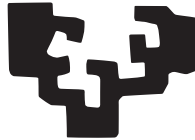


eman ta zabal zazu



Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

PhD Thesis:

**Proposal and validation of methodologies for
the categorisation of continuous variables in
the development of prediction models**

Irantzu Barrio Beraza

2015

Supervised by:

Inmaculada Arostegui

María Xosé Rodríguez-Álvarez

Nire familiari

Acknowledgements

Me gustaría empezar estas líneas dando las gracias a mis directoras Inmaculada Arostegui y María Xosé Rodríguez, sin vosotras esto nunca hubiese sido posible. Gracias por el apoyo, la paciencia, los consejos y todo lo que me habéis enseñado durante todos estos años. Inma, eskerrik asko nik ezer susmatu baino askoz lehenago lan honetan sinesteagatik eta ni gaur hemen egoteko egin duzun esfortzu eta ahalegin guztiagatik. Coté, grazas por todo o que me ensinaches e por toda a túa paciencia, as nosas maiores discusións convertéronse nos mellores resultados.

Gracias también al Doctor José María Quintana por darme la oportunidad de iniciarme en el mundo de la investigación, enseñarme y explicarme constantemente cómo es la investigación médica. Así mismo, quiero daros las gracias tanto a ti Txema como al Doctor Cristóbal Esteban, por dejarme vuestros datos, plantearme una problemática real y darme la oportunidad de intentar buscar una solución a dicho problema. I also would like to thank Luís Filipe Meira Machado for hosting me at your University and give me support whenever I needed it. Muito obrigada Luis. Gracias también a ti Vicente por confiar en mi trabajo y darme la oportunidad de mostrarlo a la comunidad internacional.

Así mismo, me gustaría agradecer todas las ayudas recibidas para poder llevar a cabo este trabajo. Empezando por los proyectos e Instituciones que han financiado mi trabajo directamente: Universidad del País Vasco UPV/EHU (GIU10/21), CIBER de Epidemiología y Salud Pública, Departamento de Educación Universidades e Investigación del Gobierno Vasco (UE09+/62), GOSIKER (BIOEF10/021) y el ISCIII Instituto de Salud Carlos III (PI06/1010); y siguiendo con los proyectos cuya financiación ha hecho posible realizar este trabajo: Departamento de Sanidad del Gobierno Vasco (200111002, 2005111005, 2005111008, 2012111008), Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT620-13),

Universidad del País Vasco UPV/EHU (UFI11/52), Ministerio de Economía y Competitividad (MTM2010-14913, MTM2013-40941-P), ISCIII Instituto de Salud Carlos III (PI061010, PI061017, PI06714, PI060326, PI060664, PI97/0326, PI020510), Red IRYSS (G03/220), REDISSEC (RD 12/0001/0001) y la Comisión de Investigación del Hospital Galdakao-Usansolo. Así mismo, quiero agradecer también a la Universidad del País Vasco UPV/EHU por las ayudas concedidas para la realización de las estancias de investigación en Guimarães, así como por la licencia concedida que ha permitido que hoy esta tesis esté terminada.

I would like to thank Frank Harrell and Camelia Sima for answering to my doubts and questions related to their research work. Additionally, I would like to thank the two expert doctors who provided a favourable report of this thesis.

Nola ez, eskerrak eman nahi dizkiet lankide guztiei. Hasteko, eskerrik beroenak zuri Arantza, edozein gauzarekin laguntzeko prest zaudelako beti eta tesiko momenturik gogorrenetan asko eskertzekoa da hori. Maider, Marijo eskerrikasko beti ni animatzen ibili zaretelako, atsedanak hartzera behartzen eta deskonektatzen laguntzen. Iraideri, egun hura ez dudalako ahaztuko eta pasilloan gurutzatu garen guztietan eman dizkidazun animo guztiengatik. Saileko lankide guztiei, azken txanpa honetan eskaini didazuen laguntasunagatik. Galdakaoko ospitaleko ikerkuntza unitateko lankide guztiei, hasierako urteetan erakutsi zenidaten guztiagatik.

Lan hau ez zen aurrera aterako etxekoan laguntzarik gabe. Eskerrik asko Hektor azken urteetan izan duzun pazientzia guztiagatik eta lan hau aurrera eramateko behar izan ditudan animo guztiak emateagatik. Eskerrik asko aita eta ama, sin vuestro apoyo incondicional hoy no estaría aquí. Eskerrik asko ere Urtzi eta San-drari askotan nire euskarria izateagatik. Pero sobre todo, gracias a todos vosotros por apoyarme en las decisiones más difíciles, creyendo mucho antes que yo en este proyecto. Azkenik, eskerrik beroenak lagunei beti hor egoteagatik, bereziki zuri Jaione, azken txanpa honetan emandako aholku zintzo guztiengatik.

Mila esker guztioi,

Muchas gracias,

Thank you very much.

Contents

Laburpena	ix
Summary	xv
1 Introduction	1
1.1 Prediction models	1
1.2 Categorisation of predictors	3
1.3 Organisation of subsequent chapters	6
2 Motivating data sets	9
2.1 The IRYSS-COPD study	9
2.2 The Stable-COPD study	12
3 Methodological background and preliminaries	17
3.1 Generalised linear model	18
3.2 Generalised additive model	19
3.3 Cox proportional hazards model	23
3.4 Discriminative ability measures	26
3.4.1 Definition	26
3.4.2 Comparison of the AUC	30
3.4.3 Optimism correction of the model's discriminative ability	31
4 Categorisation in logistic regression based on GAM	33
4.1 Proposed methodology	34
4.2 Validation and implementation	36
4.2.1 Application to the IRYSS-COPD study	36

4.2.2	Validation	38
4.2.3	Results	39
4.3	Conclusions and limitations	44
5	Categorisation methods in logistic regression based on the AUC	49
5.1	Proposed Methodology	50
5.1.1	Optimism correction	51
5.1.2	Selection of the optimal number of cut points	53
5.2	Empirical validation	55
5.2.1	Validation under known theoretical conditions	56
5.2.2	Comparison under nonlinear effects	67
5.2.3	Backward validation	72
5.3	Application to the IRYSS-COPD study	76
5.4	Conclusions	80
6	Categorisation methods in a survival model	83
6.1	Proposed Methodology	84
6.1.1	Optimism correction	85
6.1.2	Selection of the optimal number of cut points	86
6.2	Empirical validation	87
6.3	Application to the Stable-COPD study	112
6.4	Conclusions	116
7	Software development	119
7.1	An R function to calculate the average-risk category	120
7.2	The CatPredi package	123
7.2.1	catpredi.binary() function	125
7.2.2	controlcatpredi.binary() function	127
7.2.3	plot.catpredi.binary() function	128
7.2.4	catpredi.survival() function	128
7.2.5	controlcatpredi.survival() function	132
7.2.6	plot.catpredi.survival() function	133
7.2.7	comp.cutpoints.binary() function	134
7.2.8	comp.cutpoints.survival() function	135
8	Conclusions and Future Research	137

Contents	vii
References	155
A Appendix A	157

Laburpena

Gaur egun eredu auresaleek arlo askotan dute eragina, besteak beste: fisikan, meteorologian, finantzetan edo medikuntzan. Azken honetan, eredu auresaleen garrantzia gora doa erabaki prozesuen euskarri gisa eta balizko aldagai auresaleen inguruko jakintzak erabaki prozesu horietan lagungarri izaten ari da. Indibiduo baten informazio kliniko zein ez-klinikoaren arabera, gertaera kaltegarri bat izateko banakako arriskuaren estimazioaren bitartez, eredu auresale klinikoek partekatutako erabakiak hartzeko beharrezkoa den informazioa eskaini dezakete.

Eredu auresale bat garatzerakoan funtzezkoa da ereduan izango diren aldagai auresaleak ondo hautatzea, hala nola aldagai auresaleen eta erantzulearen arteko erlazioa ondo zehaztea. Aldagai jarraituen kategorizazioa ez da ontzat hartzen estatistika-ikuspuntu hutsaetik, informazioaren eta ahalmenaren galera ekar bait dezake. Gainera, badaude aldagai auresale eta erantzun aldagaiaren arteko linealtasunik inposatzen ez duten ereduak, hala nola, eredu gehigarri orokortuak (GAM izenekoak). Hala ere, ikerkuntza klinikoan eta batez ere, praktika klinikoan erabiliko diren ereduaren garapenean aldagaien kategorizazioa beharrezkoa suertatzen da. Bai medikuek zein osasun kudeaketaren arduradunek aldagai jarraituen kategorizazioaren beharra ikusten dute. Aldagaien kategorizazioa praktika klinikoan ohikoa izan arren, ez dago irizpide uniformerik mozketa puntuen kokapenari dagokionez.

Aldagaien kategorizazioaren gaia iadanik aztertu da literaturan. Baina gehienek mozketa puntu bakarra lortzea izan dute helburu. Lan honetan, eredu auresaleen garapenean erabiltzeko aldagai auresaleen kategorizazioan jartzen dugu arreta, baina betiere bi kategoria baino gehiago aintzakotzat hartuz. Horrela, informazioaren galera murriztu eta aldagai auresale eta erantzun aldagaiaren arteko erlazioa mantentzen da.

Gure helburua, erregresioan oinarritutako eredu auresaleetan aldagaiak katego-

rizatzeko metodologia baliagarria garatzea eta proposatzea da, batik bat erregresio logistikoa eta arrisku proportzionalak Cox ereduak erabiliz. Hauek baitira praktikan maiz erabiltzen diren ereduak erantzun aldagaia bitarra edo gertaera arteko denbora denean hurrenez hurren.

Birikietako butxadura kronikoaren gaixotasuna (BBKG) duten pazienteentzako eredu auresale bat garatzen geundela, zenbait aldagai jarraituen kategorizazioaren beharra sortu zen. Medikuek argi ikusten zuten zenbait aldagai, hala nola PCO_2 edo arnasketa-maiztasuna, eredu auresalean modu kategorikoan sartu behar zirela. Hala ere, ez zegoen akordiorik mozketa puntu kopuru eta beren kokapenaren inguruan. Mozketa puntuen hautua irizpide klinikoetan oinarritzen ez zirenean, maiz pertzentilen arabera aukeratzeko ziren. Hala eta guztiz ere irizpide klinikoetan oinarritzen zirenean ere, askotan ez zegoen akordiorik beraien artean. Ondorioz, arazo honi konponbide bat bilatzea erabaki genuen. Horretarako, kategorizaziorik hobereena eskaintzen zuen metodologia garatzea pentsatu genuen. Hura izan zen tesi honetan lan egiteko lehen motibazioa.

Hasierako urrats batean, X aldagai jarraitua eta Y aldagai erantzule bitarraren arteko erlazio grafikoaren arabera X kategorizatzea proposatu genuen. Bi aldagaien arteko erlazioa erakusteko erregresio logistikoa gehigarria eta P -*spline* leuntzaileak erabili ziren. X aldagai jarraitua gutxienez 3 kategorietan sailkatzea proposatu genuen, bi mozketa puntu horien kokapena batez besteko arriskuko kategoriaren limiteak izanik. Hirugarren mozketa puntu bat beharrezkoa izanez gero, irizpide klinikoan edo adierazpide grafikoan behatutako malda aldaketan oinarrituta aukeratu litzateke. Izan ere, metodologia honek muga batzuk ditu: hirugarren mozketa puntuaren hautua subjektiboa da, ez du X aldagaiaren kategorizazioa eredu anizkoitz batean bermatzen eta erantzun bitarrera mugatua dago.

Ondorioz, bigarren urrats batean metodologia orokor baten garapenean murgildu gara. Metodologia orokor honek mozketa puntu optimoak eskaintzen ditu eredu sinple zein anizkoitzean eta erantzun aldagaiaren banaketa ezberdinetarako. Hasteko erregresio logistikoa X aldagaiarentzako k mozketa puntu optimo aukeratzeko metodologia garatu dugu. Mozketa puntuen hautua eredu sinplean edo anizkoitzean egin daiteke, azken honetan $\mathbf{Z} = (Z_1, \dots, Z_p)$ beste aldagai auresaleen eragina kontuan hartuz. Y erantzun aldagaiarentzako erregresio logistikorik hobereena eskaintzen duen $\mathbf{v}_k = (x_1, \dots, x_k)$ k mozketa puntuen bektorea aurkitzean oinarritzen da gure proposamena. Izan bedi X_{cat_k} , $k + 1$ kategoria dituen eta 0tik k rako balioak hartzen dituen aldagai kategorikoa. Orduan (1) ereduaren *receiver operating characteristic* (ROC) kurbaren azpiko azalera (AUC) maximoa egiten duen

$\mathbf{v}_k = (x_1, \dots, x_k)$, k mozqueta puntu optimoen bektorea izango da.

$$\text{logit}(\pi(\mathbf{Z}, X_{cat_k})) = \beta_0 + \sum_{r=1}^p \beta_r Z_r + \sum_{q=p+1}^{p+k} \beta_q 1_{\{X_{cat_k}=q-p\}}. \quad (1)$$

AUC hori maximoa egiten duten mozqueta puntuak aurkitzeko, bi algoritmo proposatzen ditugu: *AddFor* eta *Genetic*. Batetik, *AddFor* algoritmoak mozqueta puntu bat bilatzen du aldiro. Hau da, lehendabizi $k = 1$ -entzako (1) ereduaren AUC maximoa egiten duen x_1 bilatzen du (X ren eremuan berdin tartekatutako balioen M tamainako sare batean). Behin x_1 finkatu duela, $k = 2$ -rako (1) ereduaren AUC maximoa egiten duen x_2 bilatzen du (M tamainako sarean) ($x_2 \neq x_1$). Prozesua errepikatzen da $\mathbf{v}_k = (x[1], \dots, x[k])$ bektorea lortu arte, non $x[o]$ -k ordenatutako o -garren mozqueta puntua adierazten duen. Bestalde, *Genetic* metodoak aldi berean bilatzen du (1) ereduaren AUCa maximoa egiten duen k mozqueta puntuen bektorea. Horretarako ezagunenak diren eboluzio-algoritmoak erabiltzen ditu, algoritmo genetikoak hain zuzen.

Mozqueta-puntuen hautaketaz gain, AUCaren gainestimazioaren arazoari heldu diogu. Datu berdinak erabiltzen direnez erregresio logistikoaren eredua doitzeko (mozqueta puntuen hautuan nahasia) eta AUCa estimatzeko, azken honen estimazioa alboratua egon daiteke eta zuzendu beharra dago. Testuinguru honetan bootstrap-ean oinarritutako hurbilketa bat proposatu dugu kategorizatutako aldagaiaren AUCaren gainestimazioa zuzentzeko asmoz. Bestalde, mozqueta puntu kopururik egokiena aukeratzeko, X aldagai jarraituaren bi kategorizazio konparatzeko metodoa proposatu dugu. Jakitun gara, teorikoki mozqueta puntu kopuru optimorik ez dela existitzen, ezen guztien gaintik aldagai jarraitua baitago. Hala ere, praktika klinikoan, aldagai jarraituen bertsio kategorikoak dira erabilienak, baina jakin gabe gehienetan zein kategoria kopuru hobesten den. Ondorioz, $k = l$ eta $k = l + 1$ mozqueta puntu kopuruak konparatzeko, beraien zuzendutako AUCen diferentzietan oinarritutako metodoa proposatu dugu.

Hainbat simulazio garatu ditugu proposatutako metodoak balioztatzeko helburuarekin. Lehenengo simulazio azterketa, baldintza teoriko ezagunetan oinarritu da. Honen bitartez AUCaren zuzenketaren beharra aztertu da eta estimatutako mozqueta puntuak hala nola *AddFor* eta *Genetic* algoritmoen eraginkortasuna balioztatu da. Bestalde *backward validation* deituriko simulazio azterketa egin da non mozqueta puntuak irizpide klinikoetan oinarrituta aurretiaz finkatuak daude eta hauen estimazioa balioztatu da.

