Inputazio tekniken errendimenduaren ebaluazioa bi neurketako luzeranzko datuetan

Urko Aguirre[1], *Inmaculada Arostegui*[2], *Jose M. Quintana*[1]

[1] Ikerkuntza Unitatea, Galdakao-Usansolo Ospitalea.
Red de Investigación en Servicios Sanitarios en Enfermedades Crónicas (REDISSEC)

[2] Matematika Aplikatua, Estatistika eta Ikerkuntza Operatiboa Saila, Euskal Herriko Unibertsitatea (UPV/EHU).
Red de Investigación en Servicios Sanitarios en Enfermedades Crónicas (REDISSEC)

Jasota: 30/05/2013 Onartua: 29/11/2013

Laburpena: Neurketa errepikatuetan oinarrituriko behaketa-ikerketak menpeko aldagaien aldaketak denboran zehar aztertzeko erabiltzen dira. Bi neurketa baizik bakarrik egiten ez direnean, ikerketa helburu nagusietariko bat izan daiteke menpeko aldagaiaren batez besteko aldaketa aurresaten dituzten faktoreak zehaztea. Menpeko aldagaian faltako balioak ohikoak dira ikerketa mota hauetan, behaturiko datuen analisiaren emaitzak alboratuak gerta daitezkeelarik. Lan honetan inputazio teknika desberdinak proposatuko ditugu datu-analisiak egiterakoan faltako balioei aurre egiteko aukera gisa. Hiru inputazio metodoren errendimendua aztertu dugu (*K-Nearest Neighbor*, *Propensity Score* eta *Markov Chain Monte Carlo* algoritmoak), faltako balioek datu multzo osoaren % 10a eta % 30a osatzen dutenean.

Gako hitzak: inputazioa; faltako balioak; bizi-kalitatea.

Abstract: Observational studies based on repeated measures are often used to assess the evolution of an outcome over time. When there are only two measurements one motivation of the research may be to focus on the determination of the potential predictors of the mean change of the outcome of interest. It is very common in such studies for data to be missing, which can bias the results. Different imputation techniques have been proposed as alternatives to cope with missing data. We compared three imputation methods (K-Nearest Neighbor, Propensity Score and a Markov Chain Monte Carlo algorithm) to assess their performance when handling missing data at different missingness rates (10% and 30%).

Keywords:

1. **SARRERA**

Faltako balioak agertzea arazo arrunta da ikerketa mota batzuetan [1]. Batez ere, hala da neurketa errepikatuetan oinarrituriko luzeranzko datuetan, non informazio eza garrantzizkoa gerta daitekeen.

Esate baterako, medikuntza arloko hainbat ikerketak helburu modura hartzen dute pazienteen egoera soziodemografikoa edo klinikoaren eta gaixotasunaren bilakaeraren arteko lotura aztertzea. Beraz, hau egiteko, ikerketa hauek gutxienez bi neurketa behar dituzte aldagai hauen arteko lotura aurkitzeko: lehendabizikoa, ikerketa hasieran eta bigarrena, denbora jakin bat igarota edo interbentzio baten ostean. Denboran zehar bildutako informazio honen ondorioz, maiz gertatzen da bigarren neurketan hainbat faltako balio izatea Informazio-galera hau, pazientearen bilakaerarekin, heriotzarekin edo beste arrazoi batekin uztar daiteke eta sarritan % 30 - % 40ko portzentajetara hel daiteke.

Ondorioz, ikerketan agerturiko faltako balioak pazienteen eboluzio bilakaerari egotzi ahal zaie, besteak beste, eta beraz, ikerketa osoko datu guztiak behaturik dituzten pazienteak ez dira aztertzen ari den populazioaren adierazgarri. Hau dela eta, ondorengo arazoak gerta daitezke: (a) ahalmen eta eraginkortasun galera, lagin tamainaren murrizketa dela eta; (b) datu-analisiaren zailtasuna; (d) alborapena presentzia lorturiko emaitzetan [2]. Ikerketa askotan informazio-galera hau kontuan hartzen ez delarik, datu guztiez osaturikoa azpimultzoa baizik ez da aztertzen, eta lorturiko emaitzen adierazgarritasuna kolokan jartzen da jarriz. Beraz, egungo estatistikak erronka modura hartu du egoera honen aurrean populazio guztiaren adierazgarria den laginaren datu-analisia egiteko egokiak diren estatistika metodoak proposatzea eta erabiltzea.

Ikerketa lerro honetan oso lan handia egin da azkeneko bizpahiru hamarkadetan ([3,4,5]). Little eta Rubinek [4] hiru profiletan sailkatu zituzten esparru honetako faltako balioak: informazio ezak menpeko aldagaiarekin zerikusia ez duenean, faltako balioei *missing completely at random* (*MCAR*) deritze; faltako balioak behaturiko emaitza edota faktore aske edo aurresaleekin lotura dutenean, informazio ezari *missing at random* (*MAR*) deritzo eta, horrez gain, behaturik gabe dauden datuekin lotura baldin badago, informazio ezari *missing not at random* (*MNAR*) deritzo. Hauek dira faltako balioen inguruan ageri den oinarrizko sailkapena edo profilak.

Faltako balioak daudenean, hainbat metodo daude datu-analisia egiteko. Modurik errazenean, analisitik kanpo uzten dira aldagairen batean balioren bat faltan daukaten indibiduo guztiak.

Era honetan, aztertutako aldagaietan informazioa osoa dutenen behaketekin osatutako datu multzoa baizik ez dugu kontuan hartuko kontuan hartuko dugu soilik. Metodo hau *Complete Case* (*CC*) deitzen da. Badago

beste metodo bat, *Available Case* (*AC*) izenekoa. Bigarren honetan, behaturiko informazio guztia erabiltzen da, indibiduo bakoitzerako behaturiko guztiaz baliatuz; horrela lagin tamaina maximizatu egiten da. maximizatuz. Datuak inputatzea da beste aukera bat. Inputazioak behatuta dauden datuetatik hartutako zenbakizko balioak jartzen ditu faltako balioen ordez. Arlo honetan agerturiko lan batzuetan, azpimarratzen da hobe dela inputatzea ezer ez egitea baino [6]. Lan honek helburu modura hartu du hainbat inputazio teknikaren errendimendua aztertzea simulazio ikerketa baten bitartez. Simulazioak egiteko, kontuan hartu dira aurretik aipatutako faltako balioen hiru profilak eta % 10 eta % 30eko portzentajeak. Inputazio metodoen artean, *K-nearest neighbor imputation* (*K-NNI*), *propensity score* (*PS*) eta *Markov Chain Monte Carlo* (*MCMC*) teknikak dira ezagunenetariko batzuk, eta hiru horiek dira erakutsiko ditugunak.

