, esan beharra dago konputazionalki $\text{AUC}^{G/D}$ estimatzea errezagoa dela [16].

Oharra. 3.2 atalean azaldu diren kontzeptuetan M beharrean η erabili ahal da baina, kasu horretan, $c \in \mathbb{R}$ izan beharko litzateke.

3.3 Parametroen konfiantza-tarteen kalkulua

Diskriminazio-gaitasuna neurtzeko deskribatu diren parametro guztien kasuan, haien konfiantza tartea kalkulatu ahal da. Horretarako, gehien erabiltzen den teknika *bootstrap* metodoa da [18].

Boostrap metodoa teknika sinplea izan arren oso erabilgarria da konfiantza tarteak kalkulatzeko. Funtsean, jatorrizko laginetik abiatuz laginketa berriak egiten ditu. Laginketa hauetan errepikapenak onartzen dira, gainera, lortzen diren *bootstrap* laginketa berriak hasierako laginaren indibiduo kopuru bera izan behar dute.

Demagun, B laginketa ezberdin sortu direla. Orduan, $b = 1, \ldots, B$ laginketa bakoitzerako eredua doitzen da eta diskriminazio-gaitasuna neurtzeko deskribatu diren parametro guztien estimazioak lortzen dira. Horrela, parametro bakoitzerako B estimazio ezberdin egongo dira. $(1-\alpha)\cdot 100\%$ konfiantza tartea kalkulatzeko B estimazioen $\alpha/2$ eta $1-\alpha/2$ pertzentilak kalkulatzearekin nahikoa litzateke [19].

4. Kapitulua

Simulazioak

4.1 Sarrera

Aurreko kapituluan, biziraupen eredu batean, denbora-menpeko aldagaiak kontsideratuz eta kontsideratu gabe, diskriminazio-gaitasuna neurtzeko erabili ahal diren parametro ezberdinak deskribatu dira, bakoitza kalkulatzeko erabiltzen diren estimatzaile ezberdinak aipatuz. Kapitulu honetan, zehaztutako baldintza batzuen menpe, biziraupen ereduen diskriminazio-gaitasuna neurtzeko aurreko kapituluan aipatu diren estimatzaile guztien errendimendua eta erabilgarritasuna aztertzen da. Horretarako, simulazio ikerketa bat egin da egoera ezberdinak planteatuz eta kasu bakoitzerako parametro guztiak kalkulatuz, estimatzaile ezberdinak erabiliz. Horren ostean, balio teorikoarekin konparatuko dira.

$N = 10.000$ indibiduoko populazioa simulatu da non biziraupen-denborak CPH eredu batean oinarrituz simulatzen diren, oinarrizko arrisku funtzioa $\lambda = 0.01$ parametroko banaketa esponentzialari darraion arrisku funtzioa izanik. Zehazki, ereduak honako itxura hau du:

$$h(t|\mathbf{X}) = h_0(t)e^{0.6X_1(t)+0.1X_2-0.35X_3} = 0.01e^{0.6X_1(t)+0.1X_2-0.35X_3}$$

non $X_1(t) = a + bt$ itxura duen denbora-menpeko aldagaia den, $a \sim N(0, 2)$ eta $b \sim U(0, 0.2)$ izanik. Beste bi aldagaiak ez dira denbora-menpekoak, zehazki, $X_2 \sim U(-1, 2)$ eta $X_3 \sim \exp(0.8)$ bezala definitu dira. Simulatutako arrisku funtziotik abiatuz eta [20] artikuluan deskribatzen den metodoa erabiliz, biziraupen-denborak lortu dira. (Zehaztasun gehiagorako ikusi B eranskina). Aipagarria da simulatutako populazioan denbora hilabeteetan neurtuta dagoela, biziraupen-denbora minimoa 0.006 eta maximoa 1596.754 izanik.

Behin populazioa izanda parametro bakoitzaren balio teorikoak kalkulatu dira. Populazioko datuak direnez, benetako biziraupen-denbora guztiak eskuragarri daude eta denbora limiterik ez dagoenez, ez dago zentzurarik. Hori

dela eta, parametro bakoitzerako edozein estimatzaile erabilita emaitza oso antzekoak lortzen dira.

Ondoren, deskribatu diren parametro eta estimatzaileen propietateak aztertzeko, populazioa 500 aldiz lagindu da zoriz, $n = 500$ eta $n = 1000$ tamainako laginak kontsideratuz. Horrez gain, bi modu ezberdinetan lagindu da.

Lehenengoan, $X_1(t)$ aldagaia denbora-menpekoa (DM) dela kontsideratu da, aldagaiaren balioa urtero aldatuz. Bigarrenean, ordea, X_1 aldagaiak hasierako balioa hartzen du azterketa osoan zehar. Kasuistika honetan esango dugu X_1 aldagaiak balio basala hartzen duela. Bi kasuetan X_2 eta X_3 aldagaiak ere kontuan hartu dira eta zentsura 4 urte eta 6 hilabetetan jarri da, hau da, biziraupen denbora 4 urte eta 6 hilabete baino altuagoa duten indibiduoak zentsuratu dira. Horrela, zentsura % 14,9-an finkatuz.

Egoera guztietan lehenengo CPH eredua doitu da 3 aldagaiak ereduan sartuz eta ondoren, ereduaren diskriminazio-gaitasuna neurtu da 3. kapituluan deskribatu diren estimatzaile guztiak erabiliz. Aipagarria da $\text{AUC}^{A/D}$ $(0,\ 1)$, $(1,\ 2)$, $(2,\ 3)$ eta $(3,\ 4)$ tarteetan kalkulatu dela, parentesi arteko zenbaki horiek urteak adierazten dutelarik. Bestalde, $\text{AUC}^{G/D}$ $t = 0,\ 1,\ 2$ eta 3 urtetan kalkulatu da.

4.2 Emaitzak

Emaitzak aztertutako parametroaren (C-index, $\text{AUC}^{A/D}$, $\text{AUC}^{G/D}$) eta lagin tamainaren arabera azaltzen dira jarraian. Egoera bakoitza kutxa-diagrama eta laburpen taulen bidez azalduz. Hain zuzen, 4.1. taulan eta 4.1.a irudian $n = 500$ lagin tamainarekin C-index-arentzako lortutako laburpen taula eta kutxa diagrama erakusten dira, hurrenez hurren. Berdina baina $n = 1000$ lagin tamainarekin 4.2. taulan eta 4.1.b irudian adierazita dago. $\text{AUC}^{A/D}$-rentzako $n = 500$ lagin tamainarekin lortutako emaitzak 4.3. taulan eta 4.2. irudian ageri dira eta $n = 1000$ denean, 4.4. taulan eta 4.3. irudian. Azkenik, $\text{AUC}^{G/D}$-ren emaitzak 4.5. eta 4.6. tauletan eta 4.4. eta 4.5. irudietan azalduta daude, lagin tamaina $n = 500$ eta $n = 1000$ izanik, hurrenez hurren. Kutxa-diagramak interpretatzerako orduan, kontuan hartu behar da ardatzak aztertutako parametroaren arabera aldatu egiten direla.

Lehenik eta behin, egoera guztiak kontuan hartuz, ikus dezakegu lagin tamaina handituz gero, lortutako emaitzen sakabanapena txikitzen dela, esperogarria zen moduan. Hala ere, $n = 500$ eta $n = 1000$ tamainako laginak hartuz lortu diren emaitzak oso antzekoak dira eta beraz, ondorio oso antzekoak atera ahal dira bi kasuetan.

Alde batetik, C-index-ari dagokionez, ikusten da diferentzia nabariak daudela denbora-menpeko aldagaia kontuan hartzen denean eta ez denean aintzat hartzen. X_1 basala denean, 3 estimatzaileek alborapen handiak dituzte: Harrell -0.0585 eta -0.0586, *Risk* -0.0849 eta -0.0856 eta *Mean* -0.0630 eta -0.0610, lagin tamaina $n = 500$ eta $n = 1000$ izanik, hurrenez hurren. *Mean* eta Harrell-en estimatzaileek balio oso antzekoak hartzen dituzte, *risk*-ek, aldiz, bi hauek baino alborapen handiagoak ditu. Beraz, emaitzak ikusita esan dezakegu denbora-menpeko aldagaiak erabiltzen ez direnean, aurresateko gaitasuna azpiestimatuta geratzen dela hiru estimatzaileekin, Harrell-ena izanda hiruretatik alborapenik txikiena duena.

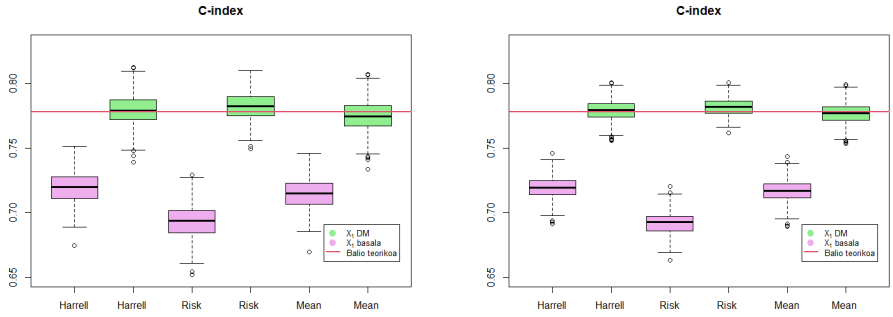
Horrez gain, X_1 denbora-menpekoa denean, Harrell-en estimatzailea balio teorikotik oso gertu gelditzen da, $n = 500$ zein $n = 1000$-ko lagin tamainak kontsideratuz. Izan ere, alborapen oso txikiak ditu. Hain zuzen, lagin tamaina $n = 500$ izanik alborapenak hauek izan dira: Harrell 0.0015, *Risk* 0.0044 eta *Mean* -0.0032 eta $n = 1000$ izanik, Harrell 0.0015, *Risk* 0.0040 eta *Mean* -0.0009. Hala ere, *risk* eta *mean* estimatzaileen alborapena ere nahiko txikia da. Aipagarria da $n = 1000$ tamainako laginak kontsideratuz *mean* estimatzaileak alborapen txikiena izan duela. Hori dela eta, esan dezakegu egoera horretan hiru estimatzaileek nahiko ondo funtzionatzen dutela, hirurekin emaitza nahiko antzekoak lortzen direlako. Ikusi 4.1. irudia eta 4.1. eta 4.2. taulak.

Bestalde, $\text{AUC}^{A/D}$-ren emaitzei erreparatuz gero, ondorio batzuk atera ditzakegu. Hasteko, esperogarria zenez, lehenengo tartean, hau da, (0, 1) tartean X_1 denbora-menpekoa zein basala kontsideratuz balio oso antzekoak lortu dira. Izan ere, tarte horretan X_1 aldagaiaren balioak berdinak dira. Gainontzeko tarteetan X_1 basalaren eta denbora-menpekoaren arteko diferentziak nabariagoak izaten doaz, diferentzia nabarienak azkenengo tartean egonik, hau da, (3, 4) tartean. Esate baterako, NNE estimatzailearekin $n = 500$ denean lortutako alborapenak lehenengo tartean 0.0002 eta 0.0050 izan dira, X_1 alde batetik basala eta bestetik denbora-menpekoa kontsideratuz, hurrenez hurren. Hurrengo tartean, hau da, (1, 2) tartean, -0.0794 eta -0.0008. (2, 3) tartean -0.2041 eta -0.0020 eta azken tartean, alborapenak -0.2791 eta 0.0019 izan dira.