Erantzun bitarraz gain, praktika klinikoan maiz erabiltzen den erantzun aldagaia da gertaeren arteko denbora. Mota honetako erantzun aldagaiak aztertzeke biziraupeneko eredurik ezagunena da arrisku proportzionaletako Cox eredua. Ondorioz, X aldagai auresale jarraitua arrisku proportzionaletako Cox ereduan kategorizatzea kontsideratu dugu. Aurretiaz aztertu egin da mozketa puntuoen estimazioa zentsuratutako datuen presentzian, baina guztietan mozketa puntu bakarra bilatzea izan dute helburu. Hala ere, aldagai auresalea bi baino kategoria gehiagotan kategorizatzea zen gure helburua, eredu simple zein anizkoitza aintzakotzat hartuz. Beraz, erregresio logistikorako garatutako metodologia arrisku proportzionaletako Cox eredura luzatzea proposatu dugu. Eredu honen diskriminazio-ahalmena neurtzeko konkordantzia-probabilitate indizea kontsideratu dugu. Are gehiago, bi estimatzaile ezberdin aztertu dira *c-index* eta *concordance probability estimator* (CPE) hain zuzen. Simulatutako datuetan oinarritutako balioztatze ikerketa garatu dugu, aldagai auresale jarraitu baten kategorizazioan CPE eta *c-index* estimatzaileen errendimendua aztertzeke. Simulazio ikerketa honetan hainbat egoera ezberdin aztertu dira. Alde batetik, mozketa puntu kopuruari dagokionez, $k = 1, 2$ eta 3 aztertu ditugu. $k = 1$ erako, X aldagai auresale jarraituaren eta T bizirik irauteko denboraren arteko arrisku-harreman hazkorra eta beherakorra aztertu dira. Horretaz gain, mozketa puntu teorikoentzako kokapen ezberdinak aztertu ditugu: a) aldagai auresalearen banaketaren erdialdean; b) arrisku handiko eremura mugituta eta c) arrisku baxuko eremura mugituta. $k = 2$ eta $k = 3$ -rentzako, X aldagai auresale jarraituaren eta T bizirik irauteko denboraren arteko erlazio lineala eta ez-lineala aztertu dira. $N = 500$ eta $N = 1000$ lagin tamainako datu baseen 500 erreplika simulatu dira.

Proposatutako metodoak BBKG zuten gaixoen bi ikerketetara aplikatu dira. BBKGaren gaizkiagotze bat duten gaixoen ikerketa den IRYSS-COPD Study, PCO_2 aldagai auresalea erregresio logistiko simple eta anizkoitzean kategorizatzeke erabili da erantzun aldagaia epe laburreko eboluzio oso txarra izanik. Eredu anizkoitzean, PCO_2 aldagaia *Glasgow coma scale* eta bihotz-maiztasuna aldagai auresaleen eragina kontuan hartuz kategorizatu da. Antzeko emaitzak lortu ditugu *AddFor* edo *Genetic* algoritmoak erabiliz, baita eredu sinplea edo anizkoitza erabili ditugunean ere. Lortutako emaitzen arabera, mozketa puntu kopururik egokiena bi izan da, beraz PCO_2 aldagaia hiru kategorietan kategorizatu da era optimo batean. Bestalde, BBKG egonkorra zuten gaixoen ikerketa, Stable-COPD Study, $\text{FEV}_{1\%}$ aldagaia arrisku proportzionaletako Cox ereduan kategorizatzeke erabili da. Kasu honetan bost urteko bizi-iraupena aztertu da. Ikerketa honetan lanean ari ziren medikuek

bi helburu mahai-gaineratu zizkiguten. Lehenengo eta behin, $FEV_{1\%}$ aldagaia lau kategorietan kategorizatzeko eskatu ziguten, hau da $k = 3$ mozketa puntu optimoak bilatzea eredu sinplea aintzat hartuz. Helburua, gure emaitzak eta literaturan aurretiaz proposatutako beste mozketa puntuak konparatzea zen. Bigarren helburua ordea, eredu anizkoitzean mozketa puntuen kokapen eta kopuru hoberena lortzea izan da, disnea eta adinaren eragina kontuan hartuz. Azken bi hauen hautua BBKG egonkorra duten gaixoen eboluzioaren aurreale onak izatean oinarritu da. Eredu sinplean lortutako emaitzak, aurreko proposamenetan lortutakoak baino hobekak izan dira. Gainera, $FEV_{1\%}$ aldagaia eredu anizkoitzean kategorizatu dugunean, mozketa puntu kopuru optimoa bi dela lortu dugu, *AddFor* eta *Genetic* algoritmoekin lortutako emaitzak berdinak izanik. Ikerketa bietan, proposatutako metodologia aplikatzerakoan lortutako emaitzak medikuengandik balioztatuak izan dira.

Azkenik, erabiltzeko erreza den pakete bat garatu dugu R softwarean, **CatPredi** deiturikoa. R-ko funtzioz osaturiko paketea da hau, aldagai jarraitu baten kategorizazio optimoa lortzea bermatzen duena. Kategorizazio hau bai ereduaren garapenaren aurretik (eredu sinplea) zein garapenean zehar (eredu anizkoitza) lor daiteke. Aukera ezberdinak garatu egin dira pakete honetan, aukeratutako eredu aurrealearen arabera. Hau da, erregresio logistikoa erantzuna bitarra denerako eta arrisku proportzionala Cox ereduaren erantzuna gertaera arteko denbora denerako. **CatPredi** paketeak hainbat emaitza eskaintzen dizkio erabiltzaileari. Alde batetik, erabiltzaileak aukeratutako kopururako, mozketa puntuen kokapena eskaintzen du. Gainera, sortutako aldagai kategorikoarekin doitutako erregresio ereduaren emaitza bueltatzen du. Halaber aldagai kategoriko horrentzako diskriminazio-ahalmeneko indizearen estimazioa hala nola zuzendutako estimazioa ematen ditu. Azkenik, mozketa puntu kopuru ezberdinentzako lortutako kategorizazio proposamenak konparatzea posiblea da **CatPredi** paketearen bidez, mozketa puntu kopuru optimoa lortuz intereseko aldagai aurrealearentzako.

Doktorego-tesi honetan, erantzun aldagaiaren banaketaren arabera, hurbilketa ezberdinak proposatu ditugu aldagai jarraituen kategorizaziorik hoberena lortzeko. Gure proposamenekin, nahi beste mozketa puntu optimo lor daitezke, bai eredu sinplean zein eredu anizkoitzean. Aurretiaz, aldagai jarraituen kategorizazioan lan egin da, baina gehienetan, mozketa puntu bakarra bilatzeko helburuarekin. Guk dakigunaren arabera, aurretiaz egindako proposamenek ez zuten aldagaiaren kategorizazioa eredu anizkoitzean kontsideratzen, ezta mozketa puntu optimoaren aukeraketa ere. Garrantzitsua da hemen aipatzea guk ez dugula kategorizazioa bera modelizaziorako irtenbide bezala proposatzen. Tesi honen helburua izan da medikuek

aldagai jarraituen kategorizatzea beharrezkoa ikusten dutenean, kategorizazio hori era egokienean egitea baimenduko duen metodologia eskaintzea.

Ondorioz, tesi honetan aldagai aurreale jarraituak kategorizatzeke baliozko metodologia garatu eta proposatu dugu, medikuek beharrezkoa ikusten dutenean erabili ahal izateko.

Summary

Prediction models are currently relevant in a number of fields such as physics, meteorology, finance or medicine, among others. In the medical field, prediction models are gaining importance as a support for decision-making whereby increased knowledge of potential predictors helps the decision-making process. Clinical prediction models may provide the necessary input for shared decision-making by estimating an individual's risk of an unfavourable event or developing a certain disease over a specific time period on the basis of his or her clinical and non-clinical profile. A vital factor in the development of prediction models is the selection of the predictors or covariates (clinical variables) to be used in the model. From a statistical perspective, categorising continuous variables is not advisable, since it may entail a loss of information and power. In addition, there are statistical modelling techniques such as the generalised additive models (GAM) which do not require any assumption of linearity between predictors and response variables, and so allow for the relationship between the predictor and the outcome to be modelled more appropriately. Yet in clinical research and, more specifically, in the development of prediction models for use in clinical practice, both clinicians and health managers call for the categorisation of continuous parameters. Despite the fact that categorisation is a common practice in clinical research, there are no unified criteria for the selection of the cut points. Previous work has been done in the categorisation of continuous variables but with the aim in almost all cases of dichotomising the predictor variable. In this dissertation, we focus on the categorisation of continuous variables to be used in the development of prediction models, considering that the use of more than two categories may be preferable. This serves to reduce the loss of information and enables the relationship between the covariate and the response variable to be retained. Our goal is to propose a methodology to categorise continuous predictor variables

in regression-based prediction models, mainly focussing on the logistic and Cox regression models which are those most widely used in the medical field for modelling dichotomous and time-to-event outcomes respectively.

The work presented in this dissertation was initially motivated by the development of a prediction model in the context of patients with chronic obstructive pulmonary disease (COPD). Clinicians agreed on the use of a categorised version of some clinical parameters such as the blood gas PCO_2 or the respiratory rate in the prediction model. However, they did not agree on the location and number of cut points. We noticed that these were usually based on quartiles and when they were based on clinical criteria there was no agreement between them. Several proposals are available in the literature, but most aimed at the selection of a single cut point. Thus, we considered developing a methodology to categorise continuous predictor variables in prediction models.

In a first stage we considered categorising a continuous predictor variable X by considering its graphical relationship with a binary response variable Y based on a logistic GAM with P-spline smoothers. We proposed to categorise X in a minimum of three categories, considering the limits of the average-risk category as the location of the cut points. The location of the third cut point, if needed, was to be based on clinical criteria or a change in the slope of the graphical display. Nevertheless, this methodology had some restrictions: the location of this third cut point was subject to subjectivity, it did not allow us to categorise X in a multivariate setting and it was limited to a binary outcome.

Thus in a second stage, we claimed for a proposal that provided with an optimal categorisation of a continuous predictor in a multivariate setting for different distributions of the response variable. We started by developing a methodology in which the location for any given k number of cut points for X could be optimally selected in a logistic regression, in addition or not to a set of other predictor variables, $\mathbf{Z} = (Z_1, \dots, Z_p)$. The proposal consisted of the selection of a vector $\mathbf{v}_k = (x_1, \dots, x_k)$ of k cut points in such a way that the best logistic predictive model was obtained for the response variable Y . Specifically, given k the number of cut points set for categorising X in $k + 1$ intervals, let us denote X_{cat_k} the corresponding categorised variable taking values from 0 to k . Then, what we propose is that the vector of k cut points $\mathbf{v}_k = (x_1, \dots, x_k)$, which maximises the area under the receiver operative characteristic (ROC) curve (AUC) of the logistic regression

model shown in equation (2) is thus the vector of the optimal k cut points.

$$\text{logit}(\pi(\mathbf{Z}, X_{\text{cat}_k})) = \beta_0 + \sum_{r=1}^p \beta_r Z_r + \sum_{q=p+1}^{p+k} \beta_q \mathbf{1}_{\{X_{\text{cat}_k}=q-p\}}. \quad (2)$$

To search for those cut points which maximise the AUC, we propose two alternative algorithms, namely *AddFor* and *Genetic*. Using the *AddFor* algorithm, one cut point is searched for at a time. In other words, this algorithm first seeks x_1 (in a grid of size M of equally spaced values in the range of X), such that the AUC of the logistic regression model shown in equation (2) for $k = 1$ will be maximised. Once x_1 has been selected, it is fixed and the algorithm proceeds to seek x_2 (in the grid of size M) ($x_2 \neq x_1$), so as to ensure that the AUC of the model in equation (2) for $k = 2$ will be maximised. The process is then repeated until the vector of k cut points, $\mathbf{v}_k = (x[1], \dots, x[k])$, has been obtained, with $x[o]$ denoting the o -th ordered cut point. On the other hand, the *Genetic* method simultaneously finds the vector of k cut points, $\mathbf{v}_k = (x_1, \dots, x_k)$, which maximises the AUC of the logistic regression model in equation (2) by using genetic algorithms.

Furthermore, we addressed the problem of overestimation of the AUC when the same data is used to fit the logistic regression model (involved in the cut point selection process) and estimate the AUC. In this context, we propose a bootstrap based approach to correct the optimism of the AUC for the categorised variable. In addition we propose a naive approach to compare two given categorisations of the predictor variable X with the aim of selecting the best number of cut points. We are aware that in theory the optimal number of cut points for the categorisation of a continuous variable does not exist, since above all the possible number of cut points, the best option would be the continuous variable. However, in clinical practice categorical versions of the continuous variables are usually preferred without it always being clear which is the best number of categories to be used. Hence, we propose an approach for selecting the best number of cut points based on the difference between the bias-corrected AUCs obtained for $k = l$ and $k = l + 1$ cut points.

Several simulation studies were conducted to empirically validate the proposed methods. The first simulation study was performed under known theoretical conditions. With this setting we studied the need of the bias correction of the AUC and validated the estimated cut points and the performance of the algorithms *AddFor* and *Genetic*. In addition, we conducted a backward validation simulation study in which we validated the estimated cut points when the cut points were scientifically

pre-established based on clinical knowledge.

In addition to the binary response variable, a common outcome in clinical practice is the time until the event occurs. The Cox proportional hazards model is the most common survival prediction model for the analysis of time-to-event data. Hence, we considered categorising a continuous predictor variable X in a Cox proportional hazard model. Previous work on the estimation of optimal cut points with censored data has been done but sought a unique cut point. However, our aim was to categorise a continuous predictor variable considering more than a unique cut point in either a univariate or a multivariate setting. Consequently, we propose to extend the methodology developed for the logistic regression to the Cox proportional hazards model. To measure the discriminative ability of the model, we considered the concordance probability index, and two different estimators were studied: the c-index and the concordance probability estimator (CPE). The algorithms used to select the optimal cut points were the ones mentioned above, *Addfor* and *Genetic* respectively. An empirical validation based on simulated data was performed to evaluate the performance of both the c-index and CPE estimators when it came to selecting the optimal cut points for the categorisation of continuous variables. Various different settings were considered for this simulation study. First of all, as far as the number of cut points is concerned, $k = 1, 2$ and 3 were considered. For $k = 1$ we considered increasing and decreasing risk relationship between the continuous predictor X and survival time T . Additionally, we considered different positions for the theoretical cut points: a) centred in the predictor's distribution; b) shifted to high risk area and c) shifted to low risk area. For $k = 2$ and $k = 3$ we considered a linear and a nonlinear relationship between the continuous predictor X and survival time T . $R = 500$ replicates of simulated data were performed for total sample sizes of $N = 500$ and $N = 1000$.

The proposed methods were applied to real data from two different Studies of patients with COPD. The IRYSS-COPD study with patients with exacerbated COPD is used to categorise the predictor variable PCO_2 in a univariate and multivariate logistic regression model where the response variable is short-term very severe evolution. In the multivariate setting, the PCO_2 was categorised adjusted by the effect the predictor variables Glasgow coma scale and heart rate. Similar cut points were obtained when the *AddFor* and *Genetic* algorithms were used, also when the univariate or multivariate setting were used. The results obtained suggested that the best number of cut points was two and hence the PCO_2 was optimally categorised into 3 categories. The Stable-COPD study was used to categorise the predictor variable

$FEV_{1\%}$ in a Cox proportional hazards model considering 5-year survival. Clinical researchers involved in the Stable-COPD study presented us with two goals. First, the aim was to categorise the predictor variable $FEV_{1\%}$ into four categories (mild, moderate, severe and very severe), i.e., $k = 3$, in a univariate setting in order to compare the results obtained with previous categorisation proposals. The second goal was to look for the best categorisation (location and number of cut points) in a multivariate setting, taking into account the effect of age and dyspnoea, which are seen as important predictors for the severity of stable COPD patients. The results obtained in a univariate setting improved those obtained from previous categorisation proposals. In addition, when we categorised the predictor variable in a multivariate setting, we obtained that two was the optimal number of cut points, hence we categorised the predictor variable $FEV_{1\%}$ in three categories being these the same when the *AddFor* or *Genetic* algorithms were used. The cut points obtained by the proposed methodology were face-validated by clinicians in both studies.