Lan hau ondoko eran banatuko da: bigarren atalean, zehazkiago azalduko dira faltako balioen profilak eta erabilitako inputazio metodoak; hirugarren atalean simulazio ikerketa aurkeztuko da, eta ondoren (laugarren atalean) emaitzak. Azkenik, lorturiko emaitzen eztabaida eta ondorioak aurkeztuko ditugu.

2. METODOLOGIA

2.1. Faltako balioen profilak

Atal honetan faltako balioak sortzen dituen mekanismo batzuk azalduko ditugu. Horren arabera, faltako balioen sailkapenak edo profilak definituko ditugu. Izan ere, ezinbestekoa da sailkapen bat faltako balioak dauden luzeranzko datuen azterketa estatistikoa sakontasunean ulertzeko.

Egoera orokorrean, demagun $\boldsymbol{Y} = \{Y_{ij}\}_{i=1,\ldots,n}^{j=1,\ldots,n_i}$ ikerketan parte hartzen duten unitate bakoitzerako ($i = 1,\ldots n$), behaturiko menpeko aldagaiaren $j = 1,\ldots,n_i$ neurketen segida dugula. Horrez gain, ondorengo $\boldsymbol{R}_{ij} = \{R_{ij}\}_{i=1,\ldots,n}^{j=1,\ldots,n_i}$ aldagai adierazlea definitzen dugu:

$$R_{ij} = \begin{cases} 1 & y_{ij} \text{ behaturik badago} \\ 0 & y_{ij} \text{ faltako balioa bada} \end{cases}.$$

Faltako balioak monotonoak edo aldakorrak izan daitezke. Hain zuzen, k neurketa bat izanik, informazio eza monotonoa dela esango dugu, baldin eta $R_{ij} = 0$, $\forall\ j \geq k$. Bestelako kasuetan, ezmonotonoa edo aldakorra dela onartuko da.

Lan honetan $j = 2$ kasu zehatzera eta faltako balio monotonoetara mugatuko gara, egindako lana bi neurketa dauden datu luzeranzkoen azterketan oinarritzen baita.

Orokorrean, ($\boldsymbol{Y}$, $\boldsymbol{R}$) datu base osoa adierazten duen zorizko bektorearen probabilitate-banaketa ondoko baterako dentsitate funtzioaren bidez dator definiturik: $f(\boldsymbol{Y}, \boldsymbol{R}|\theta, \varphi) = f(\boldsymbol{Y}|\theta) \cdot f(\boldsymbol{R}|\boldsymbol{Y}, \varphi)$, non φ eta θ parametro ezezagunak diren.

Faltako balioen mekanismoa Y bektoreak baldintzaturiko $\boldsymbol{R}$-ren banaketak zehazten du, hain zuzen, $f(\boldsymbol{R}|\boldsymbol{Y}, \varphi)$, non φ parametro ezezaguna den. Faltako balioak Y erantzun aldagaiaren menpekoak ez badira, ez behatu gabeko aleentzat ezta behaturiko aleentzat, orduan $f(\boldsymbol{R}|\boldsymbol{Y}, \varphi) = f(\boldsymbol{R}|\varphi)$ izango da, edozein $\boldsymbol{Y}$ eta φ-ren balioetarako. Kasu honetan, faltako balioak guztiz zorizkoak direla esango dugu, eta *MCAR* (*missing completely at random*) deritze. Schaferrenean [7] oinarrituz, egin daiteke $\boldsymbol{Y} = (Y_{obs}, Y_{mis})$ deskonposaketa, non Y_{obs} eta Y_{mis}, hurrenez hurren behaturiko eta faltako balioez osaturiko osaturikoa azpibektoreak diren. Deskonposaketa hau erabiliz, aurrekoa aina murriztailea ez den baldintza jar dezakegu faltako balioen gainean. Demagun faltako balioak $\boldsymbol{Y}$-ren Y_{obs} azpibektorearen menpekoak soilik direla, eta ez behatu ez den Y_{mis} azpibektorearen azpi-bektorearen menpekoak. Hau da, $f(\boldsymbol{R}|\boldsymbol{Y}, \varphi) = f(\boldsymbol{R}|Y_{obs}, \varphi)$, edozein Y_{mis} eta φ -ren balioetarako. Kasu honetan, faltako balioak zorizkoak direla esango dugu, eta *MAR* (*missing at random*) esango diegu. Azkenik, esango dugu faltako balioak ez direla zorizkoak, baldin eta $\boldsymbol{R}$-ren banaketa Y_{obs} eta Y_{mis} azpibektoreen menpekoa bada. Honelakoetan, faltako balioei *MNAR* (*missing not at random*) deritze.

Faltako balioak dauden ikerketetan datu-analisia egiterakoan, estatistikariak balioen profila antzeman eta horren arabera, egokia den metodoa erabili behar du.Datuen inputazioa da maiz erabiltzen diren metodo horietako bat. Hurrengo azpiataletan datuak inputatzeko metodo batzuen deskribapen laburra egingo dugu.

2.2. **Inputazio Metodoak**

Aurreko atalean esana dugu faltako balioak daudenean datu-analisia egiteko metodorik sinpleenak *CC* eta *AC* direla. *CC* eta *AC* metodoen ordezko aukera konplexuagoa da datuak inputatzea da, hau da, behatutako datuetatik abiatuz, faltako balioak egoztea. Inputazio metodoak bi talde nagusitan sailkatzen dira: bakunak eta anizkoitzak. Erabilitako datu kopuruek bereizten dituzte bi metodo horiek elkarrengandik hobekin: bakunek balio bakarra erabiltzen dute, eta anizkoitzek inputaziorako balio bat baino gehiago erabiltzen dituzte. Bakunen artean, *K-NNI*a da metodorik sendoena eta guk lan honetan aukeratu duguna. Metodo honek behaturiko datuen hainbat ezaugarri erabiltzen ditu modu bateratuan faltan dauden balioak ordezkatzeko [8].

Esan bezala, inputazio bakunek balio bakar bat erabiltzen dute eta honen ondorioz, lortutako emaitzak gainestimatuak gerta litezke. Horri aurre

egiteko inputazio anizkoitza delakoa garatu zen. Faltako balioak egotzirik, inputazio anizkoitzerako teknikak kontuan hartzen ditu lagin-aldakortasuna eta ez erantzunagatiko ziurgabetasuna. Teknika hauek ugariak dira, baina, esan bezala, lan honetan bi aztertuko ditugu, *MCMC* eta *PS* hain zuzen.

Ondoren, aipatutako hiru inputazio metodoak aurkeztuko ditugu era errazean, zehaztasun teknikoak alde batera utzita.