Parametro honen estimatzaileei dagokienez, ikusten da baldintza hauetan bai NNE bai KM estimatzaileek balio oso antzekoak dituztela, alborapen txikienak NNE estimatzaileak izanik. Hori dela eta, baldintza hauetan [15] artikuluan frogatuta dagoen KM estimatzailearen limitazioa ez da ikusten, lortu diren balio guztiak [0, 1] tartean bornatuta daudelako. Ikusi 4.2. eta 4.3. irudiak eta 4.3. eta 4.4. taulak.

Azkenik, $\mathrm{AUC}^{G/D}$-ren inguruan esan dezakegu erabilitako bi estimatzaileen jarrera nahiko ezberdina dela. *Mean* estimatzaileak, X_1 denbora-menpekoa hartzen denean, balio teorikotik oso gertuko balioak lortzen ditu. Izan ere, alborapen oso txikiak ditu. Adibidez, $n = 1000$ tamainako laginak kontsideratuz alborapenak -0.0089, 0.0027, -0.0024 eta 0.0002 izan dira $t = 0$, 1, 2 eta 3 urte finkatuz, hurrenez hurren. Gainera, $\mathrm{AUC}^{A/D}$-n gertatzen den antzera, hasierako unean, hau da, $t = 0$ denean, X_1 denbora-menpekoa zein basala kontsideratuz balioak oso antzekoak dira eta zehaztutako denborak gora egin ahala X_1 denbora-menpekoa eta basalaren arteko diferentziak handitzen doaz diferentziarik handiena $t = 3$ denean lortuz. Horrek esan nahi du X_1 basala denean zenbat eta denbora epe luzeagoan aurresan nahi izan, orduan eta okerrago funtzionatzen duela *mean* estimatzaileak.

Beste alde batetik, *Risk* estimatzailearen alborapena handiagoa da aztertutako edozein unetan eta X_1 basala hartzen denean lortutako balioak balio teorikotik asko urruntzen dira. Esate baterako, $n = 500$ tamainako laginak kontsideratuz, $t = 0$, 1, 2 eta 3 urte finkatuz alborapenak -0.0825, -0.0830, -0.1018 eta -0.0912 izan dira, hurrenez hurren. Aipagarria da estimatzaile honekin lortzen den sakabanapena edozein unetan txikia dela. Azkenik, X_1 basala hartuta, denborak aurrera egin ahala *risk* estimatzaileak duen alborapena oso gutxi aldatzen da. Beste estimatzailearen kasuan, aldiz, (*mean*) denborak aurrera egin ahala alborapena handitzen doa, $t = 3$-n *risk* estimatzailea baino alborapen handiagoa izanik eta ez, ordea, $t = 0$-n. Ikusi 4.4. eta 4.5. irudiak eta 4.5. eta 4.6. taulak.



(a) (b)

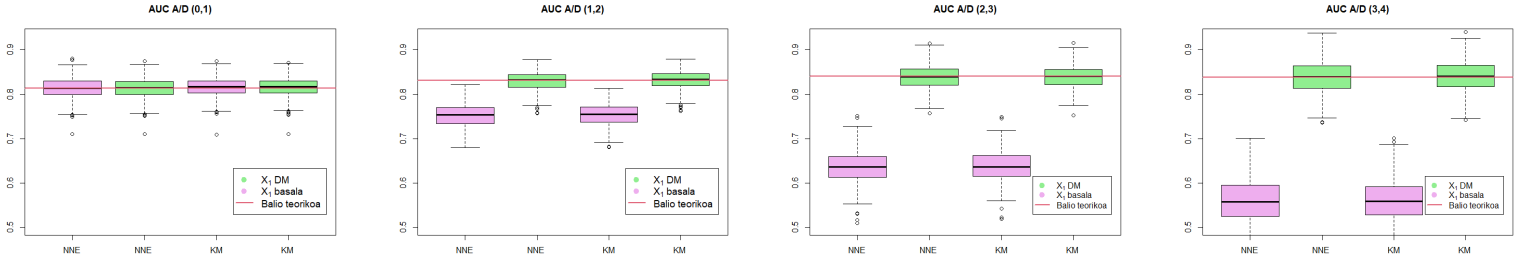
4.1. Irudia: C-index. (a) $n = 500$; (b) $n = 1000$.

Balio teorikoa	**Estimatzailea**	$\mathbf{X_1}$	**Batez bestekoa (sd)**	**Alborapena**
0.7777	Harrell	Basala	0.7193 (0.0115)	-0.0585
		DM	0.7792 (0.0117)	0.0015
	Risk	Basala	0.6928 (0.0123)	-0.0849
		DM	0.7821 (0.0104)	0.0044
	Mean	Basala	0.7147 (0.0114)	-0.0630
		DM	0.7746 (0.0116)	-0.0032

4.1. Taula: C-index-a $n = 500$ tamainako laginak kontsideratuz.

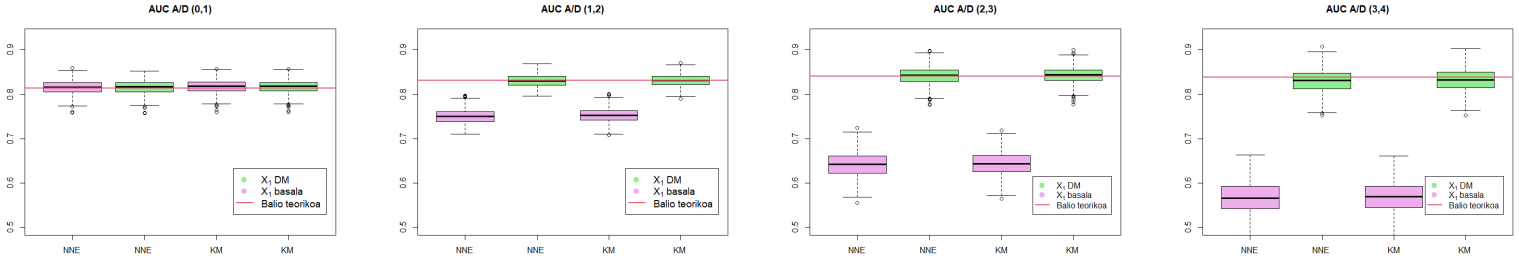
Balio teorikoa	**Estimatzailea**	$\mathbf{X_1}$	**Batez bestekoa (sd)**	**Alborapena**
0.7777	Harrell	Basala	0.7191 (0.0083)	-0.0586
		DM	0.7792 (0.0079)	0.0015
	Risk	Basala	0.6921 (0.0088)	-0.0856
		DM	0.7817 (0.0069)	0.0040
	Mean	Basala	0.7167 (0.0083)	-0.0610
		DM	0.7768 (0.0079)	-0.0009

4.2. Taula: C-index-a $n = 1000$ tamainako laginak kontsideratuz.



(a) (b) (c) (d)

4.2. Irudia: $n = 500$, $\text{AUC}^{A/D}$ tarteak: (a) (0, 1); (b) (1, 2); (c) (2, 3); (d) (3, 4).



(a) (b) (c) (d)

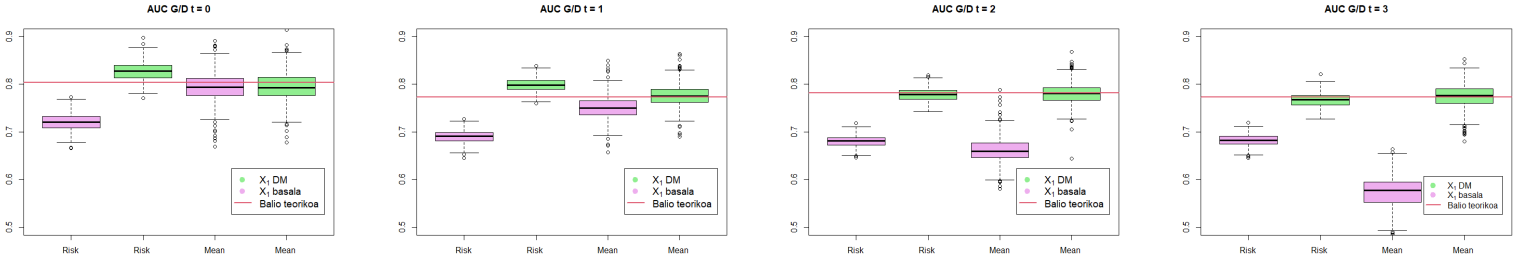
4.3. Irudia: $n = 1000$, $\text{AUC}^{A/D}$ tarteak: (a) (0, 1); (b) (1, 2); (c) (2, 3); (d) (3, 4).

Tartea	**Balio teorikoa**	**Estimatzailea**	$\mathbf{X_1}$	**Batez bestekoa (sd)**	**Alborapena**
(0, 1)	0.8137	NNE	Basala	0.8139 (0.0228)	0.0002
			DM	0.8142 (0.0228)	0.0050
		KM	Basala	0.8163 (0.0221)	0.0026
			DM	0.8163 (0.0222)	0.0026
(1, 2)	0.8313	NNE	Basala	0.7519 (0.0247)	-0.0794
			DM	0.8305 (0.0212)	-0.0008
		KM	Basala	0.7536 (0.0237)	-0.0777
			DM	0.8319 (0.0207)	0.0006
(2, 3)	0.8405	NNE	Basala	0.6364 (0.0359)	-0.2041
			DM	0.8385 (0.0256)	-0.0020
		KM	Basala	0.6374 (0.0344)	-0.2031
			DM	0.8394 (0.0245)	-0.0011
(3, 4)	0.8378	NNE	Basala	0.5587 (0.0517)	-0.2791
			DM	0.8396 (0.0346)	0.0019
		KM	Basala	0.5591 (0.0499)	-0.2787
			DM	0.8404 (0.0338)	0.0027

4.3. Taula: $\text{AUC}^{A/D}$ $n = 500$ tamainako laginak kontsideratuz.

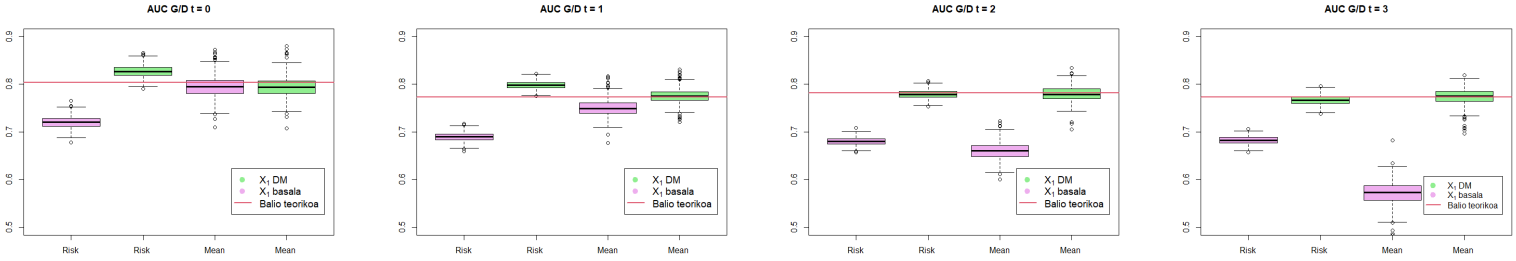
Tartea	**Balio teorikoa**	**Estimatzailea**	$\mathbf{X_1}$	**Batez bestekoa (sd)**	**Alborapena**
(0, 1)	0.8137	NNE	Basala	0.8150 (0.0156)	0.0013
			DM	0.8150 (0.0155)	0.0013
		KM	Basala	0.8170 (0.0151)	0.0033
			DM	0.8169 (0.0152)	0.0032
(1, 2)	0.8313	NNE	Basala	0.7500 (0.0165)	-0.0813
			DM	0.8299 (0.0142)	-0.0014
		KM	Basala	0.7521 (0.0161)	-0.0792
			DM	0.8310 (0.0138)	-0.0003
(2, 3)	0.8405	NNE	Basala	0.6346 (0.0254)	-0.2060
			DM	0.8385 (0.0183)	-0.0020
		KM	Basala	0.6363 (0.0250)	-0.2042
			DM	0.8393 (0.0178)	-0.0012
(3, 4)	0.8378	NNE	Basala	0.5588 (0.0351)	-0.2789
			DM	0.8386 (0.0229)	0.0009
		KM	Basala	0.5600 (0.0343)	-0.2778
			DM	0.8402 (0.0218)	0.0024