Finally, we have developed an easy-to-use package in software R, called **CatPredi**. This is a package of R functions that allows the user to categorise a continuous predictor variable either before (univariate setting) or during the development of a prediction model (multivariate setting). Different approaches have been implemented depending on the prediction model chosen, i.e., logistic regression (for binary response variables) or Cox proportional hazards model (for time to event outcomes). The **CatPredi** package provides the optimal location of cut points for a chosen number of cut points, fits the prediction model with the categorised predictor variable and returns the estimated and bias-corrected discriminative ability index for this model. Additionally, it allows to compare two categorisation proposals for different number of cut points and select the optimal number of cut points.

In this dissertation we propose different approaches for categorising continuous variables depending on the distribution of the response variable for any given number of cut points. Additionally, this categorisation approach can be applied in either a univariate or multivariate setting. Previous work on categorisation has been done but with the aim in almost all cases to dichotomise the continuous predictor variable. To the best of our knowledge, none of the previous proposals allowed the categorisation during the development of the model neither considered selecting the best number of cut points. We must note that in this dissertation we do not recommend the categorisation as a modelling solution, but our goal is to propose a valid way to do so whenever it is needed.

In conclusion, in this dissertation we propose a valid methodology for categorising

a continuous predictor variable whenever it is considered necessary by a clinical researcher.

Chapter 1

Introduction

1.1 Prediction models

Prediction models are currently relevant in a number of fields such as physics, meteorology, finance and medicine, among others. In the medical field, prediction models are gaining importance as a support for decision-making, whereby the increased knowledge of potential predictors helps the decision-making process. Decisions such as the most appropriate treatment for a disease, or whether or not a given patient should be discharged; or the development of effective, acceptable, and cost-efficient prevention strategies are based on the individual patient's risk of suffering some unfavourable event. Additionally, "shared decision-making" is now the norm whereby clinicians and patients are both actively involved in deciding therapeutic interventions or choosing medical treatments. This shared decision-making process requires us to be aware of the potential risks and advantages of each of the decisions to be taken (Steyerberg 2009). Clinical prediction models provide estimates for an individual's risk of an unfavourable event or development of a certain disease over a specific time period on the basis of a combination of a number of patient characteristics which we call variables. These variables, whose information is known, can be related to the patient, the disease or the treatment, for example. Estimation of the individual's risk of an unfavourable event by the prediction model may provide the necessary input for shared decision-making. Often, clinical prediction models are extended to include clinical prediction rules, risk scores or prognostic models.

The literature includes well-known prediction models which have been developed to predict the development of a disease, death or poor evolution caused by a current disease. The Framingham risk score, for example, was developed in 1998 to predict

coronary heart disease (Wilson et al. 1998). Today it is still a widely-used risk score with more than 7000 citations and appears in clinical guidelines such as those of the National Heart, Lung and Blood Institute (National Cholesterol Education Program 2002). Likewise, a risk score was developed to predict type II diabetes with the aim of identifying those individuals who would benefit from intensive lifestyle advice (Lindström and Tuomilehto 2003). In the context of patients with chronic obstructive pulmonary disease (COPD), the development of risk scores or prediction models is an active research area (Celli et al. 2004, Haroon et al. 2015, Make et al. 2015, Quintana et al. 2014a;b). In general, publications about clinical prediction models have increased considerably in the last few years (see Figure 1.1).

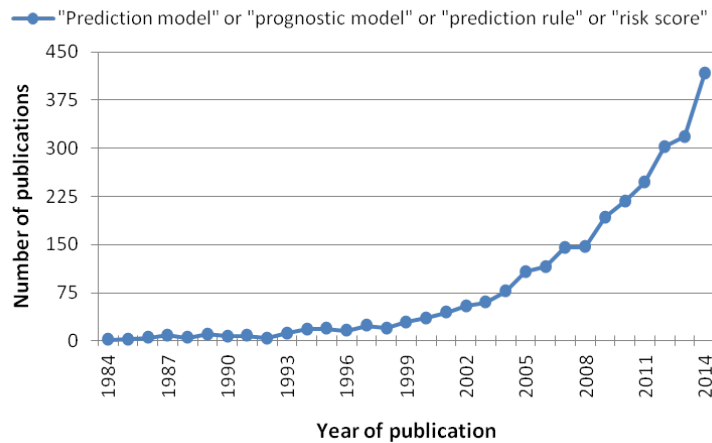


Figure 1.1: Number of articles available in Pubmed published between 1984 and 2014 with the terms “prediction model”, “prediction rule”, “prognostic model” or “risk score” in the title. The searched was performed in April 2015.

Prediction is an estimation problem together with hypothesis tests approaches. Prediction models may serve to answer estimation and hypothesis test questions while summarising the data structure. When prediction models are developed it may be necessary to make several assumptions regarding the structure of the data or the relation between covariates. For example, whether predictors effects work in an additive way or whether continuous predictors have linear effects should be tested. In this dissertation we will focus on regression-based prediction models, particularly the logistic (McCullagh and Nelder 1989) and Cox models (Cox and Oakes 1984) which are those most widely used in the medical field for modelling dichotomous

and time-to-event outcomes respectively (Steyerberg et al. 2013).

Finally, if the aim is to apply the prediction model in practice, it is important to show that it is valuable when applied to new data, which is called validation. Different validation strategies may be used in practice, i.e. internal or external validation (Steyerberg 2009). Internal validation evaluates the validity of the model when it is applied to data derived from the same sample in which they have been developed. Conversely, external validation examines the generalisability of the model to other samples. For example, we could assess the validity of the model when applied to other hospitals, countries, time periods, etc. Usually, there are no data or funding available to do external validation. Hence, when a prediction model is developed a good internal validation should be ensured at the least.

To sum up, the research on predicting outcomes from multiple variables has three main steps: development of the prediction model, validation of the model in new individuals and study of its application and impact in clinical practice. The aim of this dissertation is to propose a valid methodology for categorising continuous predictor variables whenever a clinical researcher considers it necessary to include a categorised version in the prediction model. This is framed mainly in the first phase, that is to say, in the development of the model. However, in the development of the proposed methodology we have accorded great importance to the validation of the predictive ability as well as the interpretation and practical applicability of the proposed categorical variable.

1.2 Categorisation of predictors

A vital factor in the development of prediction models is the selection of the predictors or covariates (clinical variables) to be used in the model. From a statistical perspective, categorising continuous variables is not advisable, since it may entail a loss of information and power (Altman and Lyman 1998, Cohen 1983, MacCallum et al. 2002, Royston et al. 2006) and a loss of efficiency if the correlation between the categorised and response variable is high (Taylor and Yu 2002). Additionally, there are statistical modelling techniques such as generalised additive models (GAM) (Hastie and Tibshirani 1990, Wood 2006), which do not require any assumption of linearity between predictors and response variables, and so allow for the relationship between the predictor and the outcome to be modelled more appropriately. Yet in clinical research and, more specifically, in the development of prediction models for use in clinical practice, both clinicians and health managers have called for the

categorisation of continuous parameters. Indeed, in a recent survey of the epidemiological literature, in 86% of the papers included in the study, the primary continuous predictor was categorised, and 78% used three to five categories (Turner et al. 2010). Additionally, of the seven prediction models mentioned in Section 1.1 above, in six of them categorised continuous variables are used as predictors, being the number of categories used between 3 and 5 (see Table 1.1) when more than two categories are considered. None of them use continuous variables in the prediction model but nevertheless as regards those variables studied in the univariate analysis it is not specified whether the linearity is fulfilled.

Table 1.1: Summary of the use of categorised continuous variables in prediction models

Prediction model	Use of categorised predictors	Number of categories	Selection of cut points
Framingham risk score (Wilson et al. 1998)	Yes	5	Clinical guidelines
Diabetes risk score (Lindström and Tuomilehto 2003)	Yes	3	Based on previous research
BODE - index (Celli et al. 2004)	Yes	4	Clinical guidelines
COPD mortality score (Quintana et al. 2014a)	Yes	3	Not specified
COPD prognostic severity score (Quintana et al. 2014b)	Yes	3 or 4	Based on previous research
Risk of COPD in primary care (Haroon et al. 2015)	No	-	-
COPD short-term risk (Make et al. 2015)	Yes	3	Tertiles corrected to clinically relevant cut points

There are several reasons for incorporating categorical variables in prediction models. First, in clinical practice, the implementation of the results obtained from techniques such as GAM is not always viable. It requires specific software which it is not always possible to use in consulting rooms or emergency departments. On the other hand, decisions in clinical practice are often taken on the basis of an individual patient's risk level, which is strongly related to the categorisation of that patient's

clinical variables. Yet, despite the fact that categorisation is a common practice in clinical research, there are no unified criteria for categorising continuous variables. Indeed, categorisation is very often based on percentiles, even though this is known to have drawbacks (Bennette and Vickers 2012). Moreover, even when categorisation is based on clinical criteria, it has been shown that it can vary enormously from one practitioner, hospital or even country to another. For instance, a meta-analysis conducted by Lim and Kelly (2010) showed that reported cut-off values for partial pressure of carbon dioxide in the blood (PCO_2) for hypercapnia screening ranged from 30 to 46 mmHg. In addition, an optimal categorisation may provide an understandable summary and simple interpretation of the results obtained with minimal loss of information (Gelman and Park 2009), mainly when more than two categories are considered.

Work has been done on the categorisation of continuous variables. A review of these methods shows that these have been based first, on the graphical relationship between the predictor and the outcome, second, on percentiles and, third, on the minimum p-value approach (Mazumdar and Glassman 2000). For the first, scatter plots, grouped data plots or model-based plots (Hin et al. 1999) could be used. In the third, of all possible cut points, the cut point for dichotomisation chosen is that for which the maximum chi-squared statistic (or minimum p-value) is obtained (Miller and Siegmund 1982). Altman et al. (1994) showed that the minimum p-value approach yields an increase in the false positive error rate and thus proposed a correction for the minimum p-value formulae. Latter, Faraggi and Simon (1996) proposed a cross-validation approach to estimate the minimum p-value. On the other hand, O'Brien (2004) proposed a categorisation as the partition which lead to the minimum average distance between the true and estimated expected values of the outcome for subjects in the same category. Moreover, the aim in almost all cases is to seek a single cut point, or, to put it another way, to dichotomise the continuous predictor. A particular case is the dichotomisation of an estimated probability for test markers or biomarkers for diagnostic (diseased or non-diseased) or screening purposes. In this context, the maximisation of the Youden index (Fluss et al. 2005, Youden 1950) and the point on the receiver operating characteristic (ROC) curve closest to the point (0,1) (Metz 1978, Vermont et al. 1991), have been proposed, among others. This is a research area of interest, as evidenced by recent publications (López-Ratón et al. 2014, Rota and Antolini 2014, Rota et al. 2015). However, this is not the aim of the work we present in this dissertation. We focus on the categorisation of continuous variables to be used in the development

of prediction models, considering that the use of more than two categories may be preferable. This serves to reduce the loss of information and enables the relationship between the covariate and the response variable to be retained. For example, for a variable such as blood pressure, it is not possible to classify patients into high and low risk categories by using a unique cut point, since low and high values of blood pressure are indicators of high risk. Thus, in this case at least two cut points, i.e. three categories would be needed.

In the context where the outcome of interest takes only two possible values, the search for more than one cut point has been considered, for instance, by Tsuruta and Bax (2006). Tsuruta and Bax propose a parametric method for obtaining cut points based on the overall discrimination c statistic (Harrell et al. 1982). The authors showed the optimal location of cut points in a case where the distribution of the predictor variable is known, and illustrated the proposal for application to a normal distribution. Yet, in routine clinical practice and, by extension, in medical research, variables of interest do not usually respond to either a normal or a known distribution.

While developing a multivariate prediction model for COPD patients in different studies that will be presented in detail in Chapter 2, clinical researchers and epidemiologist encouraged us to categorise several continuous variables. We realised there were no previously specified cut points for those variables and no methodology to do so. That problematic discovery motivated the work presented in this dissertation. Without recommending categorisation as an ideal solution, the aim of this work is to propose valid methods for categorising continuous variables whenever a clinical researcher considers it necessary.

1.3 Organisation of subsequent chapters

The rest of this dissertation is organised as follows. In Chapter 2 the motivating data sets are presented. More precisely, the IRYSS-COPD study of patients with exacerbated COPD (eCOPD) is presented in the first place. In the development of prediction models for eCOPD patients, the need to categorise several clinical variables obtained from the blood gasometry was presented. This fact motivated the development of a valid methodology to categorise continuous variables. In addition, the Stable-COPD study is presented. This data set of patients suffering from stable COPD (sCOPD), is used as the motivating data set for the categorisation of continuous variables when the response variable is time to event.

Chapter 3 is devoted to the presentation of the main statistical methods used throughout the dissertation. What is shown in this chapter is the basis of the proposals that are made in subsequent chapters. These include the generalised linear model (GLM), the GAM and the Cox proportional hazards (Cox PH) regression model. Furthermore, the most commonly used discriminative ability measures are presented.

Our first approximation to the categorisation of continuous variables in logistic regression models was based on a graphical display based on a GAM. This methodology is presented in Chapter 4, together with its validation and implementation in the IRYSS-COPD study.

Chapter 5 is devoted to the development of a methodology for categorising continuous predictor variables in a logistic regression setting as an improvement on what is presented in Chapter 4. In this proposal the optimal categorisation of continuous variables is based on the maximisation of the area under the ROC curve (AUC). The categorisation can be done for any number of cut points in addition to being able to select the optimal number of cut points. The methodology has been validated in theory and in practice when theoretical cut points are known. It has also been applied in the IRYSS-COPD study.

Chapter 6 is an extension to the methodology presented in Chapter 5 where the prediction model considered is a Cox PH regression model. In this setting, different discrimination index estimators have been proposed and compared as the target of maximisation. The methodology has been validated and applied in the Stable-COPD study data set.

In our opinion, the development of an easy-to-use tool to allow the implementation of the proposed methodology in practice was an important goal of this work. Hence, we developed the `CatPredi` package in software R (R Core Team 2014) which allows the user to categorise a continuous variable either in a logistic regression or a Cox PH model (methodologies presented in Chapter 5 and Chapter 6 respectively). A detailed description of this package is given in Chapter 7. Additionally, an R function is given as an easy way to implement the graphical display-based methodology presented in Chapter 4.

Chapter 8 ends this dissertation with general conclusions about the methodologies presented and the work developed so far, in addition to the objectives of future research.

Most of the scientific results provided by this work have already been presented to the scientific community as research articles or communications in conferences. At

the beginning of each chapter we present the main results related to the methodology proposed in that chapter. In addition, in Chapter 8, we enumerate all the scientific results provided by the work presented in this dissertation.

Motivating data sets

In this chapter we describe in detail the two data sets that motivated the research presented in this dissertation, which were both studies of patients with COPD. COPD is one of the most common chronic diseases, and its prevalence is expected to increase over the next few decades (Buist et al. 2008). COPD is a leading cause of death in developed countries, and patients with COPD generally suffer a substantial deterioration in their quality of life (Esteban et al. 2009). The exacerbation of COPD is defined as an event in the natural course of a patient’s COPD characterised by a change in baseline dyspnoea, cough and/or sputum, that is beyond normal day-to-day variations and that may have warranted a change in medication or treatment (Rabe et al. 2007). In the following sections we present the two research studies that have motivated the work presented in this dissertation. The first study, the IRYSS-COPD study, concerns patients with exacerbated COPD (eCOPD), and the second, the Stable-COPD study, focuses on patients with stable COPD (sCOPD).