K-NNI

K-NNI inputazio teknika sailkapen metodoen familiakoa da. Kasuen arteko antzekotasunean oinarritzen da metodo hau: bi ale izanik, zenbat eta antz gehiago, hurbilago egongo dira.

$$\{(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{R}_i)\}\ i = 1, \ldots, n,$$

non $\boldsymbol{X}_i$ koaldagai-bektorea, $\boldsymbol{Y}_i$ inputatu beharreko menpeko aldagaia eta $\boldsymbol{R}_i$ informazio eza adierazlea diren. Behaturiko datu multzoa erabiliz eta distantzia metrikoan oinarrituz, behaturik gabe dagoen balio bakoitzerako, hurbilen dauden behaturiko K ale erabiltzen dira beronen balioa inputatzeko.

K-NNI algoritmoa burutzeko bi parametro kontuan hartzen dira: erabilitako distantzia metrikoa eta inputaziorako aukeratu beharreko kasu kopurua (K). Nahiz eta doikuntza honetan metrika desberdinak erabili ahal diren, distantzia Euklidearra da erabiliena eta horregatik guk aukeratu duguna [9].

Inputazio anizkoitza

Inputazio anizkoitza hiru pausutan urratsetan banatzen da [4], laburki honela bereiztuko genituzke. Lehendabiziko pausuan faltako balioak ordezkatzen dira, faltako baliorik gabeko M datu base berri eratuz. Ondoren, m datu-base bakoitzean ($m = 1, \ldots, M$), datuen analisirako proposatuak izan diren prozedura estatistikoak aplikatzen dira (erregresio lineala, logistikoa edo biziraupena, e.a.). Azkenik, eraikitako M datu-baseetatik eratorritako emaitzak estimazio bakar batean laburbiltzen dira.

Metodo honek inputaturiko datuen bariantza bi zatitan bereizten du: barne-inputazio eta inputazio arteko bariantza. Matematikako formalizazioetara jota, hauxe dugu:

$$W = \tfrac{1}{M} \sum_{m=1}^{M} V^m \text{ barne inputazio bariantza}$$

$$B = \tfrac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta}^m - \hat{\beta}^*\right)\left(\hat{\beta}^m - \hat{\beta}^*\right)' \text{ inputazio arteko bariantza,}$$

Bertan, β jatorrizko datuekin lortutako erregresio-koefizientea izanik, $\hat{\beta}^m$ m. inputazioan lorturiko β-ren estimazioa da, V^m kobariantza matrizea eta $\hat{\beta}^*$ M datu-baseetan lortutako β-ren estimazioen batezbestekoa.

Ondorioz, bariantza osoaren adierazpena ondorengoa izango da:

$$V = W + \left(\frac{M+1}{M}\right)\mathrm{B}$$

Inputazio anizkoitza egiteko teknika ugari daude. Metodoen arteko desberdintasun nagusia aipatutako lehen urratsean datza, faltako balioak ordezkatzeko erabilitako irizpidean hain zuzen. Egindako ikerketen arabera, hurbilketa metodo hauek, faltako balioek *MAR* profila dutela suposatuz aplika daitezke [10]. Ondoren, guk lan honetarako aukeratu ditugun inputazio anizkoitzeko bi teknikak azalduko ditugu laburki.

MCMC metodoa

MCMC metodoa da inputazio anizkoitza egiteko tekniken artean ezagunena. Metodo honek, Markoven kateen bitartez faltako balioak egozten ditu [10]. Banaketa normala duen zorizko aldagai batetik abiatuz, *MCMC* metodoak bi oinarrizko pausu ditu: *I*- eta *P*-pausuak. Pausu hauek errepikatuz, *MCMC*ak era honetan lan egiten du: *I*-pausuan (inputazioa), aldagaiaren batezbestekoa eta kobariantza matrizea kalkulatzen dira informazio osoa duten indibiduoak erabiliz. Ondoren, faltako balioak simulatu egiten dira, erabilgarriak diren datuetatik abiatuta.

$$\theta^{(t)},\, Y_{obs} \rightarrow Y_{mis}^{(t+1)} \sim f(Y_{mis}|Y_{obs}, \theta^{(t)}),$$

non $\theta^{(t)}$ batez bestekoak eta kobariantza matrizeak osaturiko parametro-bektore t. pasuan, Y_{mis}, $Y_{mis}^{(t+1)}$ faltako balioez osaturiko bektorea eta Y_{obs} behaturiko balioez osaturikoa osaturiko bektorea diren.

I-pausuaren hurrengoa, *P*-pausua da (aurresate ondoko banaketa). Atal honetan batezbestekoa eta kobariantza matrizeak osaturiko bektorea simulatu eta eguneratu egiten da, behaturiko balioak eta *I*-pausuan lorturiko balio simulatuak kontuan izanik:

$$\theta^{t+1} \sim f(\theta^{(t+1)}|Y_{obs}, Y_{mis}^{(t-1)})$$

$\{Y_{mis}^{(0)}, \theta^{(0)}\}$ hasierako balioak izanik eta bi pausu hauek, Markoven kate estokatisko bat sortzen da, $\mathrm{p}(\theta, Y_{mis}|Y_{obs})$ banaketara konbergentzia lortuz. Inputazio anizkoitzak lortzeko, *I*-eta *P*-pausuak behin eta berriz errepikatzen dira banaketa egonkor bat lortu arte.

PS metodoa

Rosenbaum eta Rubin [11] ikertzaileek garatu zuten metodo estatistiko hau.

Metodo honek tratamenduaren efektua estimatzea du helburu, zorizkoak ez diren ikerketetan .

PS metodoak, definizioz, pazientea bere oinarrizko ezaugarrien arabera tratatua izateko probabilitatea adierazten du. Baliabide matematikoetara jota, honelakoak ditugu:

$$e(\boldsymbol{X}) = \boldsymbol{P}(Z = 1|\boldsymbol{X}),$$

$\boldsymbol{X}$ koaldagaiez osaturiko bektorea izanik eta $Z = 1$ pazienteak tratamendua jasotzen duela adierazten duelarik. *PS* metodoa erregresio logistiko anizkoitzaren bitartez adierazita dator.

Oro har, ikerketetan koaldagai ugari aztertzen direnean, probabilitate handia dago banako batek koaldagairen batean faltako balioak izateko. Faltako balioak ageri direnean, datu-analisia egiteko gomendagarria da *PS* metodoaren erabilpena, batez ere, informazio eza koaldagaietan zein menpeko aldagaian ageri denean.