4.4. Taula: $\text{AUC}^{A/D}$ $n = 1000$ tamainako laginak kontsideratuz.



(a) (b) (c) (d)

4.4. Irudia: $n = 500$, $\mathrm{AUC}^{G/D}$ (a) $t = 0$; (b) $t = 1$; (c) $t = 2$; (d) $t = 3$.



(a) (b) (c) (d)

4.5. Irudia: $n = 1000$, $\mathrm{AUC}^{G/D}$ (a) $t = 0$; (b) $t = 1$; (c) $t = 2$; (d) $t = 3$.

Denbora	**Balio teorikoa**	**Estimatzailea**	**X_1**	**Batez bestekoa (sd)**	**Alborapena**
t = 0	0.8032	*Risk*	Basala	0.7207 (0.0180)	-0.0825
			DM	0.8272 (0.0181)	0.0240
		Mean	Basala	0.7946 (0.0314)	-0.0086
			DM	0.7948 (0.0309)	-0.0084
t = 1	0.7731	*Risk*	Basala	0.6901 (0.0130)	-0.0830
			DM	0.7983 (0.0130)	0.0252
		Mean	Basala	0.7506 (0.0241)	-0.0225
			DM	0.7761 (0.0233)	0.0031
t = 2	0.7822	*Risk*	Basala	0.6804 (0.0111)	-0.1018
			DM	0.7784 (0.0134)	-0.0038
		Mean	Basala	0.6623 (0.0271)	-0.1199
			DM	0.7800 (0.0227)	-0.0022
t = 3	0.7735	*Risk*	Basala	0.6823 (0.0115)	-0.0912
			DM	0.7669 (0.0151)	-0.0066
		Mean	Basala	0.5733 (0.0319)	-0.2001
			DM	0.7741 (0.0254)	0.0006

4.5. Taula: $\text{AUC}^{G/D}$ $n = 500$ tamainako laginak kontsideratuz.

Denbora	**Balio teorikoa**	**Estimatzailea**	**X_1**	**Batez bestekoa (sd)**	**Alborapena**
t = 0	0.8032	*Risk*	Basala	0.7201 (0.0126)	-0.0831
			DM	0.8273 (0.0122)	0.0240
		Mean	Basala	0.7956 (0.0224)	-0.0076
			DM	0.7943 (0.0222)	-0.0089
t = 1	0.7731	*Risk*	Basala	0.6898 (0.0092)	-0.0833
			DM	0.7984 (0.0084)	0.0254
		Mean	Basala	0.7503 (0.0167)	-0.0228
			DM	0.7758 (0.0151)	0.0027
t = 2	0.7822	*Risk*	Basala	0.6803 (0.0080)	-0.1019
			DM	0.7788 (0.0090)	-0.0034
		Mean	Basala	0.6606 (0.0182)	-0.1216
			DM	0.7798 (0.0157)	-0.0024
t = 3	0.7735	*Risk*	Basala	0.6823 (0.0086)	-0.0911
			DM	0.7672 (0.0104)	-0.0063
		Mean	Basala	0.5717 (0.0240)	-0.2018
			DM	0.7737 (0.0176)	0.0002

4.6. Taula: $\text{AUC}^{G/D}$ $n = 1000$ tamainako laginak kontsideratuz.

5. Kapitulua

Aplikazioa

Aurreko kapituluetan erdietsitako emaitza teorikoak praktikan jartzeko, datubase erreal batekin lan egin da. Datu hauek Galdakao-Usansolo Unibertsitate Ospitalean egin den BGBK-ren inguruko ikerketa batekoak dira.

Helburua CPH eredu bat doitzea da gaixoen biziraupena ahalik eta xehetasun handienarekin zehazten duena.

5.1 Datu-basearen deskribapena

399 indibiduoko lagina jaso da eta bertako erantzun aldagaia heriotza da. Ikerketaren jarraipena 4 urtekoa izan da, epe horren barruan 77 paziente hil dira, zeinetarako heriotza-data neurtuta dagoenez, gertaera denborak ezagunak diren. Gainontzeko 322 pazienteak, zentzuratutako pazienteak izan dira, zentzura %81-ekoa izanik. Hauentzako ikerketa hasiera eta ikerketa arteko denbora neurtu da, horrela behatutako denbora izanik. Azpimarratu nahi da datu-basean denbora egunetan neurtuta dagoela.

Datu-base honetan lau momentu ezberdin daude, paziente bakoitzarentzako medikuak finkaturiko lau bisita, gutxi gora behera, urtean bisita bat. Aipagarria da paziente guztiek ez dituztela zertan bisita guztiak izan. Bisita bakoitzean parametro batzuk neurtu dira: *Walking test*-a (X_3), inspirazio presio maximoa (X_8) eta lau indar neurri: esku-indarra (X_4), koadrizepsaren indarra (X_5), koadrizepsaren *Lafayette* neurria (X_6) eta koadrizepsaren indarra hedaduran (X_7). Hauek denbora-menpeko aldagaiak dira. Horrez gain, denbora-menpekoak neurtu ez diren aldagaiak ere jaso dira: sexua (X_1), adina (X_2) eta Charlson-en indizea (X_9). Aldagai azaltzaile guztiak jarraituak dira sexua (X_1) izan ezik, aldagai kategorikoa dela.

Datu-basearen eta BGBK gaixotasunari buruzko informazio gehiago A eranskinean aurkitu ahal da.

5.2 Datuen analisia

Bi analisi ezberdin planteatu dira. Alde batetik, aldagai azaltzaile guztiak kontuan hartu dira baina, denbora-menpeko aldagaien kasuan haien hasierako balioa hartu da soilik. Hau da, balio basalekin lan egin da. Bestetik, aldagai aske guztiak kontuan hartu dira denbora-menpekoak diren aldagaien kasuan haiek emandako informazio guztia kontsideratuz. Beraz, kasu honetan, doituko den CPH ereduak denbora-menpeko aldagaiak izango ditu.

Aipagarria da R-n denboraren menpeko aldagaiak CPH ereduan sartzeko beharrezkoa dela datuetan moldaketa bat egitea. Gure kasuan, paziente bakoitzari zenbaki bat esleitu behar zaio ('ID') eta datu-basean paziente eta bisita bakoitzeko errenkada bat egongo da. Errenkada bakoitzean pazienteari esleitutako zenbakia eta bisita horretan neurtutako aldagai aske guztien balioa egongo da. Hau egiteko R-ko *tmerge* funtzioa erabili ahal da. Moldaketa egin ondoren datu-basearen itxura honako hau da:

ID	tstart	tstop	hil	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	0	349	0	1	56	461	35	33	26.2	118.4	32.6	1
1	349	730	0	1	56	469	46	33.5	28.6	142.8	41.5	1
1	730	758	0	1	56	484	36	37	31.1	160.1	83.7	1
1	758	1837	0	1	56	490	38	35.2	33	166.8	82.9	1
4	0	382	0	1	71	492	35	22.5	26.1	126.4	66.8	3
4	382	762	1	1	71	498	36	22.5	29.3	114.8	55.6	3

Behin analisi bakoitza egiteko erabiliko diren datuak zehaztuta, analisiari ekin diogu. Bi kasuetan egindako lehenengo gauza aldagai azaltzaile bakoitzaren adierazgarritasuna eredu bakunean aztertzea izan da. Bai Wald-en testa bai egiantz arrazoiaren testa aplikatuz, aldagai guztiak adierazgarriak izan dira, $p < 0.05$. Hori dela eta, hurrengo urratsa aldagai guztiak kontuan hartzen dituen CPH eredua doitzea izan da.

Horren ondoren, aldagai aske guztiak zituen ereduan adierazgarriak diren aldagaiekin bakarrik gelditzea izan dugu helburu. Horretarako, egiantz arrazoiaren testa erabiliz, ereduak abiaratzen joan gara. Modu honetan, ereduak konparatzen joan gara aldagai gutxiago zuten ereduekin, betiere abiaratutako bi ereduak bata bestearen azpimultzoa izanik.

Behin CPH eredua zehaztuta, zeinetan sartutako aldagai guztiak adierazgarriak diren, aldagaien arteko interakzioak aztertu ditugu. Ez da interakzio adierazgarririk aurkitu. Balio basalak kontuan hartzen dituen eredua ω_1 bezala denotatuko da eta denbora-menpeko aldagaiak dituen eredua ω_2 bidez. Eredu bakoitzak hurrengo aldagaiak ditu:

ω_1 eredua: X_1, X_3, X_5 eta X_9.
ω_2 eredua: X_1, $X_3(t)$, $X_5(t)$, $X_6(t)$, $X_7(t)$ eta X_9.

2. kapituluan azaldutakoa kontuan hartuz, CPH eredu bat doitu ostean konprobatu behar da eredu horrek arrisku proportzionalen hipotesia betetzen duela. ω_1 eta ω_2 ereduen Schoenfeld-en hondarren azterketa egin ondoren lortutako emaitzak 5.1. eta 5.2. tauletan adierazita daude, hurrenez hurren.

	ρ	AG	p-balioa
$\mathbf{X_1}$	3.36	1	0.067
$\mathbf{X_3}$	1.61	1	0.205
$\mathbf{X_5}$	0.714	1	0.398
$\mathbf{X_9}$	0.122	1	0.727
Orokorra	4.069	4	0.397

5.1. Taula: ω_1 ereduaren Schoenfeld-en hondarren azterketa. ρ aplikatutako testaren estatistikoaren balioa, AG askatasun graduak eta p-balioa izanik.

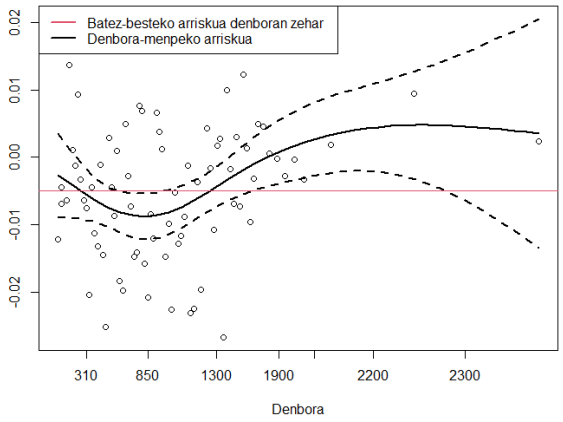
	ρ	AG	p-balioa
$\mathbf{X_1}$	3.555	1	0.059
$\mathbf{X_3(t)}$	7.098	1	0.008
$\mathbf{X_5(t)}$	0.025	1	0.875
$\mathbf{X_6(t)}$	0.843	1	0.359
$\mathbf{X_7(t)}$	0.151	1	0.697
$\mathbf{X_9}$	0.512	1	0.474
Orokorra	10.104	6	0.121

5.2. Taula: ω_2 ereduaren Schoenfeld-en hondarren azterketa. ρ aplikatutako testaren estatistikoaren balioa, AG askatasun graduak eta p-balioa izanik.