2.1 The IRYSS-COPD study

The IRYSS-COPD study (IRYSS: Red de investigación cooperativa para la Investigación en Resultados de Salud y Servicios Sanitarios - Cooperative Health Outcomes & Health Services Research Network) was created to address gaps in identifying eCOPD patients whose clinical situation is appropriate for admission to hospital, and to develop and validate severity scores for eCOPD patients (Quintana et al. 2011). In this study, a sample of 2877 episodes corresponding to 2487 patients with eCOPD attending the emergency departments (ED) of 16 participating hospitals in Spain was collected between June 2008 and September 2010. Information was

recorded as follows: at the date on which patients were evaluated at the ED; at the date on which the decision was made to admit patients or discharge them from the ED; and during follow-up after admission to hospital or discharge. Data collected upon arrival in the ED included socioeconomic data, information about the patient's respiratory function (arterial blood gases, respiratory rate, dyspnoea, forced expiratory volume in one second in percentile ($FEV_{1\%}$)), presence of other pathologies recorded in the Charlson Comorbidity Index (Charlson et al. 1987) and consciousness level measured by the Glasgow Coma Scale which was dichotomised as follows: altered consciousness defined as a score of < 15 points, unaltered consciousness as a score of 15 points (Teasdale and Jennett 1974). Additional data collected in the ED at the time a decision was made to admit or discharge the patient included the patient's symptoms, signs, and respiratory status at that point. Quintana et al. (2011) provide a detailed description of the IRYSS-COPD study. Furthermore, a description of the main selected variables is given in Table 2.1.

Table 2.1: A description of the selected variables from the IRYSS-COPD study

Variable	Available N	Mean (sd)	Range
Age	2876	72.84 (9.51)	36 - 96
Sex^a	2874		
Men		2627 (91.41%)	
Female		247 (8.59%)	
FEV_{1%}	2430	44.57 (16.76)	17 - 149
Charlson Index	2877	2.25 (1.55)	1 - 13
Glasgow^a	2874		
Normal		2797 (97.32%)	
Altered		77 (2.68%)	
Respiratory Rate	2314	25.04 (6.67)	10 - 56
PCO₂	2485	47.49 (14.08)	13 - 160
Heart Rate	2697	95.05 (18.90)	21 - 190

^aCategorical variables are shown as absolute and relative frequencies

Currently, ED physicians must rely largely on their experience and the patient's personal criteria to gauge how an eCOPD will evolve. A clinical prediction rule that could help predict eCOPD evolution would allow ED physicians to make better informed decisions about treatment. Therefore, one of the goals of the IRYSS-COPD study was to develop clinical prediction rules. To this end, two main outcomes were defined, short-term poor evolution and very severe evolution (Quintana et al. 2011). The definition of both outcomes is displayed in Table 2.2.

For the first outcome of interest, i.e. poor evolution, we first evaluated which vari-

Table 2.2: Description of the outcomes defined for the development of clinical prediction rules in the IRYSS-COPD study

Outcome	Description
Poor evolution	Includes any of the following: death, ICU admission, the need for IMV, cardiac arrest, NIMV for more than 2 days when mechanical ventilation was not needed before admission, and/or admission to an IRCU for 2 or more days
Very severe evolution	Includes any of the following: death, ICU admission, need for IMV, and/or cardiac arrest

ICU: intensive care unit ; IMV: invasive mechanical ventilation; NIMV: non-invasive mechanical ventilation; IRCU: intermediate respiratory care unit.

ables were related to the outcome in a univariate setting. We noted, among others, that clinical variables obtained from arterial blood gases such as PCO₂, PO₂ or pH or the respiratory rate (RR), were strongly related to poor evolution. However, for some of these variables not all the data needed were available in the clinical records, and even when they were available, they were not always in a desirable format, appearing, for instance, as a description of patient status rather than a numerical value, e.g. some patients' RR was recorded as "eupneic" or "taquibneic" instead of being cited as a number on a continuous scale. Despite the fact that clinicians failed to agree on the cut points to apply to each code, they did, in contrast, regard the "eupneic" and "taquibneic" patients as "normal" and "altered" respectively, which means that, although they would be able to classify such patients on a categorical scale, they would nevertheless leave them as missing data on a continuous scale. In this context, clinicians encouraged us to find the best categorisation for this variable to facilitate reconciliation of information that was partially available as a continuous variable and partially available as an ordinal variable. In order to meet this demand, we developed the initial approach described in Chapter 4.

The second goal for the IRYSS-COPD study researchers was to develop a prediction model for very severe evolution. After a preliminary analysis, the predictors selected for inclusion in the multivariate model for very severe evolution were the Glasgow Coma Scale, heart rate and the arterial blood gas PCO₂. However, the covariate PCO₂ did not have a linear relationship with the outcome and hence it had to be modelled with a smooth function or in a categorised version. The clinical researchers involved in the study opted for a categorised version of this predictor,

but there were no previously fixed cut point criteria in the literature (Lim and Kelly 2010). Furthermore, clinicians did not agree on the best number of categories for this variable. Hence, we considered developing a methodology to resolve these problems: 1) obtain the optimal cut points for any given number of cut points either in a univariate or in a multivariate setting; and 2) select the optimal categorisation by comparing different numbers of cut points. This methodology is presented in Chapter 5.

2.2 The Stable-COPD study

In this study patients being treated for COPD at five outpatient respiratory clinics affiliated with the Hospital Galdakao-Usansolo in Biscay between January 2003 and January 2004 were recruited (Esteban et al. 2014). Patients were consecutively included in the study if they had been diagnosed with COPD for at least six months and had been receiving medical care at one of the hospital respiratory outpatient facilities for at least six months. Their COPD had to be stable for six weeks before enrolment. Patients were followed for up to five years. Details of the follow-up are shown in Figure 2.1 and the main selected variables collected in this study are summarised in Table 2.3.

The main goal of this study was to develop prediction models for patients with sCOPD. Several outcomes were considered of interest. These included, among others, short-term mortality, five-year survival, frequency of hospitalisation and health-related quality of life.

An important predictor for COPD mortality or hospitalisation is $FEV_{1\%}$, which is commonly used by clinicians to diagnose and measure the severity of the disease (Vestbo et al. 2013). Recently, several multivariate prediction models which include a categorised version of $FEV_{1\%}$ among the predictor variables have shown a better survival prediction than isolated $FEV_{1\%}$. Among others, the most commonly used prediction models are the original BODE index (Celli et al. 2004), ADO index (Puhan et al. 2009), HADO index (Esteban et al. 2006), SAFE (Azarisman et al. 2007) and DOSE (Jones et al. 2009). A description of the variables included in each index is given in Table 2.4. Although all prediction models use a categorised version of the predictor variable $FEV_{1\%}$, not all of them use the same cut points. The criteria used in each index to categorise $FEV_{1\%}$ are summarised in Table 2.5. To date, the most widely-used cut points are the ones proposed by the Global Obstructive Lung Disease (GOLD) guidelines (mild ≥ 80 , moderate 50-79, severe 30-49 and very

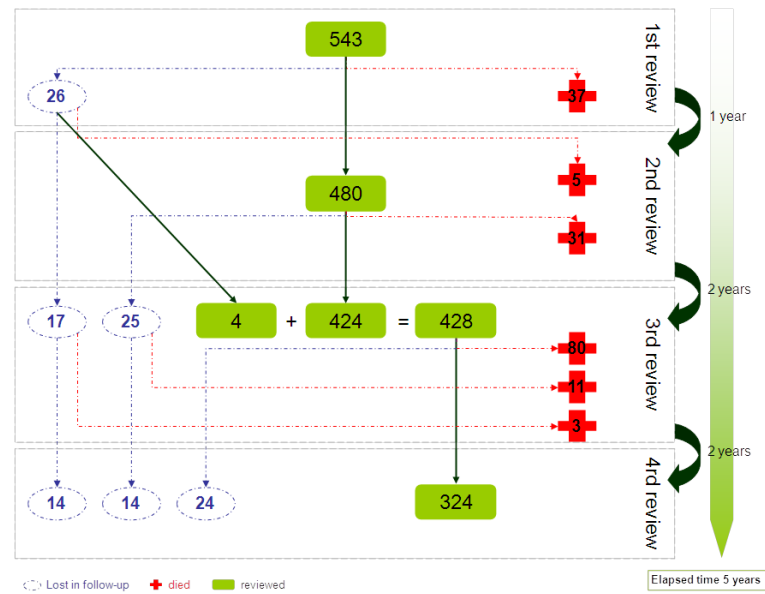


Figure 2.1: Follow-up flowchart of the Stable-COPD study. This graphic is the copyright of Dr Cristobal Esteban, principal researcher of the project and author of the article in which it is published (Esteban et al. 2014).

Table 2.3: A description of the selected variables from the Stable-COPD study.

Variable	Available N	Mean (sd)	Range
Age	543	68.32 (8.32)	33 - 86
Sex^a	543		
Men		522 (96.13%)	
Female		21 (3.87%)	
FEV₁%	543	55 (13.31)	18 - 105
BMI	543	28.28 (4.43)	16.38 - 44.04
Dyspnoea^a	543		
1		69 (12.71)	
2		264 (48.62)	
3		166 (30.57)	
4		23 (4.24)	
5		21 (3.87)	
Walking distance	543	408.89 (92.43)	46 - 644
Time until event (days)	543	1574.89 (483.43)	23 - 2045
5-year mortality	543		
Yes		167 (30.76)	

^aCategorical variables are shown as absolute and relative frequencies
Dyspnoea was measured with the modified scale of the Medical Research Council (mMRC) (Fletcher et al. 1959).

severe < 30) (Rabe et al. 2007).

Recently, Almagro et al. (2014) proposed a new categorisation of $FEV_{1\%}$ to predict five-year survival in COPD patients. This research was framed within the Collaborative Cohorts to Assess Multicomponent Indices of COPD in Spain (CO-COMICS) study. For this study a number of cohort studies of COPD patients with different stages of COPD were grouped together with the aim of assessing more accurately the survival of COPD patients.

Hence, and taking all this into account, three factors motivated us to look for the best categorisation of the variable $FEV_{1\%}$ in the prediction model developed in the Stable-COPD study. First of all, this variable is an important predictor for predicting five-year survival for sCOPD patients. Since other prediction models and especially clinical guidelines use a categorised version of this variable, the clinicians involved in the study considered it was necessary to include a categorised version of this variable in the prediction model. Second, recent research shows the importance of seeking optimal cut points for this variable. Third, to date there are no unified criteria on how to categorise the variable $FEV_{1\%}$. Consequently, we used the data set of 543 patients with sCOPD in the Stable-COPD study and in particular the $FEV_{1\%}$ predictor variable for the methodology developed in Chapter 6.

Table 2.4: Description of the existing indexes for predicting the severity of COPD patients.

Prediction Scores	Description
BODE	<ul style="list-style-type: none"> • Body mass index • Airflow obstruction measured by $FEV_{1\%}$ • Dyspnoea • Walked distance in 6 minutes
ADO	<ul style="list-style-type: none"> • Age • Dyspnoea • Airflow obstruction measured by $FEV_{1\%}$
HADO	<ul style="list-style-type: none"> • Overall health status • Level of physical activity • Dyspnoea • Airflow obstruction measured by $FEV_{1\%}$
SAFE	<ul style="list-style-type: none"> • Quality of life measured by Saint George's Respiratory Questionnaire • Airflow obstruction measured by $FEV_{1\%}$ • Walked distance in 6 minutes
DOSE	<ul style="list-style-type: none"> • Dyspnoea • Smoking status • Airflow obstruction measured by $FEV_{1\%}$ • Prior exacerbation history

Table 2.5: Airflow obstruction level measured by $FEV_{1\%}$ based on the different cut points used in the literature to categorise the continuous $FEV_{1\%}$ variable.

Criteria	Mild	Moderate	Severe	Very Severe
SAFE	≥ 80	[50 – 80)	[30 – 50)	< 30
GOLD				
DOSE		≥ 50	[30 – 50)	< 30
BODE	≥ 65	[50 – 65)	(35 – 50)	≤ 35
ADO		≥ 65	(35 – 65)	≤ 35
HADO	> 65	[50 – 65]	[35 – 50)	< 35
COCOMICS	≥ 70	(55 – 70)	(35 – 55]	≤ 35

Chapter 3

Methodological background and preliminaries

In this chapter we introduce the general notations used as well as the main statistical models we use throughout the paper, i.e. GLM, GAM and the Cox PH regression model.

Specifically, in Section 3.1 we introduce the GLM in general and the estimation method for the logistic regression model for a binary outcome in particular. The GLM assumes a linear relationship between the predictors and a function of the expected outcome, which is the logit function when a binary outcome is considered. When this linearity does not hold, an extension of the GLM appears which is known as GAM. In Section 3.2 we introduce the GAM and present the most common alternatives for the estimation of the smooth functions which are used to model the nonlinear effects of the predictor variables. The first two sections of this chapter focus on a binomial distribution of the response variable or the disease status. However, in many circumstances the disease status is not a fixed characteristic of the study. Such is the case of survival studies where the status of an individual varies with time. In such cases, the interest is then focused on the time of the occurrence of the event of interest. The most common alternative for modelling time-to-event data is the Cox PH model introduced in Section 3.3. Finally, in Section 3.4 we introduce the concordance probability as the most commonly used discriminative ability measure in prediction models. In the logistic regression setting the AUC is introduced whereas in the Cox PH model two alternative estimators for the concordance probability are presented. In addition, in this Section we discuss different approaches to compare the AUCs of two prediction models, as well as existing proposals for the bias correction in concordance probability estimators.

3.1 Generalised linear model

Suppose we have a response variable Y with some exponential family distribution and a set of predictor variables $\mathbf{Z} = (Z_1, \dots, Z_p)$ which can be either continuous or categorical variables. Often one may be interested in studying how this set of predictor variables \mathbf{Z} is related to the response variable Y . The first attempt to do so, may be to fit a GLM, assuming there exists a linear relationship between the predictors and some function of the expected outcome. Then the GLM has the form

$$g(E(Y|\mathbf{Z})) = \mathbf{Z}\boldsymbol{\beta}' = \beta_0 + \beta_1 Z_1 + \dots + \beta_p Z_p, \quad (3.1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ is the vector of the unknown regression coefficients and g is the *link* function. For ease of notation, we have included the unit term in the vector of the predictor variables, that is, $\mathbf{Z} = (1, Z_1, Z_2, \dots, Z_p)$ and represented the model as if all the variables were continuous.