Egoera honetan, hurrengo pausuak eman behar dira datu basean dauden hutsuneak betetzeko: (1) menpeko aldagaian behaturik gabe dagoen balio bakoitzerako, *PS* metodoa aplikatzen da, probabilitatea kalkulatzeko:

$$logit(P(\boldsymbol{R}_i|\boldsymbol{X}_i,\beta) = (1, \boldsymbol{X}_i)'\beta,$$

non $\boldsymbol{R}_i = 1$, $\boldsymbol{Y}_i$ behaturik gabe baldin badago, $\boldsymbol{X}_i = (X_{i1},\ldots, X_{ip})$ *i.* pazientearentzat behaturiko koaldagaien balioak diren eta β, $\boldsymbol{X}_i$ matrizearen koefizienteak diren.

Ondoren, *PS* metodoaz baliatuz lorturiko balioetan oinarrituz, behaketa bakoitza tamaina bereko taldeetan banatzen da eta ondoren, hurbilpen metodo baten bitartez datuak inputatu egiten dira.

3. **SIMULAZIOA**

Aurreko ataletan aipaturiko hiru inputazio metodoen eraginkortasuna aztertzeko eta alderatzeko simulazio ikerketa bat egin dugu. Simulazio ikerketa honek, lagunduko digu metodo hauekin lorturiko emaitzen aldakortasuna balioztatzen.

Simulazio hau egiteko birikietako butxadura kronikoaren gaixotasuna (BBKG) pairatzen duten pazienteekin egindako ikerketan jasotako datuak

erabili ditugu [12]. Ikerketa honetan BBKG pairatzen duten pazienteen bizi-kalitatea neurtu zen, ikerketaren hasieran eta urtebete beranduago hain zuzen. Bizi-kalitatea neurtzeko *St Georges Respiratory Questionnaire* [13] galdetegia erabili zen, eta lan honetan galdetegi honetatik lortutako osoko puntuazioari erreparatuko diogu. Honek, adierazten du gaixotasunaren sintomak eta aktibitate fisikoak nolako eragina duen eguneroko bizitzan. Puntuazio honen arabera, bizi-kalitatea 0 (egoera ona) eta 100 (egoera txarra) bitartean dago. Horretaz gain, bizi-kalitatean eragina eduki dezaketen buruzko beste hainbat ezaugarri bildu ziren gaixotasunari buruz eta pazienteari buruz. Horien artean, badago BBKG egonkorra duten pazienteen egoeraren larritasuna adierazten duen HADO indizea. HADO indizeak 0tik 12ra arteko heina du, puntuazio altuagoak larritasun baxuagoa adierazten dutela. Lan honek helburu modura hartu du HADO larritasun indizeak bizi-kalitatearen bilakaeran duen eragina aztertzea. Datu-analisia egiteko erabili dugun eredua erregresio lineal anizkoitza izan da, bizi-kalitatearen bi neurketak datu errepikatuak direla kontutan hartuta. Aurretik ezarritako helburua betetzeko eredu linealeko denboraren eta larritasunaren arteko elkarrekintzaren (T*HADO) beta koefizientea eta bere adierazgarritasun estatistikoa erabili ditugu.

Egoera ideal batean bi uneetako neurketa guztiak behatuko balira paziente guztientzat, analisia egiteko ez genuke inolako zalantzarik izango, metodologia guztiz zehaztuta baitago. Gehienetan, errealitateak ez du egoera ideala islatzen eta hainbat pazienteren bizi-kalitatearen bigarren neurketa falta da. Guk lan honetan proposatu eta alderatzen ditugun hiru inputazio metodoen eraginkortasuna neurtzeko, ikerketa honetan azalduriko hipotesian eta lortutako datuetan oinarritu gara. Aplikazio honetan, menpeko aldagaia bizi-kalitatea (Y_i) izan da eta aldagai askea, larritasun indizea. Lortutako datuak luzeranzko datuak dira, bi unetan neurtu direlako. Abiapuntu gisa, aipatutako ikerketatik datu guztiak zituzten 400 pazientek osaturiko lagina hartu genuen. Indibiduo guztietan behaturiko aldagai menpekoaren balioetatik bere itxaropena eta desbideratze estandarra estimatuz, banaketa normalean oinarritutako aldagai berri bat simulatu genuen, era honetara:

$$Y_i^* \sim N(\mu, \sigma),\ i = 1, \ldots, n.$$

Jatorrizko Y aldagaia bornatua dago [0, 100] tartean, eta beraz simulaturikoak ere propietate bera izan behar du. Hori dela eta, simulaturiko Y_i^* aldagaia era honetara birdefinitu genuen:

$$Y_i^{*c} = \begin{cases} 0, & Y_i^* < 0 \\ Y_i^*, & 0 \le Y_i^* \le 100 \\ 100, & Y_i^* > 100 \end{cases}$$

Behin simulatutako menpeko aldagaia (Y_i^{*c}) eraikita genuenean, behaturiko aldagai askearen balioekin uztartu genuen, horretarako, jatorrizko Y eta X-ren arteko korrelazioaz baliatuz. Era honetan, egiaztatu genuen Y_i^*-ren balio minimoak zituzten pazienteak Y_i-ren balio minimoak zituztenekin elkartzen zirela.

Menpeko aldagai simulatua era horretara definitu ostean 400 indibiduoentzat, faltako balioak sortu genituen, aurretik definituriko hiru profilak erabiliz. Izan bedi p_i informazio eza definitzen duen probabilitatea. p_i zorizkoa, koaldagaien eta hasieran behaturiko menpeko aldagaiaren funtzio baten gisara adieraz daiteke edo bestela koaldagaien eta faltako balioak dituen aldagaiaren funtzio baten gisara.

Zehazki, $p_i = P(\eta_i)$, non η_i faltako balioen banaketaren araberakoa den:

$$\begin{aligned} &MCAR: && \eta_i = \alpha, \\ &MAR: && \eta_i = \boldsymbol{X}\alpha, \\ &MNAR: && \eta_i = \boldsymbol{X}\alpha + \gamma Y_i^{*c}, \end{aligned}$$

non α, γ $\boldsymbol{X}$ bektorearen koefizienteak diren eta Y_i^{*c} simulaturiko menpeko aldagaia.

Horrez gain, profil bakoitzerako % 10 eta % 30eko faltako balioen presentzia aplikatu genuen.

Azkenik, diagnosi grafikoak egin ziren inputazio metodoen doikuntza egokitasuna aztertzeko [14] Hain zuzen, inputatutako eta behatutako balio benetakoen arteko hodei-puntuak, eta korrelazio koefizienteak kalkulatu ziren. Halaber, inputazioa egin ondorengo maiztasun-banaketarako dentsitate grafikoak irudikatu ziren.