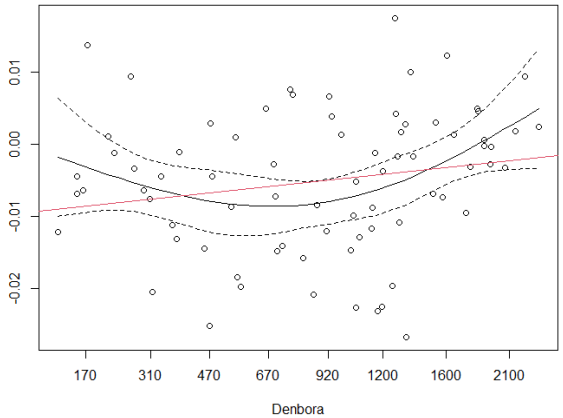
Taulei erreparatuz esan dezakegu ω_1 ereduan aldagai guztiek arrisku proportzionalen hipotesia betetzen dutela eta beraz, eredua egokitzat har dezakegula. ω_2 ereduaren kasuan, aldiz, $X_3(t)$ aldagaiak ez du hipotesia betetzen, izan ere, $p < 0.05$ da. Grafikoki ere ikusten da $X_3(t)$ aldagaiaren hondarrak denboran zehar ez direla konstanteak eta beraz, aldagaiaren efektua ez dela lineala (5.1. irudia). Arazo horri aurre egiteko, ω_2 ereduan 2.3 atalean azaldu diren metodo ezberdinak aplikatuko ditugu.

Lehenengo aukera denbora-menpeko koefizienteak erabiltzea da. ω_3 bezala denotatuko den eredua doitu dugu non ω_2 ereduan dauden aldagai guztiak sartu diren baina, $X_3(t)$ aldagaiari $\beta(t) = a + b\log(t + 365)$ itxura duen denbora-menpeko koefizientea esleituz. Modu honetan, hazard ratioak forma interpretagarria izango du, baita ikerketa egin den egun guztien linealtasuna bermatuko da ere. Doitutako eredu berrian, berriz ere, aldagai guztiak

adierazgarriak dira eta eredu honek arrisku proportzionalen hipotesia betetzen du. Izan ere, 5.2. irudian ikusten da denbora-menpeko koefiziente hau zehazten denean $X_3(t)$ aldagaiaren efektua linealtzat hartu ahal dela.



5.1. Irudia: ω_2 ereduko $X_3(t)$ aldagaiaren Schoenfeld-en hondarren azterketa.



5.2. Irudia: ω_3 ereduko $X_3(t)$ aldagaiaren Schoenfeld-en hondarren azterketa.

Bigarren aukera, erregresio *spline*-ak erabiltzea da. 5.1. irudiari erreparatuz gero, $X_3(t)$ aldagaiaren efektua zatika 3. mailako polinomio baten itxura izan ahal duela ikusten da. Hori horrela, ω_4 eredua doitu da, ω_2 ereduak dituen aldagai guztiak sartuz eta $X_3(t)$ aldagaiari *spline* kubikoa ezarriz. ω_4 ereduaren aldagai guztiak adierazgarriak dira. Eredua doitu ondoren arris-

ku proportzionalen hipotesia betetzen den ala ez aztertu da. 5.3. taulako datuen arabera esan dezakegu hipotesia betetzen dela eta ondorioz, eredua ontzat har dezakegula.

	ρ	AG	p-balioa
$\mathbf{X_1}$	2.893	1	0.089
$\mathbf{X_3(t)}$	7.19	3	0.071
$\mathbf{X_5(t)}$	0.006	1	0.936
$\mathbf{X_6(t)}$	0.926	1	0.334
$\mathbf{X_7(t)}$	0.144	1	0.702
$\mathbf{X_9}$	0.586	1	0.441
Orokorra	9.879	8	0.28

5.3. Taula: ω_4 ereduaren Schoenfeld-en hondarren azterketa. ρ aplikatutako testaren estatistikoaren balioa, AG askatasun graduak eta p-balioa izanik.

Azken aukera, $X_3(t)$ jarraitua denez, aldagaia kategorizatzea da. $X_3(t)$ kategorizatzeko irizpide bezala BODE [21] eskala sortzeko erabiltzen diren mozketa puntuak erabili dira. Hau da, *Walking test*-ean 350 metro baino gehiago egiten dituzten pazienteak *Walking test* ez-patologikoa izango dute eta 350 metro baino gutxiago egiten dituzten pazienteek, aldiz, *Walking test* patologikoa izango dute. *Walking test* patologikoaren barnean 3 maila ezberdin daude, baina, jasotako lagina ez denez oso handia ezin izan ditugu bereizketa horiek egin eta 2 kategoria soilik sortu ditugu: *Walking test* ez-patologikoa ($X_3(t) \geq 350$) eta patologikoa ($X_3(t) < 350$). Hori horrela, CPH eredu bat doitu dugu ω_2 ereduak dituen aldagai guztiak kontsideratuz baina, $X_3(t)$ kategorikoa sartuz, ω_5 denotatuko duguna. Kasu honetan ere ω_5 ereduaren aldagai guztiak adierazgarriak dira eta 5.4. taulari erreparatuz gero ondoriozta dezakegu ereduak arrisku proportzionalen hipotesia betetzen duela.

	ρ	AG	p-balioa
$\mathbf{X_1}$	2.266	1	0.13
$\mathbf{X_3\_kat(t)}$	1.947	1	0.16
$\mathbf{X_5(t)}$	0.001	1	0.98
$\mathbf{X_6(t)}$	0.678	1	0.41
$\mathbf{X_7(t)}$	0.076	1	0.78
$\mathbf{X_9}$	0.674	1	0.41
Orokorra	6.113	6	0.41

5.4. Taula: ω_5 ereduaren Schoenfeld-en hondarren azterketa. ρ aplikatutako testaren estatistikoaren balioa, AG askatasun graduak eta p-balioa izanik.

Behin ereduak doituta, eredu bakoitzaren diskriminazio-gaitasuna aztertuko da. Nahiz eta ω_2 eredua guztiz egokia ez izan, ereduaren diskriminazio-gaitasuna ere aztertuko da.

5.3 Diskriminazio-gaitasuna

Atal honetan aurreko atalean zehaztu diren ereduen diskriminazio-gaitasuna neurtuko da. Horretarako, 3. kapituluan azaldu diren parametro ezberdinak kalkulatuko dira. Horrez gain, *bootstrap* metodoa erabiliz parametroen konfiantza tartea kalkulatuko da. Atal honetako taula guztietan parentesi artean dauden balioak konfiantza tartea dira.

Ereduen C-index-a aztertzen hasiko gara. C-index-a kalkulatzeko 3 estimatzaile deskribatu dira: Harrell-ena, *risk* eta *mean*. 5.5. taulan ageri dira ereduek hartzen dituzten balioak. Orokorrean ikusten da eredu guztiek C-index nahiko altua dutela, hori dela eta, esan dezakegu, parametro honen arabera, eredu guztien aurresateko gaitasuna nahiko altua dela. Horrez gain, ez da diferentzia handirik ikusten ereduen artean, izan ere, balioak oso antzekoak dira kasu guztietan.

C-index	**Estimatzailea**		
Eredua	**Harrell**	***Risk***	***Mean***
$\boldsymbol{\omega_1}$	0.836 (0.79, 0.87)	0.802 (0.77, 0.85)	0.819 (0.76, 0.85)
$\boldsymbol{\omega_2}$	0.844 (0.79, 0.89)	0.818 (0.79, 0.86)	0.826 (0.76, 0.86)
$\boldsymbol{\omega_3}$	0.844 (0.77, 0.87)	0.780 (0.74, 0.83)	0.804 (0.74, 0.84)
$\boldsymbol{\omega_4}$	0.843 (0.79, 0.88)	0.829 (0.79, 0.87)	0.825 (0.76, 0.86)
$\boldsymbol{\omega_5}$	0.835 (0.79, 0.88)	0.804 (0.77, 0.85)	0.818 (0.76, 0.86)

5.5. Taula: Doitutako ereduen C-index-aren azterketa.

Aztertuko dugun hurrengo parametroa $\text{AUC}^{A/D}$ da. Bi estimatzaile ezberdin, NNE eta KM, erabiliz kalkulatu dugu lau tarte ezberdinetan: (0, 1), (1, 2), (2, 3) eta (3, 4) urte zehaztuz. Lortutako emaitzak 5.6. taulan adierazita daude.

Hasteko, emaitzei erreparatuz, ikusten da KM estimatzaileak ez duela ondo funtzionatzen, izan ere, kasu batzuetan 1 baino altuagoak diren balioak ematen ditu. Estimatzaile honen limitazio hau [15] artikuluan frogatuta dago. Hori horrela, NNE estimatzailearekin lortutako balioak interpretatuko ditugu.

ω_1 ereduak (0, 1) tartean gainontzeko ereduak baino aurresateko gaitasun handiagoa duela ikusten da, hala ere, beste tarte guztietan gainontzeko ereduek dutena baino baxuagoa du. Honek zentzua izan dezake ω_1 ereduak hasierako balioak soilik kontuan hartzen dituelako. Beste eredu guztien diskriminazio-gaitasuna nahiko altua da. Aipagarriak dira ω_4 eta ω_5 ereduak. Lehenengoaren kasuan, tarte guztietan balio nahiko onak ditu, ω_5

ereduak, aldiz, (0,1) tartean ez du errendimendu oso ona baina, gainontzeko tarte guztietan oso emaitza onak lortzen ditu.

$\mathbf{AUC^{A/D}}$	**NNE**			
Eredua	**(0, 1)**	**(1, 2)**	**(2, 3)**	**(3, 4)**
$\boldsymbol{\omega_1}$	0.820 (0.64, 0.93)	0.928 (0.87, 0.96)	0.877 (0.77, 0.94)	0.797 (0.71, 0.91)
$\boldsymbol{\omega_2}$	0.816 (0.63, 0.94)	0.989 (0.96, 1)	0.888 (0.82, 0.95)	0.867 (0.76, 0.95)
$\boldsymbol{\omega_3}$	0.776 (0.63, 0.90)	0.949 (0.90, 1)	0.859 (0.74, 0.94)	0.810 (0.63, 0.93)
$\boldsymbol{\omega_4}$	0.809 (0.64, 0.94)	0.990 (0.95, 1)	0.869 (0.79, 0.94)	0.841 (0.73, 0.94)
$\boldsymbol{\omega_5}$	0.745 (0.54, 0.92)	0.985 (0.93, 1)	0.889 (0.82, 0.94)	0.851 (0.73, 0.93)
	KM			
Eredua	**(0, 1)**	**(1, 2)**	**(2, 3)**	**(3, 4)**
$\boldsymbol{\omega_1}$	0.822 (0.69, 0.92)	0.927 (0.87, 0.96)	0.870 (0.79, 0.95)	0.794 (0.71, 0.90)
$\boldsymbol{\omega_2}$	0.795 (0.64, 0.93)	1.003 (0.80, 1.39)	0.832 (0.78, 0.90)	0.902 (0.80, 0.99)
$\boldsymbol{\omega_3}$	0.787 (0.66, 0.90)	0.951 (0.78, 1.33)	0.811 (0.72, 0.90)	0.864 (0.74, 0.96)
$\boldsymbol{\omega_4}$	0.803 (0.66, 0.93)	0.982 (0.79, 1.38)	0.824 (0.76, 0.89)	0.897 (0.79, 0.99)
$\boldsymbol{\omega_5}$	0.771 (0.61, 0.91)	1.009 (0.80, 1.41)	0.835 (0.78, 0.89)	0.906 (0.80, 0.99)

5.6. Taula: Doitutako ereduen $\text{AUC}^{A/D}$-ren azterketa.

Aztertuko den azken parametroa $\text{AUC}^{G/D}$ da. Kasu honetan ere bi estimatzaile ezberdin erabiliz kalkulatu dugu, *risk* eta *mean*, lau momentu ezberdinetan: $t = 0$, 1, 2 eta 3 urteetan. Emaitzak 5.7. taulan ageri dira.