In particular, when Y is a binary response variable and the link function g is the *logit* function, then the GLM is known as the logistic regression model. Let us assume that the outcome variable has been coded as zero or one, representing the absence or the presence of the event respectively. Thus the expected value for Y , $E(Y)$ is the probability of having the outcome of interest, i.e. $P(Y = 1)$. To simplify notation we use $\pi(\mathbf{Z}) = P(Y = 1|\mathbf{Z})$ to represent the conditional expectation of Y given \mathbf{Z} when Y is binary. Then, the multivariate logistic regression model for Y is written as a linear function in the logistic transformation (*logit*) of the conditional probability that the outcome is present, as shown in equation (3.2)

$$\text{logit}(\pi(\mathbf{Z})) = \ln \frac{\pi(\mathbf{Z})}{1 - \pi(\mathbf{Z})} = \beta_0 + \beta_1 Z_1 + \dots + \beta_p Z_p, \quad (3.2)$$

or equivalently,

$$\pi(\mathbf{Z}) = \frac{\exp(\mathbf{Z}\boldsymbol{\beta}')}{1 + \exp(\mathbf{Z}\boldsymbol{\beta}')}. \quad (3.3)$$

Let $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$ be a random sample drawn from (\mathbf{Z}, Y) , where \mathbf{z}_i represents the observed value of the predictor variables and y_i represents the observed binary response for subject i . The regression coefficients of model (3.2) are usually estimated by maximum likelihood. In this case, the likelihood function is given by,

$$l(\boldsymbol{\beta}) = \prod_{i=1}^N \pi(\mathbf{z}_i)^{y_i} (1 - \pi(\mathbf{z}_i))^{1-y_i}. \quad (3.4)$$

The principle of maximum likelihood states that the best value to assign to $\boldsymbol{\beta}$ is that for which equation (3.4) is maximised. Nevertheless for ease of mathematical calculation, the logarithm of the expression in equation (3.4) is maximised. This results in the *log likelihood* which is defined as,

$$L(\boldsymbol{\beta}) = \sum_{i=1}^N \{y_i \ln [\pi(\mathbf{z}_i)] + (1 - y_i) \ln [1 - \pi(\mathbf{z}_i)]\}. \quad (3.5)$$

To find the vector $\boldsymbol{\beta}$ which maximises equation (3.5), $L(\boldsymbol{\beta})$ is differentiated with respect to $\beta_0, \beta_1, \dots, \beta_{p-1}$ and β_p , which leads to the $(p + 1)$ likelihood equations that may be expressed as

$$\sum_{i=1}^N [y_i - \pi(\mathbf{z}_i)] = 0 \quad (3.6)$$

and

$$\sum_{i=1}^N z_{ij} [y_i - \pi(\mathbf{z}_i)] = 0 \quad \text{for } j = 1, \dots, p. \quad (3.7)$$

Frequently an iterative weighted least squares algorithm is used to solve the likelihood equations in (3.6) and (3.7). We will denote as $\widehat{\beta}_0$ and $\widehat{\beta}_1, \dots, \widehat{\beta}_p$ the estimated values of β_0 and β_1, \dots, β_p . More detail about estimation methods in GLM can be seen in McCullagh and Nelder (1989).

3.2 Generalised additive model

The GAM is an extension of the GLM where the modelling of the effect of the covariates is relaxed by not assuming linearity. Albeit the effect of some covariates may be assumed to be linear, the nonlinear effects are modelled using smoothing methods, such as kernel smoothers (Wand and Jones 1994), smoothing splines (Green and Silverman 1994) or regression splines (De Boor 2001, Hastie et al. 2009). In general, the model has the following structure

$$g(E(Y|\mathbf{Z})) = \beta_0 + \sum_{j=1}^p f_j(Z_j). \quad (3.8)$$

Model (3.8) is an extension of model (3.1) where $f_j(\cdot)$ are some smooth and known functions of the covariates Z_j for each $j = 1, \dots, p$. More specifically, if Y is a binary response variable, then the logistic GAM is expressed as

$$\text{logit}(\pi(\mathbf{Z})) = \beta_0 + \sum_{j=1}^p f_j(Z_j). \quad (3.9)$$

The main drawback of GAMs lies in the estimation of the smooth functions $f_j(\cdot)$, and there are different ways to address this. The most recent approaches are based on splines, which allow the GAM estimation to be reduced to the GLM context (Currie et al. 2006).

Splines are piecewise polynomials, pieces defined by a sequence of m knots $\zeta_1 < \zeta_2 < \dots < \zeta_m$, in such a way that pieces join smoothly at these knots. Splines depend on three elements: the degree of the polynomial, the number of knots and the location of these knots. There are two major approaches to smooth modelling with splines: 1) smoothing splines and 2) regression splines. Smoothing splines use as many knots as observations and incorporate a penalisation on the second derivative (Green and Silverman 1994). This implies that its implementation is not efficient when the amount of available data is very high. Regression splines can be fitted using the least squares method once the number of knots has been selected. However, the selection of the knots' location is done with complex algorithms.

An intermediate alternative for building the smooth functions, which considers both the smoothing splines' and the regression splines' advantages is the use of penalised splines also known as P-splines and introduced by Eilers and Marx (1996). P-splines use fewer knots than smoothing splines; in fact, the number of knots used in P-splines is no larger than 40 which makes it computationally more efficient than smoothing splines. Additionally, P-splines introduce more general roughness penalties which relax the importance of the knots' location. Thus, the number of knots ensures flexibility, and the penalty avoids over-fitting and ensures smoothness.

The methodology for constructing a P-spline has two main steps: 1) choose a basis for the regression, and 2) modify the likelihood function by introducing a penalty based on the differences between adjacent coefficients. The most common alternatives for the first step are truncated polynomials, thin plate regression splines (Wood 2003) and B-splines (De Boor 2001). In this dissertation we will focus on the B-spline basis. In general, a B-spline basis of degree r comprises:

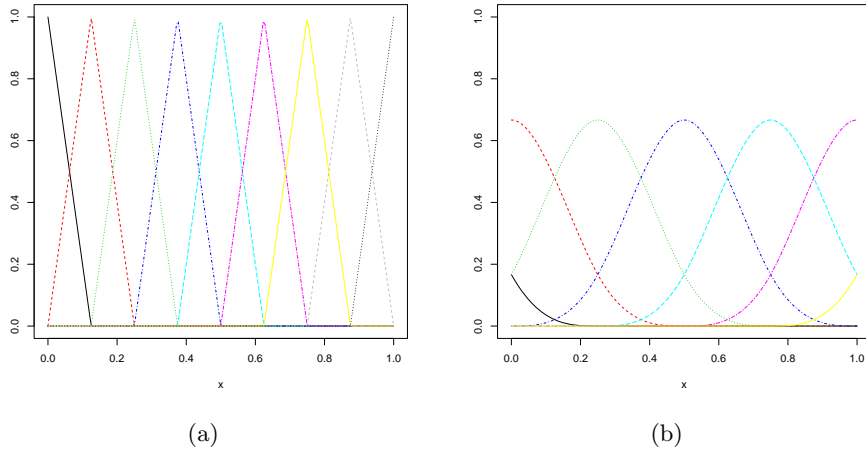


Figure 3.1: (a) B-spline basis functions of degree $r = 1$ and $m = 9$ inner knots.
(b) B-spline basis functions of degree $r = 3$ and $m = 5$ inner knots.

1. $(r + 1)$ piecewise polynomials, each of degree r
2. These $(r + 1)$ piecewise polynomials join at r inner knots.
3. At the junction point, the derivatives up to $r - 1$ order are continuous.
4. The B-spline is positive in a domain covered by $r + 2$ knots and zero otherwise.
5. Except at the frontiers, each B-spline overlaps with $2r$ neighbouring piecewise polynomials.
6. For each value z_{ij} of Z_j , there are $r + 1$ non-zero B-splines.

This means a B-spline basis is independent of the response variable. The smooth functions $f_j(\cdot)$ are represented in terms of B-spline basis functions depending on the following factors: 1) the range of the independent covariate Z_j ; 2) the number and location of the knots; and 3) the degree of the B-spline. Figure 3.1 shows two examples of B-spline basis functions with $m = 9$ inner knots and degree $r = 1$ and $m = 5$ inner knots and degree $r = 3$ respectively.

With this smoothing approach, for each GAM component, the smooth function

is reduced to a linear combination of $d_j = m_j + r_j - 1$ B-splines,

$$f_j(\cdot) = \sum_{l=1}^{d_j} B_{jl}(\cdot)\beta_{jl}, \quad (3.10)$$

where $\beta_j = (\beta_{j1}, \dots, \beta_{jd_j})$ is a vector of unknown regression coefficients and $B_{jl}(\cdot)$ is a B-spline basis.

We have seen so far how to address the representation of each smooth function $f_j(\cdot)$ in equation (3.9). However, without imposing constraints this model presents an identifiability problem, because it incorporates more than one predictor variable. We could subtract a constant ξ of any smooth function ($f_1(z_1) - \xi$), and add it to another one ($f_p(z_p) - \xi$), and the same regression model would be obtained. To avoid this problem, it is necessary to impose some restrictions. The usual way to guarantee the identification of the model is to incorporate a constant β_0 , and to “centre” the smooth functions in some way, for instance by assuming:

$$E(f_j(Z_j)) = 0 \quad \text{for } j = 1, \dots, p.$$

Given a sample $\{(z_i, y_i)\}_{i=1}^N$, the matrix representation of model (3.9) based on P-splines can be given in this way,

$$\text{logit}(\pi(\mathbf{z})) = \mathbf{B}\beta' \quad (3.11)$$

where \mathbf{z} denotes the sample predictor variables, $\beta = (\beta_0, \beta_{11}, \dots, \beta_{1d_1}, \dots, \beta_{p1}, \dots, \beta_{pd_p})$ is the coefficients vector, d_j is the number of B-splines for the j^{th} covariate, for each $j = 1, \dots, p$ and \mathbf{B} is the $N \times (1 + \sum_{j=1}^p d_j)$ regressor matrix defined as

$$\begin{pmatrix} 1 & B_{11}(z_{11}) & \cdots & B_{1d_1}(z_{11}) & \cdots & B_{p1}(z_{1p}) & \cdots & B_{pd_p}(z_{1p}) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 1 & B_{11}(z_{N1}) & \cdots & B_{1d_1}(z_{N1}) & \cdots & B_{p1}(z_{Np}) & \cdots & B_{pd_p}(z_{Np}) \end{pmatrix}$$

In general, the estimation method consists of maximising the penalised version of the log likelihood expressed as:

$$L^* = L(\boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^p \lambda_j \boldsymbol{\beta}_j \mathbf{P}_j \boldsymbol{\beta}_j' \quad (3.12)$$

where the term $L(\boldsymbol{\beta})$ represents the log likelihood of the vector of the response variable Y as presented in (3.5). For each $j = 1, \dots, p$ the $\lambda_j \geq 0$ are the smoothing parameters and \mathbf{P}_j is a $d_j \times d_j$ dimension matrix that defines the penalty for the j^{th} smooth function. The estimation method is explained in detail in Marx and Eilers (1998).

3.3 Cox proportional hazards model

As we mentioned before, survival analysis is used to analyse data in which the time until the event of interest occurs is the response variable, which is frequently called *survival time* or *event time*. This response variable is generally continuous, but survival analysis allows that the variable is not fully determined for some subjects. For example, in a survival study that explores five-year mortality after surgery for colon cancer, if a patient is still alive at five years, it is known that survival is greater than five years, but the exact value is not known. Hence that patient's survival time is *censored* on the right. Censoring can also occur when an individual is lost to follow-up. Different types of censorship exist, but in this dissertation we will focus on right-censoring in which individuals' survival time is observed only if the event occurs before a certain time, but allow different censoring times between individuals.

Even if there is no censoring, there are several reasons to use survival analysis to model the time until the event rather than standard linear regression models. On the one hand, time to event is restricted to be positive, usually with a skewed distribution, and thus it does not meet the normality assumption. On the other hand, the probability of surviving past a certain time is often more interesting than the expected survival time.

Let T be a non-negative random variable representing the time to the event of interest and let us denote by C the random right-censoring variable. Instead of defining the statistical model for the response T in terms of the expected survival time, it is worth defining it in terms of the survival function, $S(t)$, given by

$$S(t) = P(T > t) = 1 - F(t), \quad (3.13)$$

where $F(t)$ is the cumulative distribution function for T . As an example, Figure 3.2 represents the estimated survival function (using Kaplan-Meier estimator (Kaplan and Meier 1958)) for the Stable-COPD study presented in Chapter 2. The hazard function, $h(t)$, also called instantaneous event rate, is defined as

$$h(t) = \lim_{\nu \rightarrow 0} \frac{P(t < T \leq t + \nu | T > t)}{\nu}, \quad (3.14)$$

and its integral can be based on the survival function such that

$$\int_0^t h(v) dv = -\ln S(t). \quad (3.15)$$

Thus, the hazard function at time t is related to the probability that the event will occur in a small interval around t , given that the event has not occurred before time t .

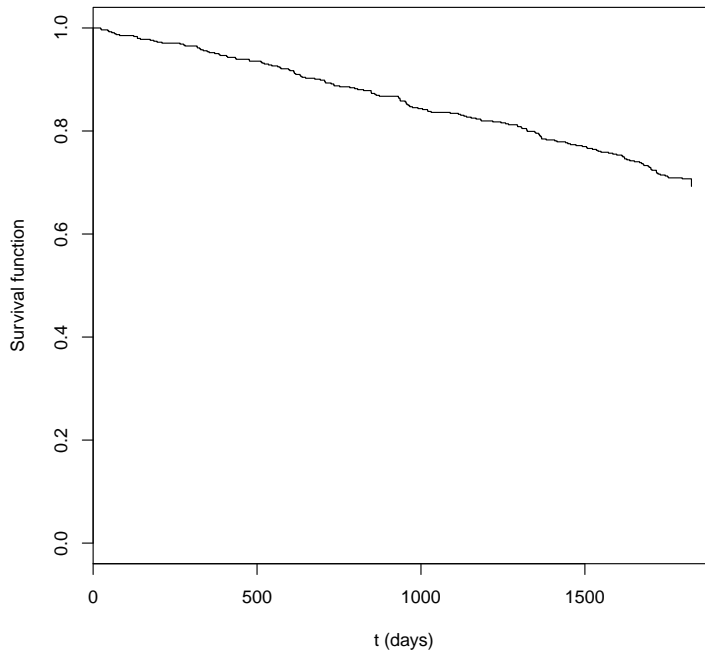


Figure 3.2: Survival function for patients in the Stable-COPD study.

Let $\mathbf{Z} = (Z_1, \dots, Z_p)$ be a set of predictor variables in which we are interested in

terms of studying the relationship with the survival time T . The most widely used survival regression specification allows the hazard function $h(t)$ to be multiplied by $\exp(\mathbf{Z}\boldsymbol{\beta}')$. Thus, the hazard function for T in a time t given the covariates \mathbf{Z} is given by,

$$h(t|\mathbf{Z}) = h(t) \exp(\mathbf{Z}\boldsymbol{\beta}') \quad (3.16)$$

where $\boldsymbol{\beta}$ is the regression coefficients vector. This expression is called the *proportional hazards* (PH) model. If a parametric hazard function is used for $h(t)$ then the model is called the parametric proportional hazards model. Commonly used parametric forms for the hazard function are based on the exponential and Weibull distributions. For the former the hazard function takes a constant value and for the latter the hazard function can be expressed as $h(t) = \lambda\gamma t^{\gamma-1}$, where λ and γ are usually called scale and shape parameters respectively. Conversely, the hazard function can also be left completely unspecified in equation (3.16), yielding the Cox semiparametric proportional hazards model (Cox 1972). This model is the most commonly used regression model for analysing survival data.

Note that the model in equation (3.16) can be rewritten as

$$\ln h(t|\mathbf{Z}) = \ln h(t) + \mathbf{Z}\boldsymbol{\beta}'. \quad (3.17)$$

This implies that the effect of the covariates \mathbf{Z} is assumed to be the same at all values of t since $\ln h(t)$ can be separated from $\mathbf{Z}\boldsymbol{\beta}'$. The regression coefficient for Z_j , β_j , is the increase in the logarithm of the hazard at any fixed time t if Z_j is increased by one unit and all the rest of the covariates are kept constant (assuming Z_j is continuous). Usually the effect a covariate Z_j has on the response variable time to event is measured by the hazard ratio (HR), which is estimated as the exponential of the regression coefficient β_j .

For estimating and testing the regression coefficients β_0, \dots, β_p , the Cox PH model is as efficient as the proportional hazards parametric model even when all the assumptions of the parametric model are satisfied (Efron 1977). The estimation of the regression coefficients β_0, \dots, β_p is carried out by the maximization of the partial likelihood function proposed by Cox (1972). The construction of the partial likelihood is based on the observed events but does not explicitly consider the censored individuals.