Simulaturiko egoera desberdinen arabera, sei testuinguru genituen, faltako balioen hiru profil eta bi portzentajeen arabera. Horietariko bakoitzerako proposaturiko hiru inputazio teknikak erabili genituen faltako balioak inputatzeko. Hiru inputazio teknika horien bidez lortutako datu base «osoak» eredu misto orokortuak erabiliz aztertu genituen [15].

Ikerketa honetako helburua inputazio metodoen ebaluazioa eta konparaketa zen. Hori dela eta, ereduko elkarrekintzaren beta koefizienteak kalkulatu eta alderatu genituen definituak zeuden faltako balioen profila eta portzentaje desberdinen arabera lortutako 6 testuinguruetarako eta inputazio metodo bakoitzerako. Beta koefizienteak alderatzeko hurrengo parametroak erabili genituen [16]:

— Alborapen erlatiboa honako eran definitu da: $\frac{\hat{\beta} - \beta}{\beta}$, non β jatorrizko datuen bidez lortutako estimazioa den eta $\hat{\beta}$ simulazioko egoera bakoitzean lortutako estimazioa.

— *Alborapen estandarizatua*: Benetako alborapenaren eta simulazioetan lorturiko beta koefizientearen errore estandarraren arteko zatiketa gisa definitzen da:
$\frac{\bar{\hat{\beta}} - \beta}{SE(\hat{\beta})}$, non β jatorrizko datuekin lortutako balioa den, $\hat{\beta}$ simulazioetan lortutako estimazioa, eta $SE(\hat{\beta})$ bere errore estandarra.
Estatistiko honen adierazgarritasun estatistikoa aztertzeko, bere banaketa $(n - 2)$ askatasun graduko t dela onartu zen.

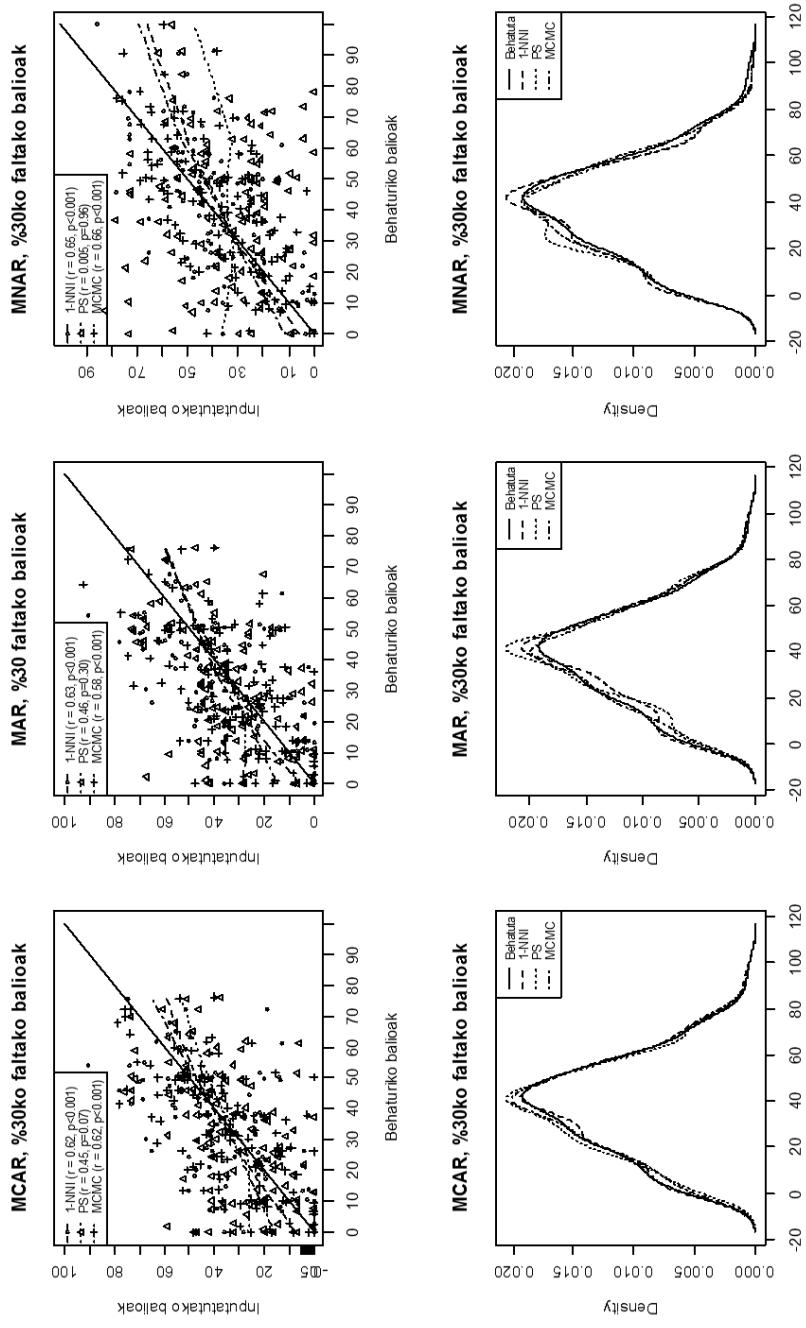
Datu-analisi guztiak SAS 9.2 eta grafikoak R 2.15 bertsioa erabiliz egin ziren.