$\mathbf{AUC^{G/D}}$	***Risk***			
Eredua	**t = 0**	**t = 1**	**t = 2**	**t = 3**
$\boldsymbol{\omega_1}$	0.833 (0.79, 0.88)	0.828 (0.79, 0.88)	0.817 (0.78, 0.87)	0.782 (0.75, 0.84)
$\boldsymbol{\omega_2}$	0.823 (0.78, 0.87)	0.823 (0.79, 0.87)	0.826 (0.79, 0.87)	0.806 (0.77, 0.85)
$\boldsymbol{\omega_3}$	0.776 (0.72, 0.83)	0.780 (0.73, 0.83)	0.786 (0.74, 0.84)	0.773 (0.73, 0.82)
$\boldsymbol{\omega_4}$	0.832 (0.79, 0.88)	0.834 (0.79, 0.88)	0.836 (0.80, 0.88)	0.820 (0.79, 0.86)
$\boldsymbol{\omega_5}$	0.817 (0.77, 0.87)	0.812 (0.77, 0.86)	0.805 (0.77, 0.85)	0.792 (0.76, 0.83)
	Mean			
Eredua	**t = 0**	**t = 1**	**t = 2**	**t = 3**
$\boldsymbol{\omega_1}$	0.846 (0.69, 0.94)	0.862 (0.80, 0.96)	0.890 (0.79, 0.92)	0.826 (0.78, 0.91)
$\boldsymbol{\omega_2}$	0.835 (0.69, 0.94)	0.868 (0.79, 0.97)	0.907 (0.84, 0.94)	0.858 (0.81, 0.94)
$\boldsymbol{\omega_3}$	0.823 (0.66, 0.93)	0.858 (0.79, 0.95)	0.886 (0.79, 0.93)	0.834 (0.79, 0.93)
$\boldsymbol{\omega_4}$	0.839 (0.69, 0.94)	0.870 (0.80, 0.97)	0.901 (0.83, 0.93)	0.851 (0.80, 0.93)
$\boldsymbol{\omega_5}$	0.821 (0.67, 0.93)	0.852 (0.77, 0.96)	0.898 (0.84, 0.92)	0.854 (0.79, 0.91)

5.7. Taula: Doitutako ereduen $\text{AUC}^{G/D}$-ren azterketa.

Orokorrean, bi estimatzaileekin lortzen diren balioak nahiko ezberdinak dira. Aurreko kapituluan ikusi dugu *mean* estimatzaileak hobeto funtzionatzen duela eta [16] artikuluan esaten da gehienetan estimatzaile horrekin

lan egitea hobesten dela. Hori dela eta, ondorioak *mean* estimatzailearekin lortu diren emaitzekin aterako ditugu.

$\text{AUC}^{A/D}$-rekin gertatzen den antzera, ω_1 ereduak hasieran diskriminazio-gaitasun altuena du eta denborak aurrera egin ahala beste ereduekin konparatuz emaitza txarragoak lortzen ditu. Oraingoan, gainontzeko ereduek edozein unetan aurresateko gaitasun altua dute eta berriz ere, aipagarriak dira ω_4 eta ω_5 ereduak, emaitza oso onak lortzen dituztelako edozein momentutan.

Laburbilduz, doitu diren eredu guztiek diskriminazio-gaitasun altua dute. Hala ere, ikusi da denbora-menpeko aldagaiak dituzten ereduetan, denborak aurrera egin ahala, haien aurresateko gaitasuna oso ona izaten jarraitzen duela. Hori dela eta, aukera izanda, komenigarriagoa da denbora-menpeko aldagaiak kontsideratzea. Emaitzei erreparatuz, ω_4 eta ω_5 ereduen diskriminazio-gaitasuna oso ona da kasu guztietan eta eredu bat aukeratu behar izatekotan horietako bat aukeratuko genuke. Ziurrenik, ω_5 eredua hautatuko genuke beste eredua baino interpretazio errezagoa duelako.

5.4 Ereduaren interpretazioa

ω_5 eredua doitzerako orduan lortutako koefizienteen balioak hauek dira:

	$\boldsymbol{\beta}$	**w**	$\boldsymbol{p}$	$\mathbf{e}^{\boldsymbol{\beta}}$
$\mathbf{X_1}$ **(Emakumeak)**	-1.318	-3.546	< 0.001	0.268
$\mathbf{X_3\_kat(t)}$ **(Patologikoa)**	0.829	3.201	0.001	2.29
$\mathbf{X_5(t)}$	-0.081	-3.54	< 0.001	0.922
$\mathbf{X_6(t)}$	0.056	2.043	0.041	1.057
$\mathbf{X_7(t)}$	-0.012	-2.474	0.013	0.988
$\mathbf{X_9}$	0.253	3.32	< 0.001	1.288

Emaitzak ikusita esan dezakegu beste aldagai guztiak konstante mantenduta arriskuaren murrizketa egongo dela pazientea gizona izan beharrean emakumea bada, $\text{HR} = e^{\beta_1} = 0.268 < 1$ delako. $X_5(t)$-en balioa handitzen denean, beste aldagai guztiak konstante mantenduz, heriotzaren arriskuaren murrizketa ere egongo da, izan ere, HR-a ere 1 baino txikiagoa da. Berdina gertatzen da $X_7(t)$ aldagaiarekin.

Bestalde, aldagai guztiak konstante mantenduta heriotzarako arriskuaren gehikuntza egongo da pazienteak *Walking test* ez-patologikoa izan beharrean *Walking test* patologikoa badu, $\text{HR} = e^{\beta_3} = 2.29 > 1$ delako. Horrez gain, X_9-ren balioak gora egiten badu, beste aldagai guztiak konstante mantenduz, kasu honetan ere, $\text{HR} > 1$ izango da, beraz, arriskuaren gehikuntza egongo da. Ondorio berdinera hel genezake $X_6(t)$ aldagaiarekin.

6. Kapitulua

Ondorioak

Lanaren helburuak zeintzuk diren eta aurreko kapituluetan jasotako informazio guztia kontuan hartuz, hainbat ondorio atera dira. Hurrengo lerroetan azalduta daude.

Hasteko, lanean zehar Cox-en arrisku proportzionalen eredu bat zelan doitzen den aztertu da eta eredu horietan denbora-menpeko koefizienteak gehitu ahal direla ikusi da. Jasotako emaitzetan ikusi da, orokorrean, CPH eredu batean denbora-menpeko aldagaiak sartuz gero ereduaren diskriminazio-gaitasunak gora egiten duela. Hori dela eta, komenigarria ikusten dugu ahal den kasu guztietan CPH ereduetan denbora-menpeko aldagaiak sartzea. Honela, eredua informazio gehiago kontuan hartzen egongo da eta ereduaren hasiera uneko aurresateko gaitasuna denboran zehar mantendu ahalko da.

Ereduen diskriminazio-gaitasuna neurtzeko proposatu diren parametroei dagokionez, C-index-a parametro orokorra dela ikusi da. Eredu baten aurresateko gaitasuna era orokorrean neurtzen du, azterketa egin deneko momentu guztietarako. Parametro honen desabantaila denborarekin ez duela bereizketarik egiten da. Hau da, gertatu ahal da eredu batek hasieran, $t = 0$ denean, diskriminazio-gaitasun oso altua izatea eta denborak aurrera egin ahala aurresateko gaitasuna galtzen joatea. Hori horrela, lortutako C-indexaren balioa nahiko ona izan ahalko da nahiz eta ereduaren errendimendua azterketa egin den denbora osoan zehar ona ez izan. Parametro honen abantaila da zenbaki bakar bat bueltatzen duenez, C-index-arekin ereduen arteko konparaketak modu sinple batean egin daitezkela.

Bestalde, C-index-a estimatzeko proposatutako estimatzaileen inguruan esan dezakegu hirurek antzeko jarrera erakutsi dutela, hala ere, Harrell-en estimatzaileak alborapen txikienak izan ditu, konputazionalki azkarrena izan da eta interpretazio intuitiboena du.

Denboraren-menpeko AUC-en kasuan, zehaztutako denbora-tarte edo momentu bakoitzerako balio bat ematen dute, eta beraz, aukera ematen dute ereduen diskriminazio-gaitasuna denboran zehar mantentzen den ala ez aztertzeko. Bi denbora-menpeko AUC ezberdin deskribatu dira: $\text{AUC}^{A/D}$ eta $\text{AUC}^{G/D}$.

Lehenengoa estimatzeko aztertu diren bi estimatzaileen artean argi ikusi da batek, zehazki NNE-k, ondo funtzionatzen duela eta informazio baliagarria ematen duela ereduen aurresateko gaitasuna aztertzeko orduan. Hala ere, konputazionalki kostu handia du eta lagin-tamaina handitzen denean kostua asko handitzen da.

$\text{AUC}^{G/D}$-ren inguruan esan dezakegu deskribatu diren bi estimatzaileen artetik ikusi dela *mean*-ek *risk*-ek baino estimazio hobeak ematen dituela, hala ere, *risk*-ek sakabanapen txikiagoak izan ditu. Denbora-menpeko AUC mota honek, konputazionali aurrekoak baino kostu txikiagoa du. Horrez gain, parametro honek aukera ematen du denbora zehatz bat ezarri gabe balio bat bueltatzeko, C-index-a dena. Hori horrela izanik ere, iraupen laburreko biziraupena iragarri nahi denean komenigarriagoa da $\text{AUC}^{A/D}$ erabiltzea, izan ere, balio teorikotik alborapen txikiagoak lortu ditu.

Azkenik, 5. kapituluan doitu den ω_5 ereduari dagokionez, uste dugu lagintamaina handiagoa balitz eta $X_3(t)$ aldagaiarentzako 2 kategoria beharrean 4 kategoria sortu ahalko bagenitu, ereduaren diskriminazio-gaitasunak hobera egingo zuela. Horrez gain, eredu berean $X_6(t)$ eta $X_7(t)$ aldagaiak dituguneanean, $X_6(t)$ aldagaiaren koefizientea positiboa da eta medikoki hau gertatzeak ez du zentzu askorik. Gainera, bakarrik gertatzen da $X_7(t)$ aldagaia sartzen denean. Bi aldagai hauen artean interakzioa dagoen aztertu da baina ez da adierazgarria izan. Hala ere, interesgarria ikusten dugu etorkizunean *spline*-ekin interakzioak aztertzea. Honela, bi aldagai hauen artean antzematen den efektu bitxia era matematiko batean azaltzeko aukera izango genuke.

A. Eranskina

Biriketako gaixotasun buxatzaile kronikoari buruz

Biriketako gaixotasun buxatzaile kronikoa (BGBK; *enfermedad pulmonar obstructiva crónica*, EPOC, gazteleraz) aire-fluxuaren murrizpen progresiboa eragiten duen gaixotasun prebenigarri eta tratagarria da. Gas edo partikula narrigarrien aurrean (tabakoa bereziki) aire bideek eta birikek sortzen duten erantzun inflamatorioa izango da honen eragilea. Klinikoki 2 fenotipo bereizten dira: bronkitis kronikoa (aire bideen gaixotasun inflamatorio kronikoa) eta enfisema (birika parenkimaren suntsipena). Gaixotasun honetan ohikoak dira exazerbazioak eta gaixoen komorbilitateak lotuta daude gaixotasunaren larritasunarekin. Gaixoen komorbilitateak neurtzeko Charlson-en indizea (X_9) erabili ohi da.