Let $\{\mathbf{z}_i, y_i, \delta_i\}_{i=1}^N$ be a sample of size N , where \mathbf{z}_i represents the observed value of the predictor variables for subject i , y_i represents the observed follow-up time for

subject i , being the minimum between the censoring (c_i) and the event time (t_i), i.e. $y_i = \min(t_i, c_i)$, and δ_i represents whether subject i is an event ($\delta_i = 1$) or is censored ($\delta_i = 0$). Thus, $\delta_i = I(t_i \leq c_i)$.

Assume for now that there are no tied event times in the sample, so the partial likelihood is defined as,

$$l(\beta) = \prod_{i:\delta_i=1} \frac{e^{\mathbf{z}_i\boldsymbol{\beta}'}}{\sum_{l:y_l \geq y_i} e^{\mathbf{z}_l\boldsymbol{\beta}'}} \quad (3.18)$$

and the log partial likelihood

$$L(\beta) = \sum_{i:\delta_i=1} \left\{ \mathbf{z}_i\boldsymbol{\beta}' - \ln \left[\sum_{l:y_l \geq y_i} e^{\mathbf{z}_l\boldsymbol{\beta}'} \right] \right\}. \quad (3.19)$$

When there are tied event times in the sample, the maximisation of (3.19) becomes very time-consuming or not feasible. In such a case, two approximations have been proposed in the literature. The first approximation was proposed by Breslow (1974) and is a good approximation of the partial likelihood when the number of ties is not large. However, when the number of ties is large, the approximation proposed by Efron (1977) is preferred since it is more accurate than Breslow's approximation (Harrell 2001). More details of these approximations can be seen in Harrell (2001).

3.4 Discriminative ability measures

3.4.1 Definition

In general, the concordance probability between the observed response variable and the predicted outcome is a measure widely used to assess the predictive discriminative ability of a prediction regression model. Suppose one has a response variable R (in logistic regression the response variable is represented as Y and in survival analysis as T) and a set of predictor variables \mathbf{Z} . Then, in general a regression model can be written as a linear function of \mathbf{Z} for some function of R such that

$$m(R|\mathbf{Z}) = \beta_0 + \beta_1 Z_1 + \dots + \beta_p Z_p. \quad (3.20)$$

For example, if R is a binary variable and the logistic regression model is used, then m is the *logit* function of the expectation of R . If R is a time-to-event variable, then m is the logarithm of the hazard function. Given two independent copies

(\mathbf{Z}_1, R_1) and (\mathbf{Z}_2, R_2) of the random vector (\mathbf{Z}, R) , the concordance probability is defined as the probability that predictions and outcomes are concordant, that is,

$$\mathfrak{C} = P(m(R_2|\mathbf{Z}_2) > m(R_1|\mathbf{Z}_1)|R_2 > R_1). \quad (3.21)$$

If \mathfrak{C} takes a value of 0.5 then the model provides random predictions whereas a value of $\mathfrak{C} = 1$ represents a perfectly discriminative model.

For the specific logistic regression setting, the area under a receiver operating characteristic (ROC) curve (AUC) is a measure widely used to assess the discriminative ability of the model. The ROC curve is defined as

$$ROC(\cdot) = \{(FPR(c), TPR(c)), c \in (0, 1)\}, \quad (3.22)$$

where $FPR(c) = P(\pi(\mathbf{Z}) \geq c|Y = 0)$ is the *false positive rate* and $TPR(c) = P(\pi(\mathbf{Z}) \geq c|Y = 1)$ is the *true positive rate* for a given threshold c . Equivalently, the ROC curve can also be written as

$$ROC(\cdot) = \{(t, ROC(t)), t \in (0, 1)\}, \quad (3.23)$$

where the ROC function maps t to $TPR(c)$, and c is the threshold corresponding to $FPR(c) = t$. Thus, the ROC curve is a monotone increasing function mapping $(0, 1)$ onto $(0, 1)$ (Pepe 2003). An example of the ROC curve is given in Figure 3.3.

A widely used measure of the ROC curve is the AUC which is defined as

$$AUC = \int_0^1 ROC(t)dt. \quad (3.24)$$

A model which discriminates individuals perfectly into events and non-events will have an $AUC = 1$, whereas a non-discriminative model will have an $AUC = 0.5$. In this specific setting in which the outcome is a binary response variable, Bamber (1975) and Hanley and McNeil (1982) showed that the AUC is identical to the concordance probability. Thus, the AUC can be interpreted as the probability that the predictions from a randomly selected pair of event and non-event subjects are correctly ordered (Pepe 2003).

Given a sample $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, the estimation of the AUC is frequently calculated by the Mann-Whitney statistic (Pepe 2003). Let $\hat{\pi}_i = \widehat{\pi(\mathbf{z}_i)} = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \dots + \hat{\beta}_p z_{ip})$ be the estimated probability for subject i , then the estimated AUC is given as:

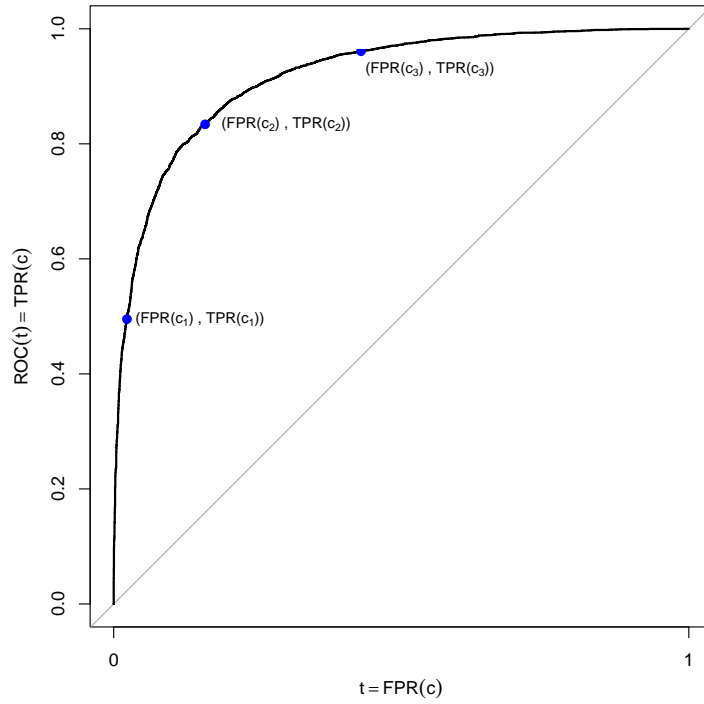


Figure 3.3: An example of a ROC curve. Values for (FPR,TPR) are shown for three threshold points, c_1, c_2, c_3 as an example.

$$\widehat{AUC} = \frac{1}{N_0 N_1} \sum_{l \in D_{Y=0}} \sum_{m \in D_{Y=1}} I[\widehat{\pi}_l, \widehat{\pi}_m], \quad (3.25)$$

where $D_{Y=1}$ and $D_{Y=0}$ are the sets of subjects with $Y = 1$ and $Y = 0$, respectively, N_1 and N_0 are the sizes of these sets and $I[\bullet]$ is the indicator function adjusted for ties

$$I[\widehat{\pi}_l, \widehat{\pi}_m] = \begin{cases} 1 & \text{if } \widehat{\pi}_l < \widehat{\pi}_m \\ 0.5 & \text{if } \widehat{\pi}_l = \widehat{\pi}_m \\ 0 & \text{if } \widehat{\pi}_l > \widehat{\pi}_m . \end{cases} \quad (3.26)$$

In a setting where the outcome is time to event, the concordance probability is defined as the probability that, of two randomly chosen individuals, the one with higher predicted probability of survival will outlive the one with lower predicted probability. In the presence of censorship, a problem arises with the comparison

of predicted survival times for a pair of individuals for which the event has not been observed. Harrell et al. (1982) proposed an estimator for the concordance probability called the c -index which is defined as “the proportion of all pairs of patients for which we could determine the ordering of survival times such that the predictions are concordant”. More specifically, Harrell et al. classified the pairs of individuals as *usable* or *unusable*. A pair of individuals is considered unusable if both had the event at the same time, or if one had the event and the other not, but was not followed long enough to determine whether would outlast the one with the event. Thus, the c -index estimator proposed by Harrell is the proportion of usable individual pairs in which the predictors and the outcomes are concordant and is computed by forming all pairs of observed data where the shorter follow-up time is an event.

Given a sample $\{\mathbf{z}_i, y_i, \delta_i\}_{i=1}^N$, consider all pairs $((\mathbf{z}_i, y_i, \delta_i), (\mathbf{z}_l, y_l, \delta_l))$ where the shorter follow-up time is an event. Then the c -index estimator proposed by Harrell is defined as

$$c = \frac{\sum_{i < l} \sum \{I(y_i < y_l)I(\mathbf{z}_i \hat{\boldsymbol{\beta}}' > \mathbf{z}_l \hat{\boldsymbol{\beta}}')I(\delta_i = 1) + I(y_l < y_i)I(\mathbf{z}_l \hat{\boldsymbol{\beta}}' > \mathbf{z}_i \hat{\boldsymbol{\beta}}')I(\delta_l = 1)\}}{\sum_{i < l} \sum \{I(y_i < y_l)I(\delta_i = 1) + I(y_l < y_i)I(\delta_l = 1)\}}, \quad (3.27)$$

where $\hat{\boldsymbol{\beta}}$ represents the partial likelihood estimate of $\boldsymbol{\beta}$.

Even though it is widely used in practice, as pointed out by Gönen and Heller (2005), the c -index estimator proposed by Harrell et al. (1982) is biased and the bias increases with the censoring rate. Hence, Gönen and Heller (2005) proposed an alternative estimator called the concordance probability estimator (CPE), which under the proportional hazards assumption is a consistent estimator of the concordance probability. This estimator is defined as

$$CPE = \frac{2}{N(N-1)} \sum_{i < l} \sum \left\{ \frac{I(\mathbf{z}_{li} \hat{\boldsymbol{\beta}}' < 0)}{1 + e^{\mathbf{z}_{li} \hat{\boldsymbol{\beta}}'}} + \frac{I(\mathbf{z}_{il} \hat{\boldsymbol{\beta}}' < 0)}{1 + e^{\mathbf{z}_{il} \hat{\boldsymbol{\beta}}'}} \right\}, \quad (3.28)$$

where \mathbf{z}_{il} represents the pairwise difference $\mathbf{z}_i - \mathbf{z}_l$.

Other estimators for the concordance probability have been proposed in the literature: Begg et al. (2000), Heagerty and Zheng (2005), Song and Zhou (2008) and Uno et al. (2011). In this dissertation we focused on the original Harrell c -index estimator and the CPE for two main reasons. Schmid and Potapov (2012) carried out

a comparison of different discrimination indexes and none of the estimators proved to be stable in all scenarios. Schmid and Potapov (2012) concluded that “censoring rates and model misspecification have non-negligible effects on the behaviour of estimators of discrimination indexes”. In addition, work has been done on the comparison of these two estimators in dichotomising a continuous predictor in a Cox PH model (Sima and Gönen 2013) and we intended to extend this research to the search for more than one cut point.

3.4.2 Comparison of the AUC

In prediction, often the interest lies in comparing the discriminative ability of two prediction models, with the aim of selecting the best one. Several methods have been proposed in the literature to compare correlated ROC curves, that is, ROC curves which are estimated on the same set of individuals. Of these, the methods proposed by Bandos et al. (2005; 2006), Braun and Alonzo (2008), DeLong et al. (1988), Hanley and McNeil (1983) are based on the comparison of the AUC whereas the methods proposed by Moise et al. (1988) and Venkatraman and Begg (1996) are based on the comparison of the ROC shape.

Additionally, other measures have been proposed to compare the discriminative ability of different prediction models or assess the incremental value of a new predictor variable (Steyerberg et al. 2010). Among others, these include reclassification tables (Cook 2007), reclassification test (Cook 2008), net reclassification improvement (NRI) and integrated discrimination improvement (IDI) (Pencina et al. 2008). Furthermore, decision analytic measures have also been proposed, which include decision curves (Vickers et al. 2008, Vickers and Elkin 2006) which plot the net benefit attained by decisions based on model predictions.

The most widely used method for comparing AUCs in practice is the one proposed by DeLong et al. (1988). Their test addresses a nonparametric comparison of areas under correlated ROC curves based on the U-statistics theory. This test is the one implemented in the SAS[®] statistical software and it is also implemented in several libraries of the R software (R Core Team 2014) such as the `pROC` package (Robin et al. 2011).

However, several authors have criticised the use of the DeLong’s test to evaluate the increment in discriminative ability owed to a new predictor variable by comparing the AUCs derived from the prediction model with and without the new predictor variable (Demler et al. 2012, Seshan et al. 2013, Vickers et al. 2011). Vickers et al.

(2011) demonstrated by means of simulations that the DeLong's test has a conservative test size and lower power than the Wald test when the aim is to assess the increase in discriminative ability of a new predictor variable. Additionally, Demler et al. (2012) concluded, based on numerical simulations, that the DeLong's test should not be used to compare two correlated AUCs of models which have been developed and validated in the same data and Seshan et al. (2013) showed that the use of the DeLong's test to compare two nested models is invalid. Nevertheless, they all agree that the DeLong's test has valid statistical properties when the aim is to compare two dependent prediction tests.

Thus, we considered the DeLong's test as a first approximation for the comparison of two categorical proposals in this dissertation. Moreover, we considered the IDI measure for validation purposes in the application to the IRYSS-COPD study in Chapter 5 because of the similarities with the specific objectives in this case.

3.4.3 Optimism correction of the model's discriminative ability

Over-fitting is a major problem in regression modelling, especially if the aim is to make predictions for new individuals. In our context, when fitting a regression model (either logistic regression or Cox PH model) our interest recalls on the estimation of the discriminative ability of such model. However, when the discriminative ability is estimated on the same data that were used to fit the model, the discrimination index obtained is overestimated.

Several approaches are used in practice to obtain unbiased estimates of the discriminative ability index, such as data splitting, cross-validation and bootstrap (Steyerberg 2009). Airola et al. (2011) carried out an experimental comparison of cross-validation techniques for estimating the AUC and proposed leave-pair-out cross-validation as the preferred method for bias-corrected estimation of the AUC. On the other hand, Harrell (2001) stated that the bootstrap provides the most efficient estimates for discrimination indexes. Harrell (2001) and Steyerberg (2009) proposed a bootstrap-based optimism correction of the discriminative ability of a regression model which was based on the original proposal made by Efron and Tibshirani (1993). Let \mathfrak{c} represent a discrimination index estimator of any of the models seen so far. Then the bootstrap-based optimism correction proposed by Harrell and Steyerberg can be summarised as follows.

Step 1. Fit the regression model on the basis of the sample $\{(z_i, y_i)\}_{i=1}^N$ and compute the corresponding discrimination index \mathfrak{c} . Let us denote this *apparent* \mathfrak{c}

as \mathbf{c}_{app} .

Step 2. For $b = 1, \dots, B$, generate the bootstrap resample $\{(z_{ib}^*, y_{ib}^*)\}_{i=1}^N$ by drawing a random sample of size N with replacement from the original sample. Fit the regression model to the bootstrap resample, obtain the estimated regression coefficients $\widehat{\beta}_0^b, \widehat{\beta}_1^b, \dots, \widehat{\beta}_p^b$, and compute the corresponding discrimination index, \mathbf{c}_{boot}^b for $b = 1, \dots, B$.

Step 3. Obtain the linear predictor for the original sample based on the fitted regression model obtained in Step 2, $\widehat{\beta}_0^b + \widehat{\beta}_1^b z_{i1} + \dots + \widehat{\beta}_p^b z_{ip}$, and compute the discrimination index and denote it by \mathbf{c}_o^b for $b = 1, \dots, B$.