4. **EMAITZAK**

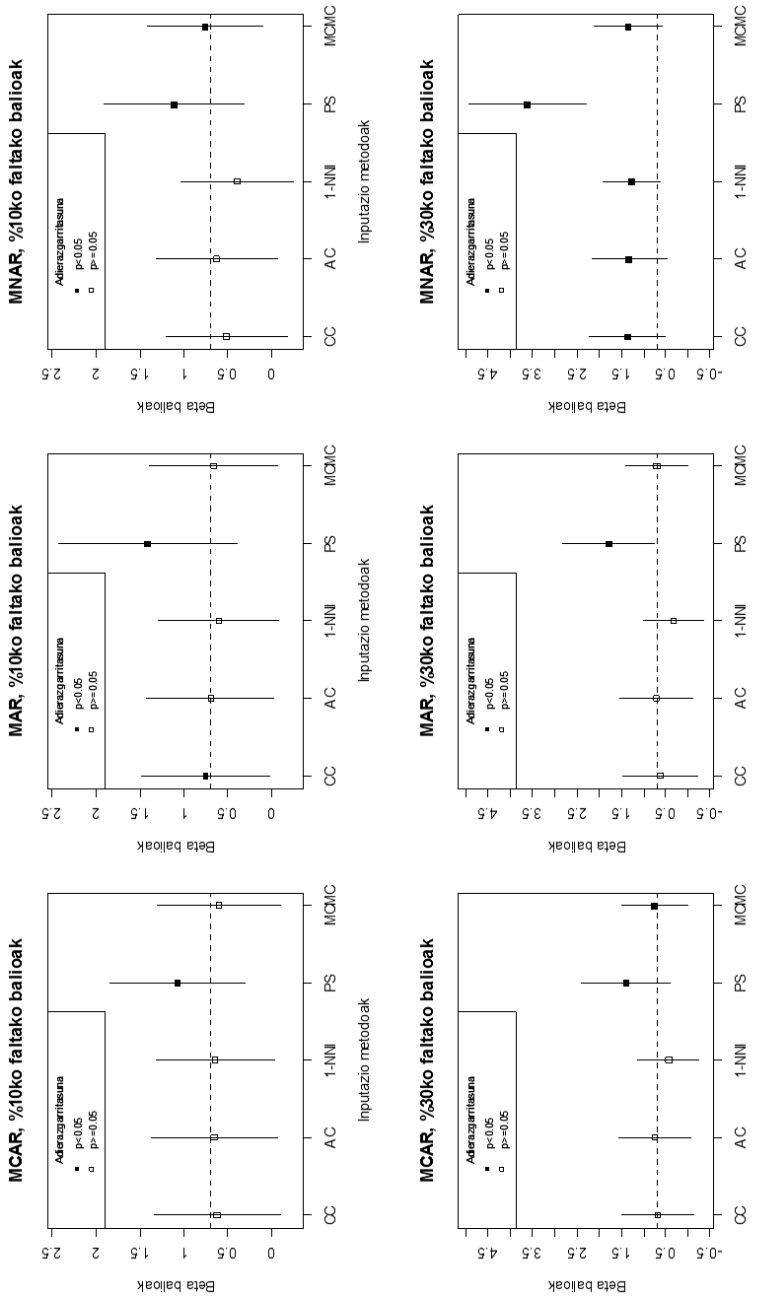
Lehendabizi, behaturiko datuen eta inputaturiko datuen arteko adostasuna erakusten dugu I. irudiko goiko zatian, bien arteko hodei-grafikoak azalduz. Irudi horren beheko zatian, inputatutako datuen maiztasun-banaketaren dentsitate-funtzioak ageri dira. Irudi honetan behatutako balioen % 30a faltan zegoenean lortutako emaitzak erakusten dira, baina %10eko portzentajerako antzeko emaitzak lortu genituen. *MCAR* banaketan oinarrituz, hiru metodoek antzeko portaera erakusten zuten, nahiz eta *PS* metodoak korrelazio txikiena adierazi (r = 0,45). Faltako balioen profila *MAR* edo *MNAR* zen kasuetan, datuen inputazioan egindako lana ez da ona metodo baterako ere, nahiz eta *MCMC* eta 1-*NNI* metodoen bidez *PS* metodoarekin baino emaitza hobeak lortu ziren.

II. irudian denboraren eta larritasunaren arteko elkarrekintzaren beta koefizienterako estimaturiko balioa aurkezten da ondoko kasuetarako: jatorrizko datuetan (inputazio gabe) eta proposaturiko hiru metodoekin inputaturiko datuetan. Irudian erakusten dira faltako balioen hiru profilen eta bi portzentajeen araberako emaitzak. Inputaturiko datuekin lortutako beta koefizienteez gain, beren % 95eko konfiantza tarteak ere agertzen dira irudian, jatorrizko datuetatik abiatuz lortutako koefizientearekin alderatuz, eta beren adierazgarritasun estatistikoa erakutsiz. Kasu guztietan *PS* metodoaren bitartez estimatutako beta koefizientea zen jatorrizko datuetatik lortutako baliotik urrunen zegoena. Faltako balioen profila *MCAR* edo *MAR* zenean, beste bi metodoen bidez kin lortutako diferentziak ez ziren estatistikoki adierazgarriak jatorrizko datuekin lortutakoekiko, informazio falta % 10 zein %30 ekoa izan. *MNAR* egoeran geundenean ordea, faltako balioen portzentajea % 10ekoa zenean emaitzak ez ziren txarrak, baina % 30ekoa zenean diferentzia guztiak ziren estatistikoki adierazgarriak.

I. taulan ageri dira simulazio ikerketan lortutako alborapen erlatiboa eta estandarizatuaren balioak, faltako balioen hiru profilen eta bi portzentajeen arabera. Halaber, taula horretan agertzen dira faltako balioen inputaziorik gabe egindako *CC* eta *AC* ohiko metodoetatik lortutako emaitzak ere.



I. irudia. *Goiko zatia:* Inputaturiko eta behaturiko balioen arteko komunztadura grafikoki (hodei-puntua) eta analitikoki (korrelazio koefizientea) adierazita. *Beheko zatia:* Behaturiko eta inputaturiko datuen banaketa enpirikoa (dentsitate grafikoa).



II. irudia. Beta koefizientearen estimazioak faltako balioen profila eta portzentajeen arabera. Lerro etenak jatorrizko datuekin lortutako beta koefizientearen balioa adierazten du (β: 0.686 eta bere errore estandarra: SE(β): 0.346; p-balioa = 0.048). Lerro bertikalak % 95ko konfiantza tarteak adierazten dituzte.

I. taula. Simulazioetan lorturiko beta koefizientea $\hat{\beta}$, alborapen erlatiboa eta estandarizatua, azterturiko inputazio teknika, faltako balioen profila eta % 10 eta % 30 informazio ezaren arabera. Jatorrizko datuetatik lortutako beta koefizientearen estimazioa: β: 0.686 eta bere errore estandarra: SE(β): 0.346 (p-balioa = 0.048).