Charlson-en indizeak (*Charlson Comorbidity Index*), indibiduo bakoitzak pairatzen dituen gaixotasunak neurtzen ditu. Gaixotasun bakoitzak pisu desberdina du, hau da, indize hau kalkulatzerakoan gaixotasun arinenek larriek baino pisu gutxiago dute eta indizeak zenbat eta balio altuagoak hartu, orduan eta gaixoago egongo da pazientea [22].

BGBK gaixoaren balorazioa egiteko batetik diagnostikoa eta bestetik arriskuen estratifikazioa egingo da. Diagnostikorako: arrisku faktoreak, arnas sintomatologia eta espirometria (bronkodilatazio osteko buxadura patroia, FEV1/FVC PBD $< 0,7$) eduki beharko ditugu. Bestalde, arriskuaren estratifikazioa egiteko: buxadura patroiaren larritasuna, disnea maila (MRC eskalaren bidez neurtuta) eta exazerbazioak kontuan hartuko dira [23].

Horrez gain, BODE (*The body-mass index, airflow obstruction, dyspnea, and exercise capacity index*) eskala pronostikoa existitzen da [21]. Azken hau biziraupen prediktorea da, pazientearen balorazio globala eta hilkortasun arriskuaren balorazioa egitea baimenduko duena. BODE eskalan GMI,

FEV1, disnea eta *Walking test*-a (6 minututan ahalik eta distantzia handiena egitea oinez), datu-basean X_3 aldagaia dena, neurtuko dira bakoitzaren emaitzen arabera.

BGBK gaixoek, exazerbazioak edukiko dituzte, hau da, ezegonkortasun klinikoko gertaera akutuak jasango dituzte. Hauek eragin negatiboa izango dute bizi-kalitatean, pronostikoan, birika narriaduran, hilkortasunean eta osasun publikorako kostua nabarmenki igoko dute.

BGBK-ak paziente askotan arnas gutxiegitasuna eragingo du, hau da, hipoxemia egoera. Ondorioz, ehunek, bereziki periferikoek oxigeno gutxiago edukiko dute ehun-hipoxia eraginez. Hipoxia egoera denboran mantenduz ehunen atrofia progresiboa gertatuko da. Hau era ez-zuzen batean pazienteen indar neurrien bidez neur daiteke, paziente bakoitzak erakusten duen masa-muskular edo indar galera gaixotasunaren larritasunarekin erlazionatuz. Gainera, indar neurrien neurketa prozedura azkarra eta ez-inbaditzailea da, kostu gutxikoa.

Hori dela eta, BGBK pazienteen indar neurriak neurtzea erabaki da gaixotasunaren eboluzioaren adierazle on bat izan daitekeen baloratzeko. Jasotako datu-basean 4 neurri ezberdin daude: esku indarra kg-tan neurtuta (X_4), koadrizepsaren indarra (X_5), *Lafayette* indar neurtzailearekin jasotako koadrizepsaren indarra hedaduran (X_6) eta koadrizepsaren indarra hedaduran *biodex* aparatuarekin neurtuta (X_7).

X_8, inspirazio presio maximoa, muskulu inspiratorioen indar globala neurtzen duen indizea da. Gaixotasun berarengatik zein muskulu inspiratorioen ahuleziagatik gaitzaren eboluzioan zehar inspirazio presio maximoa murriztuz joango da eta honekin batera arnas-bolumenak eta organismora heltzen den oxigeno bolumena murriztuz joango da gurpil-zoro egoera bat eraginez.

Horrez gain, aldagai guztietan Estatistika Deskribatzailea aplikatu da jasotako datu guztiak zuzenak zirela ziurtatzeko eta balio-galduak (NA) ondo identifikatu ahal izateko hurrengo emaitzak lortuz:

Oharra. Denbora-menpekoak diren aldagaien kasuan bakarrik lehenengo bisitan jaso diren datuen emaitzak adieraziko dira.

- **Aldagai kategorikoa:**

Aldagaia	**Kategoriak**	**NA**
$\mathbf{X_1}$	Gizonak: 293 (%73.43) Emakumeak: 106 (%26.57)	0

- **Aldagai jarraituak:**

Aldagaia	**Minimoa**	**Maximoa**	**Batez bestekoa (sd)**	**NA**
$\mathbf{X_2}$	31	83	64.16 (8.42)	0
$\mathbf{X_3}$	120	692	476.1 (108.16)	0
$\mathbf{X_4}$	10	64	32.4 (9.12)	0
$\mathbf{X_5}$	2	50	22.94 (9.1)	4 (%0.01)
$\mathbf{X_6}$	8.1	54.3	27.34 (6.94)	1 (%0.002)
$\mathbf{X_7}$	49.4	314.9	133.77 (50.21)	3 (%0.008)
$\mathbf{X_8}$	8.2	127	59.93 (21.92)	1 (%0.002)
$\mathbf{X_9}$	1	7	1.74 (1.06)	0

B. Eranskina

Biziraupen-denborak simulatzen

Literaturan aukera ezberdinak deskribatu dira CPH eredu bateko biziraupen-denborak simulatzeko. Gehienetan, h_0, oinarrizko arrisku funtzioa, banaketa ezagun bati darraion arrisku funtzioa dela asumitzen da eta hortik abiatuz, eta aldagaien eta koefizienteen balioak zehaztuz, biziraupen-denborak simulatzen dira. Denbora-menpeko aldagaiak kontuan hartu nahi direnean ez da hain erreza simulazio hauek egitea eta beraz, garapen matematiko sakonago bat egin behar da. [20] artikuluan modu bat deskribatzen da denbora-menpeko aldagaiak izanik biziraupen-denborak simulatzeko. Deskribatzen den formulan Lambert-en W funtzioa erabiltzen da.

Definizioa. Izan bedi $\begin{array}{rccc} f\colon & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) = x \cdot e^x \end{array}$ aplikazioa. *Lambert-en W funtzioa*, W bidez denotatuko duguna, f funtzioaren alderantzizkoa da. $W(x) \cdot e^{W(x)} = x$ eta $[-1/e, \infty)$ tartean definituta dago [24].

Biziraupen-denborak simulatzeko aurrefinkatu diren balioak ondorengoak dira:

- $X_1(t) = a + bt$ non $a \sim N(0, 2)$ eta $b \sim U(0, 0.2)$ diren.
- $X_2 \sim U(-1, 2)$.
- $X_3 \sim \exp(0.8)$.
- $\boldsymbol{\beta} = (\beta_1,\ \beta_2,\ \beta_3) = (0.6,\ 0.1,\ -0.35)$.
- h_0, oinarrizko arrisku funtzioa, $\lambda = 0.01$ parametroko banaketa esponentzialari darraion arrisku funtzioa da.

Biziraupen-denborak simulatzeko honako formula hau erabili da, $u \sim U(0, 1)$ izanik:

$$T = \frac{1}{\beta_1 b} W\left(\frac{-\beta_1 b \log(u)}{\lambda e^{\beta_1 a + \beta_2 X_2 + \beta_3 X_3}}\right) = \frac{1}{0.6b} W\left(\frac{-0.6b \log(u)}{0.01 e^{0.6a + 0.1X_2 - 0.35X_3}}\right)$$