Once the above process has been completed, the optimism O of the original discrimination index is calculated as follows

$$O = \frac{1}{B} \sum_{b=1}^B (\mathbf{c}_{boot}^b - \mathbf{c}_o^b) \quad (3.29)$$


and the bias-corrected discrimination index is then computed as $\mathbf{c}_{app} - O$.


We have used this bootstrap-based optimism correction as the basis of our proposal for the optimism correction presented in Chapter 5 and Chapter 6 below.

Chapter 4

Categorisation in logistic regression based on GAM

The work presented in this chapter has been previously published and partially presented at an international conference.

 Barrio, I., Arostegui, I., Quintana, J.M., and IRYSS-COPD Group. *Use of generalised additive models to categorise continuous variables in clinical prediction.* BMC Medical Research Methodology 2013; **13**:83.

 HTA in Integrated Care for a Patient Centered System. *Continuous variables categorization to apply in the development of predictive models for patients with COPD exacerbation.* Barrio, I., Arostegui, I., Quintana, J.M., Esteban, C., and IRYSS-COPD Group. *Contribution.* Bilbao June 2012.

In this chapter we present a graphical-based methodology to categorise continuous predictors to be used in clinical prediction models. The proposal is based on the use of GAMs with P-spline smoothers to determine the relationship between the continuous predictor and the outcome. This proposal is based on the work done by Hin et al. (1999), but it considers the need for more than two categories. The proposed method consists of creating at least one average-risk category along with high- and low-risk categories based on the GAM smooth function. In this chapter we present the development of this methodology along with a validation and an implementation to the IRYSS-COPD study. The rest of the chapter is organised as follows. In Section 4.1 we introduce the proposed methodology. Section 4.2 is devoted to the validation and the implementation of the proposed methodology. Finally, we point out some conclusions and limitations in Section 4.3.

4.1 Proposed methodology

Our proposal consists of categorising continuous variables by using the GAM with P-spline smoothers presented in Chapter 3, Section 3.2 above. Without loss of generality, let us assume that there is a continuous predictor variable X which we wish to categorise and a response variable Y with some exponential family distribution. In such a case, the GAM defined in equation (3.8) is fitted with X as the covariate and Y as the response variable:

$$g(E(Y|X)) = \beta_0 + f(X). \quad (4.1)$$

Specifically, if we consider a dichotomous outcome Y , the link to be used in the GAM regression model will be the *logit*. In such a case, the model described in equation (4.1) for one continuous covariate X would be more precisely specified by the following expression:

$$\text{logit}(\pi(X)) = \beta_0 + f(X). \quad (4.2)$$

Note that although we have represented by Z the covariates in Chapter 3, henceforth we will specifically represent by X the continuous predictor variable which we want to categorise.

The aim of this method is to categorise the covariate X in terms of the influence it has on the response variable Y . The number of categories as well as the location of the cut points will depend on the graphical relationship obtained by using the GAM with P-spline smoothers. On the basis of this model, the graphical display shows the relationship between X and $f(X)$, where X is plotted on the horizontal axis and the smooth function f is plotted on the vertical axis. $f(X)$ is the centred mean function where the centring coefficient is β_0 , and $f(X) = 0$ refers to the average value of the covariate. We therefore start by creating an average-risk category around this average-risk point, together with as many high- and low-risk categories as are required to capture the relationship between X and $f(X)$, as outlined in detail below.

We consider an average-risk category by building an interval around the point $x_0 \in X$ such that $f(x_0) = 0$. To do so, we calculate the value for x_0 , by computing the inverse of f , and then the estimated value $\hat{\pi}_0$, such that:

$$\hat{\pi}_0 = \text{logit}^{-1}(\beta_0 + f(x_0)) = \text{logit}^{-1}(\beta_0)$$

and its 95% confidence interval $(\hat{\pi}_{0_{inf}}, \hat{\pi}_{0_{sup}})$ given by

$$\hat{\pi}_{0_{inf}} = \hat{\pi}_0 - 1.96se(\hat{\pi}_0)$$

and

$$\hat{\pi}_{0_{sup}} = \hat{\pi}_0 + 1.96se(\hat{\pi}_0)$$

where $se(\hat{\pi}_0)$ is the estimated standard error of the expected response given by the GAM evaluated at point $\hat{\pi}_0$.

Finally, we obtain the interval $(x_{0_{inf}}, x_{0_{sup}})$ by reversing the process, such that

$$f^{-1}(\text{logit}(\hat{\pi}_{0_{inf}}) - \beta_0) = x_{0_{inf}}$$

and

$$f^{-1}(\text{logit}(\hat{\pi}_{0_{sup}}) - \alpha_0) = x_{0_{sup}}$$

That is, the points $x_{0_{inf}}$ and $x_{0_{sup}}$ are thus the cut points that determine the average-risk category.

If x_0 is not unique, i.e., if the graph displayed crosses the vertical axis more than once at point 0, then there will be more than one average-risk category, provided that the band at x_0 , is not too wide (with the band being the confidence interval shown in the graph). In other words, if x_{01} and x_{02} are two values for which the graph crosses the vertical axis at point 0, two average-risk categories will be considered, as long as $(x_{01_{inf}}, x_{01_{sup}})$ and $(x_{02_{inf}}, x_{02_{sup}})$ do not overlap. If the last happens, we hypothesise that it may result from two situations. The first is that one of the two intervals is based on a very small sample size which leads to a non-accurate and hence very wide interval. The second is the overlapping of two intervals of similar size. Under the first circumstance, we suggest to dismiss the interval based on a very small sample. However, if the second circumstance happens, we will consider the union of both intervals as the average-risk category.

Once the average-risk category has been defined, the following two possible scenarios are considered for creating high- and low-risk categories:

1. The relationship shown on the graph between the covariate and the outcome given by the GAM is linear along the entire range of X . Under this scenario, we propose to categorise X into a minimum of three categories, with the cut points for the three being the limits of the average-risk category. This hypothetical situation is depicted in Figure 4.1(a).

Moreover, if more categories are needed to ensure that the linear relationship between the covariate and the outcome is adequately retained, these could be created by considering appropriate cut points, preferably based on clinical criteria, in any of the designated high-risk or low-risk categories; or,

2. The relationship shown on the graph between the covariate and the outcome given by the GAM is not linear, which means that there is either a jump or a change in the slope. First, we propose to proceed as described above for the first three categories labelled “average-risk”, “low-risk” and “high-risk”. Second, the points at which the slope change occurs will be deemed to be extra cut points. Consequently, this will lead to the corresponding low-risk or high-risk categories, or both, being re-categorised as very low- and low- or very high- and high-risk categories respectively. The selection of these extra cut points will be made on the basis of graphical visualisation of the slope and the clinical significance of the cut point in question. This hypothetical situation of one extra cut point is depicted in Figure 4.1(b)

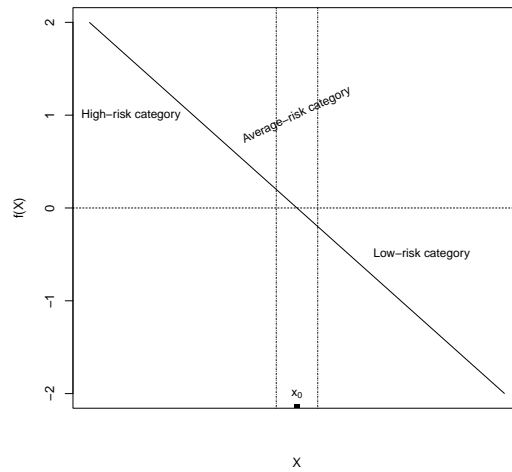
In both cases, the need for more than the proposed minimum number of categories is evaluated by comparing the results of adding more categories to the original continuous covariate, using the validation criterion explained in detail below.

4.2 Validation and implementation

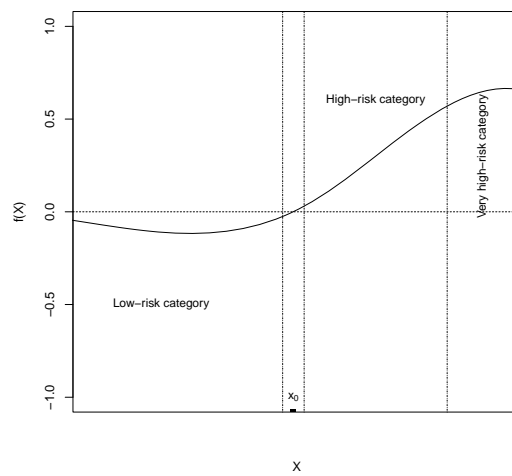
We considered the prospective cohort of patients with eCOPD presented in Chapter 2, Section 2.1 as the real data set for validation and implementation of the proposed methodology. First, we present that part of the IRYSS-COPD study selected for validation and implementation methods. Then the criteria for method validation are presented and, finally, the obtained results are shown.

4.2.1 Application to the IRYSS-COPD study

By way of an illustration of the application of the proposed methodology, we selected part of the IRYSS-COPD study: specifically, one dichotomous outcome; poor evolution in the first seven days from arrival at the ED; and two continuous predictor variables, namely, the blood gas parameter, PCO_2 , and the RR. Exacerbated COPD is a severe condition quite commonly seen at EDs, where proper decision-making tools are vitally important for performing the necessary diagnoses and implementing the treatments that are urgently required. Among the basic diagnostic tests used for



(a) Linear relationship



(b) Nonlinear relationship

Figure 4.1: Graphical representation of two hypothetical shapes between the predictor and the outcome with GAM. From left to right: a) linear relationship and b) nonlinear relationship.

classifying the severity of exacerbated COPD in such patients, arterial blood gases are the main tool. PCO_2 is a highly valuable item of information drawn from arterial blood gases (Quintana et al. 2014b); similarly, another key item of information is proper assessment of patients' respiratory rate (RR), something that is invariably affected in these cases. Furthermore, these two variables represent the two possible theoretical scenarios described above.

4.2.2 Validation

The total sample was randomly divided into a derivation (60%) and a validation (40%) sample. Cut points were obtained with the derivation sample and the validation sample was used for method evaluation purposes. As we mentioned in Chapter 1, split-sample validation is commonly used in the prediction modelling process. We chose to do it that way in order to mimic the development of prediction models used by clinical researchers.

The method for categorising continuous covariates was evaluated by comparing the performance of the proposed categorical predictor in the model with that of the original continuous variable modelled by a GAM as the best option in the same model. In addition, we also compared the proposed categorisation with the dichotomised variable suggested by Hin et al. (1999), which considers x_0 such that $f(x_0) = 0$ is the cut point for dichotomisation.

To compare models using different approaches to represent the same covariate, two criteria were selected: the first was the Akaike Information Criterion (AIC), a well-known, classical method for comparing two models (Akaike 1974); the second method of evaluation was based on the specific model defined in equation (4.2) and the study's designated purpose. In this particular case, we desired to evaluate the predictive ability of the model selected. We thus proposed to use the AUC as the parameter that quantifies a logistic model's discriminative ability as presented in Section 3.4. The AUCs for two ROC curves were compared with the DeLong test (DeLong et al. 1988).

Additionally, the goodness-of-fit of the proposed categorisation was evaluated by means of the Hosmer-Lemeshow test, which assesses the concordance between observed and expected event rates in a logistic regression model (Hosmer and Lemeshow 2000). Finally, the need for additional categories, in excess of a minimum of three, was also checked by testing for statistically significant differences in risk between additional and adjacent categories.

Finally, we performed a sensitivity analysis in order to assess the impact which sample size may have on the width of the average-risk category. We recalculated the average-risk category for PCO_2 and for samples of size 200, 400, 600, 800, 1000 and 1200 obtained resampling without replacement from the original sample.

All statistical analyses were performed with the (64 bit) R 3.0.1 software package (R Core Team 2014). The `mgcv` (Wood 2006), `BB` (Varadhan and Gilbert 2009) and `pROC` (Robin et al. 2011) libraries were specifically used to compute the GAM, obtain cut points and compare AUCs with the DeLong test respectively. The R code used to implement the proposed methodology was developed by the authors and is shown in Section 7.1.

4.2.3 Results

Categorisation process

RR: The relationship between the RR and poor evolution, as plotted by an additive logistic regression model with smoothing P-splines, is depicted in Figure 4.2. It can be seen that the relationship between the RR and poor evolution was linear and that there was only one value for which $f(x_0) = 0$ ($x_0 = 22$). Application of the proposed methodology to determine the limits of the average-risk category showed this category to be (20-24). It was therefore decided that the RR would be classified in three categories (Figure 4.2), with a high risk of poor evolution for values above 24 and a low risk of poor evolution for those below 20: accordingly, our final proposal for classifying the RR into three categories was ≤ 20 ; (20,24]; >24 . Internal limits were open at left and close at right by convenience.

In the search for an optimal fit to the original model, the need for a fourth category was explored. Taking the number of individuals with an RR above 24 and the available clinical information about the disease into consideration (Quintana et al. 2008), we selected an additional cut point of 30. The following four-category RR version, namely, ≤ 20 ; (20,24]; (24,30]; >30 , was thus also tested. The need for an extra cut point below 20 was not checked because of the small sample size in this category.

In addition, the RR variable was dichotomised as proposed by Hin et al. (1999), whereby an RR value for which there is an average risk of poor evolution is taken as the cut point, which in our case was 22: consequently, our dichotomous RR proposal was ≤ 22 ; >22 .

PCO_2 : the relationship between PCO_2 and poor evolution, as plotted by an

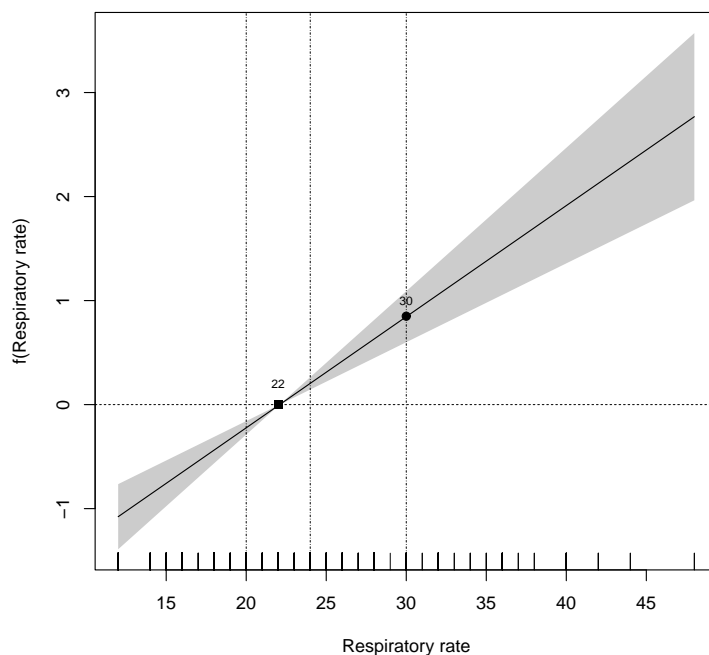


Figure 4.2: Graphical representation of the cut points obtained for the respiratory rate, based on the relationship between the continuous predictor respiratory rate and poor evolution.

additive logistic regression model with smoothing P-splines, is shown in Figure 4.3. In this case, the relationship did not prove linear, and showed a trend towards a less steep slope for higher values. We started by calculating an average-risk category: this was (43-52), meaning that there was a high risk of poor evolution for values above 52 and a low risk of poor evolution for those below 43. The need to select more cut points was then explored. From 40 to 43, the relationship was linear, and below this there was no significance because the confidence interval was too wide. Above 52, however, there were several points where there was a slope change. Although all values above 80 were dismissed, since the confidence interval was too wide and there were very few patients with PCO_2 values as high as this, graphical examination of values below 80 nevertheless showed 65 to be a reasonable cut point for distinguishing between the high- and very high-risk categories. Finally, we decided on the four-category PCO_2 proposal shown in Figure 4.3, namely, ≤ 43 ; (43-52]; (52-65]; >65 .