Banaketa % behaturik gabe	Inputazio metodoa	$\hat{\beta}$	$SE(\hat{\beta})$	$\frac{\hat{\beta}-\beta}{\beta}$	$\frac{\bar{\hat{\beta}}-\beta}{SE(\hat{\beta})}$	p-balioa
MCAR						
%10						
	CC	0,620	0,369	–0,096	–0,179	0,858
	AC	0,650	0,367	–0,052	–0,097	0,923
	1-*NNI*	0,641	0,344	–0,066	–0,132	0,895
	PS	1,070	0,392	0,559	0,980	0,328
	MCMC	0,602	0,356	–0,122	–0,235	0,815
%30						
	CC	0,682	0,417	–0,006	–0,010	0,992
	AC	0,753	0,411	0,098	0,164	0,87
	1-*NNI*	0,432	0,353	–0,370	–0,719	0,473
	PS	1,386	0,491	1,021	1,426	0,155
	MCMC	0,749	0,377	0,092	0,168	0,867
MAR						
%10						
	CC	0,749	0,370	0,092	0,171	0,864
	AC	0,694	0,368	0,011	0,021	0,983
	1-*NNI*	0,603	0,349	–0,121	–0,237	0,813
	PS	1,409	0,497	1,054	1,454	0,147
	MCMC	0,659	0,369	–0,039	–0,072	0,943
%30						
	CC	0,627	0,422	–0,086	–0,140	0,889
	AC	0,726	0,414	0,058	0,096	0,924
	1-*NNI*	0,331	0,352	–0,518	–1,008	0,314
	PS	1,772	0,506	1,583	2,147	0,032
	MCMC	0,701	0,362	0,022	0,041	0,967
MNAR						
%10						
	CC	0,513	0,354	–0,252	–0,489	0,624
	AC	0,626	0,352	–0,087	–0,170	0,865
	1-*NNI*	0,392	0,327	–0,429	–0,901	0,368
	PS	1,109	0,400	0,617	1,057	0,291
	MCMC	0,759	0,338	0,106	0,215	0,830
%30						
	CC	1,365	0,433	0,989	1,568	0,117
	AC	1,324	0,425	0,930	1,501	0,134
	1-*NNI*	1,281	0,328	0,867	1,810	0,071
	PS	3,608	0,635	4,259	4,598	<0,001
	MCMC	1,350	0,394	0,968	1,684	0,093

MCAR: Missing Completely at Random; *MAR*: Missing at Random; *MNAR*: Missing Not at Random; *CC*: Complete Case; *AC*: Available Case; 1-*NNI*: 1 Nearest Neigbourg imputation; *PS*: propensity score; *MCMC*: Markov Chain Monte Carlo.

$\frac{\hat{\beta}-\beta}{\beta}$: alborapen erlatiboa; $\frac{\bar{\hat{\beta}}-\beta}{SE(\hat{\beta})}$: alborapen estandarizatua; p-balioa: alborapen estandarizatuaren adierazgarritasuna, (n–2) askatasun graduko t banaketatik abiatuz.

Kasu guztietan, *PS* metodoak erakutsi zuen alborapen estandarizaturik handiena. Hain zuzen, balioen % 30a falta zenean *MAR* egoeran, alborapen hori estatistikoki adierazgarria zen ($p = 0.032$). Datuen % 30a falta zenean, faltako balioen profila *MCAR* edo *MAR* izanik, 1-*NNI* teknikak erakutsitako alborapen estandarizatua ere handia zen.

Alborapen erlatiboari dagokionez ere, *PS* metodoarekin beta koefizientearen balio gainestimatuak lortu ziren kasu guztietan. Faltako balioen portzentajea % 30ekoa zenean, *MCAR* edo *MAR* profiletan, *CC* eta 1-*NNI* tekniken bitartez_jatorrizko datuekin lortutako balioa baino estimazio txikiagoak lortu ziren. Aldiz, *MNAR* egoeran, metodo guztiek emaitza gainestimatuak aurkeztu zituzten.

5. **EZTABAIDA**

Osasun arloko edozein ikerketatan garrantzi handikoa da menpeko aldagaiaren bilakaera denboran zehar aztertzea eta pazienteen ezaugarriekin duen lotura ikertzea.

Datuen analisirako metodologiaren konplexutasuna dela eta, maiz ebaluazio hau zaila izaten da, eta are zailago bilakatzen da faltako balioak ageri direnean.

Hainbat metodo estatistiko garatu dira luzeranzko edo errepikatuzko datuetan oinarrituriko ikerketetan faltako balioak daudenean datu-analisia egiteko burutzeko. Lan honetan kasu bakunenean oinarritu gara, hain zuzen menpeko aldagaiaren bi neurketa egiten diren kasuan. Horrek arlo askotako egoera arrunta islatzen du, esate baterako ikertzaileek interbentzio baten ondorioz izandako menpeko aldagairen aldaketan interesatuta daudenean. Arrunta da era berean honelakoetan faltako balioak izatea, gehien bat bigarren neurketan. Egoera honen aurrean, teknika ugari azaltzen dira bibliografian datu-analisia aurrera eramateko. Teknika gehienak faltan dauden datuen inputazioan oinarritzen dira, eta modu batera edo bestera, faltan dagoen informazio «asmatu» egitea proposatzen da. Teknika gehienak inplementatzeko nahiko konplexuak izaten dira. Baina, askotan teknikarik egokiena aukeratzea izaten da ikertzailearentzat erronkarik handiena, gehienetan inplementazioa bera baino handiagoa. Hutsune honetan ikertzaileei laguntza txiki bat eman nahian egin dugu lan hau. Egoera honetan, datuen analisirako teknika aukeratzeko funtsezkoak diren bi gai daude, lehenik faltako balioen zenbatekoa eta bigarrenik informazio-galeraren eredua edo profila, zorizkotasunari dagokionez. Bi elementu hauen araberako simulazio ikerketa aurkezten dugu lan honetan, datuen analisirako literaturan gehien erabiltzen diren teknikak alderatuz, hain zuzen *K-NNI*, *PS* eta *MCMC*. Simulazio ikerketaren emaitzak aurkezten ditugun arren, simulazioak datu errealetan oinarrituta egin dira. Beraz, esan genezake lortzen di-

ren emaitzek errealitatean gerta daitezkeen egoera ezberdinak nahiko ondo islatzen dituztela.

Desberdintasun ugari ikusten dira emaitzetan. *MCAR* egoeran, *AC* eta *CC* metodoek *K-NNI* eta *PS* teknikek baino alborapen txikiagoa erakutsi zuten. Datu-analisia behaturiko datuekin soilik egiteak, hain zuzen, CC, lagintamaina eta estimazioen efizientzia txikiagotzen ditu.

Efizientzia-gutxitze honek estimazioen adierazgarritasunari eragiten dio [4], eta *CC* eta *AC* metodoetan desbideratze estandar handiak lortzen dira. Hau dela eta, alborapenak txikiak izango dira. Horregatik, *CC* edo *AC* metodoen erabilera behaturik gabeko portzentajea txikia denean bakarrik gomendatzen da. *MAR* egoeran, aldiz, metodoen eraginkortasuna desberdina zen informazio-galera % 10ekoa edo % 30ekoa izanda. Galera % 10koa zenean *AC* metodoak agertzen zuen alborapenik txikiena. Informazio ezaren portzentajea txikia denean, datuak inputatu gabe eredu mistoak erabiltzea da gomendagarriena. Faltako balioen portzentajea % 30era aldatuz, *MCMC* metodoa zen jatorrizko datuen estimaziora gehien hurbiltzen zena. Berez, aurretik esan dugun bezalaxe, *MCMC* metodoa *MAR* egoerarako garatua dagoen metodoa da eta beraz, gure ikerketan lorturiko ondorioak beste ikertzaileen emaitzekin bat datoz [1]. Faltako balioen profila *MNAR* zenean, azterturiko metodoen bitartez kalkulatutako alborapenak handiak ziren. Aztertutako metodo hauek ez daude profil honetarako garatuak.