C. Eranskina

R-ko kodea

```
# Kargatu behar diren pakete eta funtzioak:
library(survival) #ereduak doitzeko
library(survivalROC) #diskriminazio-gaitasuna neurtzeko
library(risksetROC) #diskriminazio-gaitasuna neurtzeko
library(lamW) #Lambert  W funtzioa erabili ahal izateko
# meanrankROC paketeko funtzioak kargatu
source("MeanRank.q")
source("NNE-estimate.q")
source("NNE-CrossValidation.q")
source("interpolate.q")
source("dynamicTP.q")
source("NNE-estimate_TPR.q")
source("dynamicIntegrateAUC.R")
# Harrell-en C-index-a kalkulatzeko funtzioa:
cindex.categorization <-
  function (x, y) {
    if (!is.Surv(y))
      y <- Surv(y)
    i <- is.na(x) | is.na(y)
    if (any(i)) {
      x <- x[!i]
      y <- y[!i, ]
    }
    k <- survConcordance.fit(y, x)
    cindex <- (k[1] + k[3]/2)/sum(k[1:3])
    cindex
  }

# Simulazioak ----------------------------------------

## Datu-basea simulatu
set.seed(2598)
N <- 10000
```

```
#Aldagaiak sortu
#X1(t)=a+bt
a <- rnorm(N, mean = 0, sd=2)
b <- runif(N, min=0, max = 0.2)
X2 <- runif(N, min=-1, max=2)
X3 <- rexp(N,0.8)
betak <- c(0.6, 0.1, -0.35) #koefizienteen bektorea

#Biziraupen denborak simulatu
u <- runif(N)
lambda <- 0.01
Tl <- lambertW0((-betak[1]*b*log(u))/(lambda*exp(
  betak[2]*X2+betak[3]*X3+betak[1]*a)))/(betak[1]*b)

#Datu-basea sortu X1 DM
simdb<- data.frame( id= 1:N,t0=rep(0,N),t1=rep(12,N),
                    t2=rep(24,N),t3=rep(36,N),
                    t4=rep(48,N), tfal=Tl,fal=rep(1,N),
                    X1_0=a, X1_1=a+b*12, X1_2=a+b*24,
                    X1_3=a+b*36,X1_4=a+b*48,
                    X2=X2, X3=X3)

simdbt<-simdb[,c(1,14,15)]
simdbt <- tmerge(simdbt, simdb, id=id,
                 fal= event(tfal,fal))
simdbt <- tmerge(simdbt, simdb, id=id, X1=tdc(t0,X1_0))
simdbt <- tmerge(simdbt, simdb, id=id, X1=tdc(t1,X1_1))
simdbt <- tmerge(simdbt, simdb, id=id, X1=tdc(t2,X1_2))
simdbt <- tmerge(simdbt, simdb, id=id, X1=tdc(t3,X1_3))
simdbt <- tmerge(simdbt, simdb, id=id, X1=tdc(t4,X1_4))

#Datu-basea sortu X1 basala
simdb1 <- data.frame(id= 1:N,tstart=rep(0,N), tstop=Tl,
                     fal=rep(1,N), X1=a, X2=X2, X3=X3)

## Simulazioak egin

simulazioa <- function(datuaks,datuakt, nn, tzents,
                       B=500, tart){
  mat_dg <- matrix(nrow = B, ncol = 38)
  tarteak <- c(0, 1, 2, 3)*tart
  tarteluzera <- 1
  bandwidths <- 0.05 + c(1:80)/200
  IMSEs <- vector(length=length(bandwidths))

  set.seed(2917)
  seeds <- round(runif(B)*10000)
  for(b in 1:B){
    set.seed(seeds[b])
```

```
# 1) Lagindu
id_sim <- sample(unique(datuaks$id), size = nn,
                 replace = FALSE)
datsim <- NULL
datsimt <- NULL
for(j in 1:length(id_sim)){
  datsim <- rbind(datsim,
                  datuaks[which(
                    datuaks$id==id_sim[j]),])
  datsimt <- rbind(datsimt,
                   datuakt[which(
                     datuakt$id==id_sim[j]),])}
# zentsuratu
datsim$fal <- as.numeric(datsim$tstop <=tzents)
datsim$tstop <- pmin(datsim$tstop,tzents)

for(i in 1:length(datsimt[,1])){
  if(datsimt$tstart[i]==tail(tarteak,n=1)+tart
     & datsimt$tstop[i]>tzents){
    datsimt$fal[i] <- 0
    datsimt$tstop[i] <- tzents
  }
}

# 2) Ereduak doitu

ers <- coxph(Surv(tstop,fal) ~ X1+X2+X3, data=datsim)
datsim$marker_b <- predict(ers, type = 'lp')

ert <- coxph(Surv(tstart,tstop,fal) ~ X1+X2+X3,
             data=datsimt)
datsimt$marker_t <- predict(ert, type = 'lp')


# 3) Diskriminazio gaitasuna

#C-index Harrell
mat_dg[b,1] <-  ers$concordance[6]

mat_dg[b,1+19] <-  ert$concordance[6]

#AUC A/D
for(j in 1:length(tarteak)) {
  datDG <- subset(datsim, tstop >= (tarteak[j]))

  datDGt <- subset(datsimt, tstart <= (tarteak[j])
                   & tstop > (tarteak[j]))

  behkop <- nrow(datDG)
```

```
out1 <- survivalROC(
  Stime=datDG$tstop, status=datDG$fal,
  marker= datDG$marker_b,
  predict.time = (tarteak[j] + tarteluzera*tart),
  method = "NNE", span = 0.04 * behkop^(-0.2))
mat_dg[b,1+j] <- out1$AUC

behkop <- nrow(datDGt)

out1 <- survivalROC(
  Stime=datDGt$tstop, status=datDGt$fal,
  marker= datDGt$marker_t ,
  predict.time = (tarteak[j] + tarteluzera*tart),
  method = "NNE", span = 0.04 * behkop^(-0.2))
mat_dg[b,1+j+19] <- out1$AUC

#AUC A/D KM

out1 <- survivalROC(
  datDG$tstop, datDG$fal, marker= datDG$marker_b,
  predict.time = (tarteak[j] + tarteluzera*tart),
  method = "KM")
mat_dg[b,5+j] <- out1$AUC


out1 <- survivalROC(
  Stime=datDGt$tstop, status=datDGt$fal,
  marker= datDGt$marker_t ,
  predict.time = (tarteak[j] + tarteluzera*tart),
  method = "KM")
mat_dg[b,5+j+19] <- out1$AUC


# AUC G/D risk

out1 <- risksetROC(
  Stime = datsim$tstop, entry = datsim$tstart,
  status = datsim$fal, marker = datsim$marker_b,
  predict.time = tarteak[j], plot = F)
mat_dg[b,9+j] <- out1$AUC


out1 <- risksetROC(
  Stime = datsimt$tstop, entry = datsimt$tstart,
  status = datsimt$fal, marker = datsimt$marker_t,
  predict.time = tarteak[j], plot = F)
mat_dg[b,9+j+19] <- out1$AUC
```

```
}
# AUC G/D mean
mmm <- MeanRank(survival.time = datsim$tstop,
                start = datsim$tstart,
                survival.status = datsim$fal,
                marker = datsim$marker_b )
for(j in 1:length(bandwidths)) {
  nnnC <- nne.CrossValidate(x=mmm$time,
  y=mmm$mean.rank, lambda=bandwidths[j])
  IMSEs[j] <- nnnC$IMSE
}

currLambdaOS <- mean(
    bandwidths[which(IMSEs==min(IMSEs, na.rm=T))])

nnn <- nne(
    x= mmm$time, y= mmm$mean.rank,
    lambda=currLambdaOS, nControls=mmm$nControls)

mat_dg[b,14:17] <- sapply(
    tarteak,
    function(x){ interpolate(x = nnn$x, y=nnn$nne,
                                target=x)})

mmm <- MeanRank(survival.time = datsimt$tstop,
                start = datsimt$tstart,
                survival.status = datsimt$fal,
                marker = datsimt$marker_t )
for(j in 1:length(bandwidths)) {
  nnnC <- nne.CrossValidate(x=mmm$time,
  y=mmm$mean.rank, lambda=bandwidths[j])
  IMSEs[j] <- nnnC$IMSE
}

currLambdaOS <- mean(
    bandwidths[which(IMSEs==min(IMSEs, na.rm=T))])

nnn <- nne(
    x= mmm$time, y= mmm$mean.rank,
    lambda=currLambdaOS, nControls=mmm$nControls)

mat_dg[b,33:36] <- sapply(
    tarteak,
    function(x){ interpolate(x = nnn$x, y=nnn$nne,
                                target=x)})


# C-index risk
```

```
    cind <- risksetAUC(
        Stime = datsim$tstop, status = datsim$fal,
        marker = datsim$marker_b, tmax=tzents, plot = F)
    mat_dg[b,18] <- cind$Cindex

    cind <- risksetAUC(
        Stime = datsimt$tstop, entry = datsimt$tstart,
        status = datsimt$fal, marker = datsimt$marker_t,
        tmax=tzents, plot = F)
    mat_dg[b,37] <- cind$Cindex

    # C-index meanrankROC

    mat_dg[b,19] <- dynamicIntegrateAUC(
      survival.time = datsim$tstop,
      survival.status = datsim$fal,
      marker = datsim$marker_b, cutoffTime = tzents)

    mat_dg[b,38] <- dynamicIntegrateAUC(
      survival.time = datsimt$tstop,
      start = datsimt$tstart,
      survival.status = datsimt$fal,
      marker = datsimt$marker_t,
      cutoffTime = tzents)


  }
  data.frame(mat_dg)
}


sim_500 <- simulazioa(simdb1, simdbt, nn=500,
                      tzents = 120, tart = 12)
sim_1000 <- simulazioa(simdb1, simdbt, nn=1000,
                       tzents = 120, tart = 12)

# Konfinatza tarteen kalkulua (bootstrap) ---------------

B <- 500

mat_kt <- matrix(nrow = B, ncol = 19*5)
colnames(mat_kt) <- rep(c('C␣Harrell', 'AUC␣C/D␣(0,1)',
                          'AUC␣C/D␣(1,2)',
                          'AUC␣C/D␣(2,3)',
                          'AUC␣C/D␣(3,4)',
                          'AUC␣C/D␣KM␣(0,1)',
                          'AUC␣C/D␣KM␣(1,2)',
                          'AUC␣C/D␣KM␣(2,3)',
```

```
                              'AUC␣C/D␣KM␣(3,4)',
                              'AUC␣G/D␣t=0␣risk',
                              'AUC␣G/D␣t=1␣risk',
                              'AUC␣G/D␣t=2␣risk',
                              'AUC␣G/D␣t=3␣risk',
                              'AUC␣G/D␣t=0␣mean',
                              'AUC␣G/D␣t=1␣mean',
                              'AUC␣G/D␣t=2␣mean',
                              'AUC␣G/D␣t=3␣mean',
                              'C␣risk', 'C␣mean'), 5)
egunak <- 365.25

tarteak <- c(0, 1, 2, 3)*egunak

tarteluzera <- 1

bandwidths <- 0.05 + c(1:80)/200
IMSEs <- vector(length=length(bandwidths))

set.seed(13)
seeds <- round(runif(B)*10000)
for (b in 1:B){
  set.seed(seeds[b])

  id_boot <- sample(
    unique(db11$ID), size = length(unique(db11$ID)),
    replace =T)
  boot_1 <- NULL
  boot_t <- NULL
  for(j in 1:length(id_boot)){
    boot_1 <- rbind(boot_1, db11[which(
      db11$ID==id_boot[j]),])
    boot_t <- rbind(boot_t, dbt1[which(
      dbt1$ID==id_boot[j]),])}

  boot_1$marker_e1 <- predict(e1, type = 'lp',
                              newdata = boot_1)
  boot_t$marker_e2 <- predict(e2, type = 'lp',
                              newdata = boot_t)
  boot_t$marker_e3 <- predict(e3, type = 'lp',
                              newdata = boot_t)
  boot_t$marker_e4 <- predict(e4, type = 'lp',
                              newdata = boot_t)
  boot_t$marker_e5 <- predict(e5, type = 'lp',
                              newdata = boot_t)


  #C-index Harrell
  mat_kt[b,1] <-  cindex.categorization(
```

```
    boot_1$marker_e1, Surv(boot_1$TEND, boot_1$FAL))
  mat_kt[b,1+19] <- cindex.categorization(
    boot_t$marker_e2, Surv(boot_t$tstart,boot_t$tstop,
                           boot_t$hil))
  mat_kt[b,1+19*2] <- cindex.categorization(
    boot_t$marker_e3, Surv(boot_t$tstart,boot_t$tstop,
                           boot_t$hil))
  mat_kt[b,1+19*3] <- cindex.categorization(
    boot_t$marker_e4, Surv(boot_t$tstart,boot_t$tstop,
                           boot_t$hil))
  mat_kt[b,1+19*4] <- cindex.categorization(
    boot_t$marker_e5, Surv(boot_t$tstart,boot_t$tstop,
                           boot_t$hil))


  #AUC A/D
  for(j in 1:length(tarteak)) {
    datBas <- subset(boot_1, TEND >= (tarteak[j]))
    datDM <- subset(boot_t, tstart <= (tarteak[j])
                    & tstop > (tarteak[j]))

    behkop <- nrow(datBas)
    out1 <- survivalROC(
      datBas$TEND, datBas$FAL, marker= datBas$marker_e1,
      predict.time = (tarteak[j] + tarteluzera*egunak),
      method = "NNE", span = 0.04 * behkop^(-0.2))
    mat_kt[b,1+j] <- out1$AUC

    behkop <- nrow(datDM)

    out1 <- survivalROC(
      datDM$tstop, datDM$hil, marker= datDM$marker_e2,
      predict.time = (tarteak[j] + tarteluzera*egunak),
      method = "NNE", span = 0.04 * behkop^(-0.2))
    mat_kt[b,1+j+19] <- out1$AUC

    out1 <- survivalROC(
      datDM$tstop, datDM$hil, marker= datDM$marker_e3,
      predict.time = (tarteak[j] + tarteluzera*egunak),
      method = "NNE", span = 0.04 * behkop^(-0.2))
    mat_kt[b,1+j+19*2] <- out1$AUC

    out1 <- survivalROC(
      datDM$tstop, datDM$hil, marker= datDM$marker_e4,
      predict.time = (tarteak[j] + tarteluzera*egunak),
      method = "NNE", span = 0.04 * behkop^(-0.2))
    mat_kt[b,1+j+19*3] <- out1$AUC

    out1 <- survivalROC(
```

```
    datDM$tstop, datDM$hil, marker= datDM$marker_e5,
    predict.time = (tarteak[j] + tarteluzera*egunak),
    method = "NNE", span = 0.04 * behkop^(-0.2))
mat_kt[b,1+j+19*4] <- out1$AUC

#AUC A/D KM

out1 <- survivalROC(
    datBas$TEND, datBas$FAL, marker= datBas$marker_e1,
    predict.time = (tarteak[j] + tarteluzera*egunak),
    method = "KM" )
mat_kt[b,5+j] <- out1$AUC

out1 <- survivalROC(
    datDM$tstop, datDM$hil, marker= datDM$marker_e2,
    predict.time = (tarteak[j] + tarteluzera*egunak),
    method = "KM")
mat_kt[b,5+j+19] <- out1$AUC

out1 <- survivalROC(
    datDM$tstop, datDM$hil, marker= datDM$marker_e3,
    predict.time = (tarteak[j] + tarteluzera*egunak),
    method = "KM")
mat_kt[b,5+j+19*2] <- out1$AUC

out1 <- survivalROC(
    datDM$tstop, datDM$hil, marker= datDM$marker_e4,
    predict.time = (tarteak[j] + tarteluzera*egunak),
    method = "KM")
mat_kt[b,5+j+19*3] <- out1$AUC

out1 <- survivalROC(
    datDM$tstop, datDM$hil, marker= datDM$marker_e5,
    predict.time = (tarteak[j] + tarteluzera*egunak),
    method = "KM")
mat_kt[b,5+j+19*4] <- out1$AUC

# AUC G/D risk

out1 <- risksetROC(S
    time = boot_1$TEND, status = boot_1$FAL,
    marker = boot_1$marker_e1,
    predict.time = tarteak[j], plot=F)
mat_kt[b,9+j] <- out1$AUC

out1 <- risksetROC(
    Stime = boot_t$tstop, entry = boot_t$tstart,
    status = boot_t$hil, marker = boot_t$marker_e2,
    predict.time = tarteak[j], plot = F)
```

```
    mat_kt[b,9+j+19] <- out1$AUC

    out1 <- risksetROC(
        Stime = boot_t$tstop, entry = boot_t$tstart,
       status = boot_t$hil, marker = boot_t$marker_e3,
        predict.time = tarteak[j], plot = F)
    mat_kt[b,9+j+19*2] <- out1$AUC

    out1 <- risksetROC(
        Stime = boot_t$tstop, entry = boot_t$tstart,
        status = boot_t$hil, marker = boot_t$marker_e4,
        predict.time = tarteak[j], plot = F)
    mat_kt[b,9+j+19*3] <- out1$AUC

    out1 <- risksetROC(
        Stime = boot_t$tstop, entry = boot_t$tstart,
        status = boot_t$hil, marker = boot_t$marker_e5,
        predict.time = tarteak[j], plot = F)
    mat_kt[b,9+j+19*4] <- out1$AUC

  }
  # AUC G/D mean

  #Eredua 1

  mmm <- MeanRank(
    survival.time = boot_1$TEND,
    survival.status = boot_1$FAL,
    marker = boot_1$marker_e1)

  for(j in 1:length(bandwidths)) {
    nnnC <- nne.CrossValidate(
      x=mmm$time, y=mmm$mean.rank, lambda=bandwidths[j])
    IMSEs[j] <- nnnC$IMSE
  }

  currLambdaOS <- mean(bandwidths[which(
    IMSEs==min(IMSEs, na.rm=T))])

  nnn <- nne(x= mmm$time, y= mmm$mean.rank,
             lambda=currLambdaOS,nControls=mmm$nControls)
  mat_kt[b,14:17] <- sapply(
    tarteak,
    function(x){interpolate(x = nnn$x, y=nnn$nne,
                            target=x)})

  #Eredua2
  mmm <- MeanRank(
    survival.time = boot_t$tstop, start = boot_t$tstart,
```

```
  survival.status = boot_t$hil,
  marker = boot_t$marker_e2)


for(j in 1:length(bandwidths)) {
  nnnC <- nne.CrossValidate(
    x=mmm$time, y=mmm$mean.rank, lambda=bandwidths[j])
  IMSEs[j] <- nnnC$IMSE
}

currLambdaOS <- mean(bandwidths[which(
  IMSEs==min(IMSEs, na.rm=T))])

nnn <- nne(x= mmm$time, y= mmm$mean.rank,
           lambda=currLambdaOS,nControls=mmm$nControls)
mat_kt[b,33:36] <- sapply(
  tarteak,
  function(x){interpolate(x = nnn$x, y=nnn$nne,
                          target=x)})

#eredua 3
mmm <- MeanRank(
  survival.time = boot_t$tstop, start = boot_t$tstart,
  survival.status = boot_t$hil,
  marker = boot_t$marker_e3)

for(j in 1:length(bandwidths)) {
  nnnC <- nne.CrossValidate(
    x=mmm$time, y=mmm$mean.rank, lambda=bandwidths[j])
  IMSEs[j] <- nnnC$IMSE
}

currLambdaOS <- mean(bandwidths[which(
  IMSEs==min(IMSEs, na.rm=T))])

nnn <- nne(x= mmm$time, y= mmm$mean.rank,
           lambda=currLambdaOS,nControls=mmm$nControls)
mat_kt[b,52:55] <- sapply(
  tarteak,
  function(x){interpolate(x = nnn$x, y=nnn$nne,
                          target=x)})

#eredua 4
mmm <- MeanRank(
  survival.time = boot_t$tstop, start = boot_t$tstart,
  survival.status = boot_t$hil,
  marker = boot_t$marker_e4)

for(j in 1:length(bandwidths)) {
```

```
  nnnC <- nne.CrossValidate(
    x=mmm$time, y=mmm$mean.rank, lambda=bandwidths[j])
  IMSEs[j] <- nnnC$IMSE
}

currLambdaOS <- mean(bandwidths[which(
  IMSEs==min(IMSEs, na.rm=T))])

nnn <- nne(x= mmm$time, y= mmm$mean.rank,
           lambda=currLambdaOS,nControls=mmm$nControls)
mat_kt[b,71:74] <- sapply(
  tarteak,
  function(x){interpolate(x = nnn$x, y=nnn$nne,
                          target=x)})

#eredua 5
mmm <- MeanRank(
  survival.time = boot_t$tstop, start = boot_t$tstart,
  survival.status = boot_t$hil,
  marker = boot_t$marker_e5)

for(j in 1:length(bandwidths)) {
  nnnC <- nne.CrossValidate(
    x=mmm$time, y=mmm$mean.rank, lambda=bandwidths[j])
  IMSEs[j] <- nnnC$IMSE
}

currLambdaOS <- mean(bandwidths[which(
  IMSEs==min(IMSEs, na.rm=T))])

nnn <- nne(x= mmm$time, y= mmm$mean.rank,
           lambda=currLambdaOS,nControls=mmm$nControls)
mat_kt[b,90:93] <- sapply(
  tarteak,
  function(x){interpolate(x = nnn$x, y=nnn$nne,
                          target=x)})

# C-index risk

cind <- risksetAUC(
  Stime = boot_1$TEND, status = boot_1$FAL,
  marker = boot_1$marker_e1, plot=F,tmax = egunak*5)
mat_kt[b,18] <- cind$Cindex

cind <- risksetAUC(
  Stime = boot_t$tstop, entry = boot_t$tstart,
  status = boot_t$hil, marker = boot_t$marker_e2,
  tmax=egunak*5, plot = F)
mat_kt[b,37] <- cind$Cindex
```

```
  cind <- risksetAUC(
    Stime = boot_t$tstop, entry = boot_t$tstart,
    status = boot_t$hil, marker = boot_t$marker_e3,
    tmax=egunak*5, plot = F)
  mat_kt[b,56] <- cind$Cindex

  cind <- risksetAUC(
    Stime = boot_t$tstop, entry = boot_t$tstart,
    status = boot_t$hil, marker = boot_t$marker_e4,
    tmax=egunak*5, plot = F)
  mat_kt[b,75] <- cind$Cindex

  cind <- risksetAUC(
    Stime = boot_t$tstop, entry = boot_t$tstart,
    status = boot_t$hil, marker = boot_t$marker_e5,
    tmax=egunak*5, plot = F)
  mat_kt[b,94] <- cind$Cindex

  # C-index mean

  mat_kt[b,19] <- dynamicIntegrateAUC(
    survival.time = boot_1$TEND,
    survival.status = boot_1$FAL,
    marker = boot_1$marker_e1, cutoffTime = egunak*5)

  mat_kt[b,38] <- dynamicIntegrateAUC(
    survival.time = boot_t$tstop, start = boot_t$tstart,
    survival.status = boot_t$hil,
    marker = boot_t$marker_e2, cutoffTime = egunak*5)

  mat_kt[b,57] <- dynamicIntegrateAUC(
    survival.time = boot_t$tstop, start = boot_t$tstart,
    survival.status = boot_t$hil,
    marker = boot_t$marker_e3, cutoffTime = egunak*5)

  mat_kt[b,76] <- dynamicIntegrateAUC(
    survival.time = boot_t$tstop, start = boot_t$tstart,
    survival.status = boot_t$hil,
    marker = boot_t$marker_e4, cutoffTime = egunak*5)

  mat_kt[b,95] <- dynamicIntegrateAUC(
    survival.time = boot_t$tstop, start = boot_t$tstart,
    survival.status = boot_t$hil,
    marker = boot_t$marker_e5, cutoffTime = egunak*5)
}

kt <- round(apply(mat_kt, 2, quantile,
                  probs=c(0.025,0.975), na.rm=TRUE),2)
```