PS inputazio metodoak emaitza alboratuak agertu zituen kasu guztietan. Horrek metodo horren erabilpenaren hausnarketa sakonagoa eskatzen du. Teknika estatistiko hau *propensity* delakoan oinarrituta dago, eta horrek esan nahi du behaketak zenbaki finkoko taldeetan banaturik egon behar direla. Kasu honetan, sakonean aztertu beharko dugu baldintza hori ez betetzeari egotzi ote dakiokeen *PS* metodoaren errendimendu eskasaren arrazoia.

K-NNI metodoa ere aztertu dugu lan honetan. Oro har, lorturiko emaitzetan, 1-*NNI*-k , *PS* metodoak baino alborapen txikiagoak eta *MCMC*, *CC* edo *AC* metodoak baino alborapen handiagoak aurkeztu zituen. Lan honetan, erabilpenaren erraztasuna sinpletasuna bermatu nahian metodo honen eraginkortasuna $K = 1$ kasurako soilik aztertu zen. *K-NNI* metodoa erabiltzerakoan alborapena eta bariantzaren arteko oreka K parametroak zehazten du, $K = 1$ baliorako, alborapen txikia eta bariantza handia lortzen delarik. Etorkizuneko ikerketetan, gomendagarria izango litzateke K parametro honen balio desberdinetarako *K-NNI* metodoaren eraginkortasuna aztertzea.

Lan hau egiteko azaldu dugun datuen analisia menpeko aldagaiaren banaketaren normaltasunean oinarrituta dago. Erakusten dugun aplikazioan BBKG pairatzen duten gaixoetan osasunari lotutako bizi-kalitatea neurtzeko maiz erabiltzen den *St George* indizea izan da menpeko aldagaia.

Kasu honetan, aldagaiaren normaltasuna kontrastatu dugun arren, mota honetako aldagaietan banaketa normala ontzat hartzea ez da beti errealista, hain zuzen eskala bornatuak izatearen ondorioz. Horrez gain, berdin suposatu dugu koaldagaietan ez dagoela faltako baliorik. Halaber, emaitza hauek errepikapen bakarreko simulazio batetik eratorriak dira. Emaitza eta ondorio sendoagoak lortzeko simulazioetan hainbat errepikapen egitea gomendatzen da.

Gure emaitzek, erakutsi dute *MCAR* (%10 eta %30ko informazio-galera) edo *MAR* (%10 informazio-galera) profiletan, *AC* teknikarekin lorturiko beta estimazioak zirela alborapenik txikiena zeukatenak.

Aldiz, *MAR* egoeran eta behaturik gabeko informazioa % 30a zenean, *MCMC* metodoa zen eraginkorrena. Faltako balioen profila *MNAR* zenean, eta batez ere faltako balioen portzentajea altua denean, datuen analisi sakona egitea gomendatzen da inolako inputazio teknikarik proposatu baino lehen.

ESKER ONAK

Lan honetan erakusten den metodologiaren garapena bideragarria izan da MTM2010-14913, IT620-13 eta UFI11/52 diru-laguntzei esker. Gure eskerrik beroenak eman nahi dizkiogu Cristobal Esteban jaunari datu horiek erabiltzen uzteagatik.

6. BIBLIOGRAFIA

[1] ALTMAN D.G. eta BLAND J.M. 2007. «Missing data». *British Medical Journal*, **334,** (7590):424.

[2] BARNARD J. eta MENG X. 1999. «Applications of multiple imputation in medical studies: From AIDS to NHANES». *Statistical Methods in Medical Research*, **8**, 17-36.

[3] LAIRD N.M. 1988. «Missing data in longitudinal studies». *Statistics in Medicine*, **7**, 305-315.

[4] LITTLE R.J.A. eta RUBIN D.B. 2002. *Statistical analysis with missing data*. Wiley, New York.

[5] ROBINS J., ROTNITZKY A. eta ZHAO L. 1994. «Estimation of regression coefficients when some regressors are not always observed». *Journal of the American Statistical Association*, **89**, 846-866.

[6] JANSSEN K.J.M., DONDERS A., HARREL F., VERGOUWE Y., CHEN Q., GROBBEE D. eta MOONS K. 2010. «Missing covariate data in medical research: To impute is better than to ignore». *Journal of Clinical Epidemiology.* **63**, 721-727.

[7] SCHAFER J.L. eta GRAHAM J.W. 2002. «Missing data: Our View of State of Art». *Psychological Methods* **7**,147-177.
[8] GARCÍA-LAENCINA P.J., SANCHO-GÓMEZ J.L., FIGUEIRAS-VIDAS A. eta NERLYSEN M. 2009. «K-Nearest neighbours with mutual information for simultenous classification eta missing data imputation». *Neurocomputing, **72**, 1483-1493.*
[9] TROYANSKAYA O., CANTOR M., SHERLOCK G., BROWN P., HASTIE T., TIBISHIRANI R., BOTSTEIN D. eta ALTMAN DG. 2001. «Missing value estimation methods for DNA microarrays». *Bioinformatics*, **17**, 520-525.
[10] MOLENBERGHS G. eta KENWARD M.G. 2007. *Missing Data in Clinical studies*. John Wiley & Sons, West Sussex, Engleta.
[11] ROSENBAUM P. eta RUBIN D. 1983. «The central role of the Propensity Score in Observational Studies for Causal Effects». *Biometrika*, **7**, 41-55.
[12] ESTEBAN C., QUINTANA J.M., ABURTO M., MORAZA J., CAPELASTEGUI, A. 2006. «A simple score for assessing stable chronic obstructive pulmonary disease». *QJM: An International Journal of Medicine,* **99**, 751-759.
[13] JONES P.W., QUIRK P.H., BAYESTOCK C.M. 1992. «A self-complete measure of health status for chronic airflow limitation. The St George's Respiratory Questionnaire». *American Journal of Respiratory eta Critical Care Medicine*, **145**,1321-7.
[14] ABAYOMI K., GELMAN A. eta LEVY M. 2008. «Diagnostics for multivariate imputations». *Applied Statistics*, **57**, 3:273-291.
[15] MOLENBERGHS G. eta VERBEKE G. 2000. Linear *mixed models for Longitudinal data*. Springer, New York.
[16] BURTON A., ALTMAN D.G., ROYSTON P. eta HOLDER R. 2006. «The design of simulation studies in medical statistics». *Statistics in Medicine*, **25**, 4279-4292.