Bibliografia

[1] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972.

[2] DR Cox eta D Oakes. *Analysis of Survival Data*. Chapman and Hall/CRC, 1984.

[3] FE Harrel. *Regression Modeling Strategies*. Springer, 2015.

[4] DF Moore. *Applied Survival Analysis Using R*. Springer, 2016.

[5] DG Kleinbaum eta M Klein. *The Cox Proportional Hazards Model and Its Characteristics*. Springer, 2012.

[6] M Sestelo. A short course on survival analysis applied to the financial industry, 2017.

[7] Z Zhang, J Reinikainen, K Adedayo Adeleke, ME Pieterse eta CGM Groothuis-Oudshoorn. Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine*, 6(7), 2018.

[8] T Therneau, C Crowson eta E Atkinson. Using time dependent covariates and time dependent coefficients in the cox model. *Survival Vignettes*, 2(3):1–30, 2023.

[9] H Heinzl eta A Kaider. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. C*omputer methods and programs in biomedicine*, 54(3):201-208, 1997.

[10] E Longato, M Vettoretti eta B Di Camillo. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, 108:103496, 2020.

[11] A Alabdallah, M Ohlsson, S Pashami eta T Rögnvaldsson. The concordance index decomposition – A measure for a deeper understanding of survival prediction models. *arXiv preprint arXiv:2203.00144*, 2022.

[12] FE Harrell, RM Califf, DB Pryor, KL Lee eta RA Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 1982.

[13] MS Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

[14] D Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415, 1975.

[15] P Blanche, JF Dartigues eta H Jacqmin-Gadda. Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5):687–704, 2013.

[16] A Bansal eta PJ Heagerty. A tutorial on evaluating the time-varying discrimination accuracy of survival models used in dynamic decision making. *Medical Decision Making*, 38(8):904–916, 2018.

[17] L Chambless eta G Diao. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*, 25:3474–3486, 2006.

[18] A Bansal eta PJ Heagerty. A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. *Diagnostic and prognostic research*, 3:1–13, 2019.

[19] B Efron eta RJ Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1993.

[20] JS Ngwa, HJ Cabral, DM Cheng, DR Gagnon, MP LaValley eta LA Cupples. Generating survival times with time-varying covariates using the Lambert W function. *Communications in Statistics-Simulation and Computation*, 51(1):135-153, 2019.

[21] BR Celli, CG Cote, JM Marin, C Casanova, M Montes de Oca, RA Mendez, V Pinto Plata eta HJ Cabral. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *New England Journal of Medicine*, 350(10):1005–1012, 2004.

[22] ME Charlson, P Pompei, KL Ales eta CR MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, 40(5):373-383, 1987.

[23] CR Borstnar eta F Cardellach. *Farreras Rozman. Medicina Interna*. Elsevier Health Sciences, 2020.

[24] TP Dence. A Brief Look into the Lambert W Function, *Scientific Research Publishing*, 4:887-892, 2013.

[25] L Antolini, P Boracchi eta E Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

[26] S Cygu, H Seow, J Dushoff eta BM Bolker. Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. *Scientific Reports*, 13(1):1370, 2023.