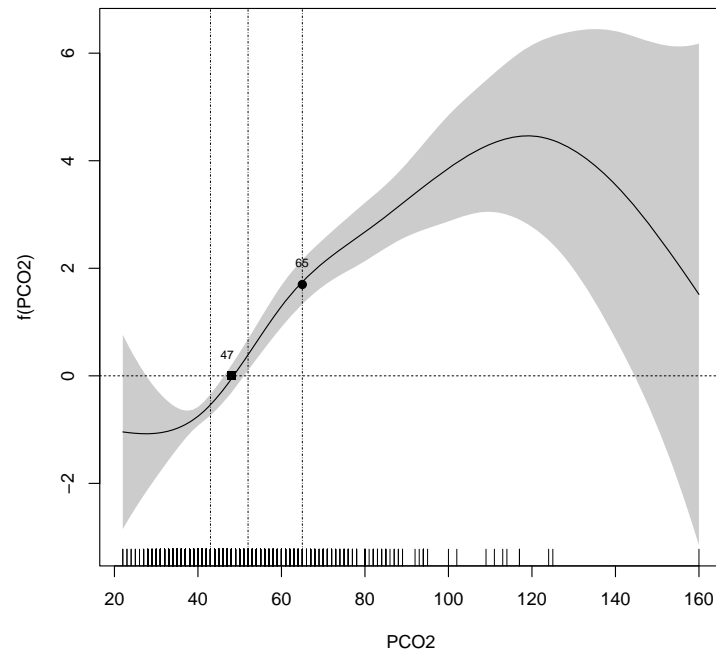


Figure 4.3: Graphical representation of the cut points obtained for the PCO_2 , based on the relationship between the continuous predictor PCO_2 and poor evolution.

As in the case of the RR above, the PCO_2 variable was also dichotomised, whereby a PCO_2 value for which there was an average risk of poor evolution was taken as the cut point. Since the value in this particular instance was 47, the dichotomous PCO_2 proposal was therefore ≤ 47 ; >47 .

Validation

We considered the assessment parameters AIC and AUC obtained from the continuous predictor in a GAM as the best option and those obtained with the dichotomised option as perhaps needing improvement. Detailed results of the validation process are shown in Table 4.1.

RR: Our approach proposed that the RR be classified into a minimum of three categories, for which the following values were obtained: AIC=314.5 and AUC=0.638 for the three-category option versus AIC=317.1 and AUC=0.634 for

Table 4.1: Categorisation of the respiratory rate (RR) and PCO₂ covariates from the IRYSS-COPD study based on the proposed methodology.

Variable	Derivation		Validation		
	Cut points		AIC	AUC	p-value*
RR	Continuous†		317.10	0.634	-
RR	Dichotomised	≤ 22	318.10	0.594	0.079
		> 22			
RR	3-category	≤ 20	314.50	0.638	0.8198
		(20 – 24] > 24			
RR	4-category	≤ 20	316.2	0.640	0.6833
		(20 – 24]			
		(24 – 30]			
		> 30			
PCO ₂	Continuous†		250.26	0.825	-
PCO ₂	Dichotomised	≤ 47	281.50	0.742	≤.0001
		> 47			
PCO ₂	3-category	≤ 43	270.76	0.779	0.0002
		(43 – 52] > 52			
PCO ₂	4-category	≤ 43	258.11	0.810	0.1148
		(43 – 52]			
		(52 – 65]			
		> 65			

* Corresponding to DeLong's test for comparing the AUC of each model with the continuous option

†AIC and AUC were calculated from the GAM

the continuous predictor, with no statistically significant differences being found between the two AUCs ($p = 0.8198$). The respective values for the dichotomous predictor were AIC=318.1 and AUC=0.594. Statistically significant differences in AUCs were observed between the dichotomous and proposed three-category approaches ($p = 0.049$).

Lastly, the four-category option yielded an AIC of 316.2 and an AUC of 0.64. No statistically significant differences in AUCs were observed when the four-category option was compared with both the continuous ($p = 0.6833$) and the three-category approaches ($p = 0.5968$). Moreover, when the model for the four-category option was fitted, however, non-statistical differences were found between the estimated parameters for the (24,30] and >30 categories ($p = 0.074$). Detailed results of the fitted model are shown in Table 4.2.

Additionally, the models for the four-category and three-category options were both well calibrated (Hosmer-Lemeshow test p -values > 0.05 in both cases).

Table 4.2: Results of the fitted logistic regression models with the four-category option for the respiratory rate (RR) and PCO_2 covariates from the IRYSS-COPD study, showing estimates of the β coefficients, their 95% confidence intervals and the p -values of their significance.

Category	Estimate	95%CI	p-value
RR ≤ 20	-1.76	(-2.37 , -1.16)	< 0.0001
RR (20 – 24]	-1.22	(-1.86 , -0.58)	0.0002
RR (24 – 30]	-0.56	(-1.18 , 0.06)	0.074
RR > 30	-	-	-
Hosmer-Lemeshow test p -value > 0.05			
$\text{PCO}_2 \leq 43$	-3.48	(-4.18 , -2.86)	< 0.0001
PCO_2 (43 – 52]	-2.62	(-3.27 , -2.03)	< 0.0001
PCO_2 (52 – 65]	-1.44	(-1.97 , -0.93)	< 0.0001
$\text{PCO}_2 > 65$	-	-	-
Hosmer-Lemeshow test p -value > 0.05			

PCO₂: Our approach proposed that the PCO_2 variable be classified into four categories, for which the following values were obtained: AIC=258.1 and AUC=0.81 for the four-category option versus AIC=250.26 and AUC=0.825 for the continuous predictor, with no statistically significant differences between the two AUCs ($p = 0.1148$). The respective values for the dichotomous predictor were AIC=281.5 and AUC=0.742. Statistically significant differences in AUCs were observed, not only between the continuous and dichotomous approaches ($p < 0.0001$), but also between the dichotomous and proposed four-category approaches ($p = 0.0001$). In addition, we verified the need to create a fourth category, by comparing the four-category against the three-category predictor (our minimum proposal). Statistically significant differences were found between both AUCs ($p = 0.0004$), with the AUC value for the three-category predictor being 0.779.

Furthermore, when the model for the four-category option was fitted, statistically significant differences were found between the estimated parameters for the (52-65] and > 65 categories ($p < 0.0001$). Lastly, the Hosmer-Lemeshow test assessed the goodness-of-fit of both the three- and four-category options ($p > 0.05$). Detailed results are shown in Table 4.2.

Additionally, we tested the performance of the two categorised variables in a

multivariate logistic regression model. Comparison between the multivariate model with RR and PCO_2 as continuous and that with these two variables classified in three and four categories respectively yielded no statistically significant differences in AUCs (AUC=0.827 for the former versus AUC=0.814 for the latter; $p = 0.7021$).

Finally, we performed a sensitivity analysis to assess the impact which the sample size has on the average-risk category width. We considered sub-samples of the original sample of sizes 200, 400, 600, 800, 1000 and 1200 and calculated the average-risk category for each of them. We realised that, although for sample sizes greater than 400 results remained stable, the average-risk category for a 200 sample size became much greater. In our opinion, when sample size is small (< 200), it is hard to detect the functional relationship between the predictor and the outcome accurately, and so there is a high variability in the selected cut points which results in a very wide interval for the average-risk category. Results from the sensitivity analysis for sample size variation for PCO_2 are shown in Figure 4.4.

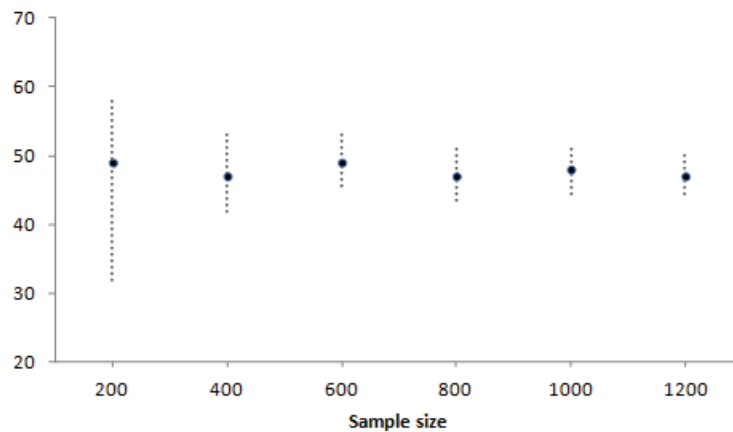


Figure 4.4: Graphical representation of the average-risk category width and location for PCO_2 based on sample size.

4.3 Conclusions and limitations

Our aim with the proposal presented in this chapter was to furnish a method of categorisation for clinical parameters selected as predictors, taking the following three facts into account: 1) categorisation would depend on the outcome of interest

and, by extension, on the model selected for analysis; 2) any loss of information would be minimal compared with the continuous predictor; and 3) the method would provide clinicians with a convenient and easily interpretable categorical predictor.

Studying the relationship between the predictor and the outcome was absolutely necessary in order to accommodate the first two facts. We decided to start by plotting the relationship graphically. GAM functions were selected because they are a powerful technique for estimating the relationship between continuous predictor variables and outcomes (Hastie and Tibshirani 1990), with no need for any assumptions about this relationship. P-spline smoothers are suggested in the literature as the most convenient technique for estimating smooth functions (Rice and Wu 2001). When developing the proposed methodology, we considered the method suggested by Hin et al. (1999) as the first approach to our designated objective: their proposal consists of dichotomising the predictor variable by using GAMs, taking the value for which an average risk is obtained as the cut point. We felt that a specific cut point could be highly dependent on the sample size or even the random sample itself, and that an interval, rather than a single point, might therefore be a wiser choice for an average-risk category. Although sample size would also affect the length of any interval, the latter would nevertheless provide more information to clinicians than would a single point. On the other hand, ensuring a minimum loss of information vis-à-vis the continuous variable was one of our stated goals, and so we hypothesised that two categories were possibly not enough.

This proposal, motivated in part by the work of Hin et al. (1999), occupies the middle ground between their approach and the original continuous predictor. We have seen that the categorisation proposal presented in this chapter does not lose critical information from the original predictor, respects the relationship between the original predictor and the outcome, and offers validated results with better predictive ability than the dichotomous approach. Moreover, our proposal starts by suggesting a minimum number of three categories, and offering the necessary cut points to ensure that such a categorisation is a good approximation of the continuous option. We have shown that, in general, this approach improves Hin et al.'s proposal (1999) in terms of fitting and prediction. The proposal includes a method for building an interval around the average-risk point using the inverse of the 95% confidence interval for the expected response. Although more complex techniques could provide other alternatives, in our opinion this is a simple and easily understood method that shows the advantages and usefulness of a three-category approach. In any given case, the need for more categories can be evaluated by researchers, de-

pending upon the relationship between the predictor and the outcome, sample size and clinical knowledge of the problem. Moreover, any improvement resulting from the addition of more categories can be statistically tested. Although this is an illustrative example, in the application presented here we selected four categories for PCO_2 and three categories for RR. We fitted the logistic regression models for the categorised variables and assessed the goodness of fit of those models by means of the Hosmer-Lemeshow test (Hosmer and Lemeshow 2000) because it is a test often used for binary outcomes (Steyerberg 2009). However, studying the effect the number of categories have on the goodness of fit of the model was out of the scope of this dissertation. A deeper analysis would be needed to study how categorisation affects calibration and in this context, other tests in addition to the Hosmer-Lemeshow test should be studied.

Nevertheless, we noticed this proposal had several limitations which should be taken into account. The first limitation of the proposed categorisation lies in the fact that it depends on the outcome and so its use cannot be recommended in every situation. This means that one might obtain different categorisation proposals for the same predictor, if one were to consider different outcomes or different modelling approaches. Although this characteristic of the proposal could be seen as a strength in the specific modelling situation, it must however be carefully reviewed when different modelling situations are being considered. We have previously mentioned that the width of the average-risk category will depend on sample size, which is an obvious limitation. Sensitivity analysis showed that for sample sizes above 200 results were quite stable, whereas for size 200 the interval was much wider. In our opinion, for a moderate sample size of 200 there were probably not enough data to catch the relationship between the predictor and the response variable, and so the average-risk category became very wide. In a simulation study with samples of size 200, we found that in 90% of them there were no differences between our proposal and the method suggested by Hin et al. (1999). Therefore, in this case we would recommend checking the performance of the dichotomised option first, merely for simplicity. Nevertheless, our proposal includes assessing the need for that third category in each case, and it compares the two versus the three categories approaches. The third limitation of our proposal resides in the subjectivity implied in the selection of extra cut points, in cases where more than three categories were necessary. We have given an outline of the way to do this in two different situations; and indeed, the addition of extra cut points in one of the specific applications was shown to improve the final result. However, we have also shown that improvement is


progressive, increasing as more cut points are added, basically because comparison is made with the continuous predictor. Cut points are selected on the basis of the predictor/outcome relationship given by the graphical display, which means that as more cut points are added, not only will the categorical and the continuous predictors be more similar, but the selected categorisation will also be more data-dependent. Apart from statistical significance, therefore, an important part of researchers' work will be to seek a balance between loss of information and practicality.


Considering all these limitations we thought of improving this categorisation proposal by means of a more mathematical methodology, in which, first, the number of cut points would not be limited to three or four. Second, the search for the cut points would not be exclusively based on a graphical display but on the optimisation of the discriminative ability of the prediction model, leading to an optimal number of cut points. And finally, but no less important, an approach for selecting optimal cut points for a continuous predictor in a multivariate setting was needed. Therefore, we address these requirements in the methodology proposed in Chapter 5 below.

Chapter 5

Categorisation methods in logistic regression based on the AUC

The work in this chapter has been previously partially presented at an international conference and is being reviewed in an international journal

 Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. A new approach to categorising continuous variables in prediction models: proposal and validation. *Statistical Methods in Medical Research (under review)*

 *International Workshop on Statistical Modelling. Location of optimal cut-points to categorize continuous variables in clinical studies. Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Contribution. Prague July 2012.*

In the previous chapter we proposed a method to categorise continuous predictor variables in a logistic regression model based on GAM. We showed that the proposal started by suggesting a minimum number of three categories, and determining if a fourth category was needed based on the graphical relationship and clinical knowledge. On the other hand, the approach presented in Chapter 4 did not allow for the categorisation of the continuous covariate in a multivariate setting.

With the aim of improving the limitations of the previous proposal, in this chapter we present a methodology to categorise continuous predictor variables in a logistic regression setting by maximising the AUC. The goal is to provide an optimal location for any given number of cut points as opposed to both the subjectivity of relying on a graphical display and being limited to two or three cut points. In addition, this methodology allows for the categorisation of a continuous predictor variable in a multivariate setting as suggested by Mazumdar et al. (2003). Finally,

as an alternative to the DeLong's test, we propose a new approach to select the optimal number of categories.

The rest of this chapter is organised as follows. In Section 5.1 we introduce the proposed methodology to categorise continuous variables by maximising the AUC. We also present a proposal to correct the bias of the AUC together with an approach for the selection of the optimal number of cut points. In Section 5.2 we present an empirical validation of the proposed methodology when the optimal cut points are known in theory or in practice. Section 5.3 is devoted to the application of the proposed methodology to the IRYSS-COPD study. Finally, we point out some conclusions and limitations in Section 5.4.

5.1 Proposed Methodology

Lets assume we have the simplest situation in which one has a dichotomous response variable Y , and a continuous predictor X that one wishes to categorise. Then, our proposal consists of categorising X such that the best logistic predictive model is obtained for Y . Specifically, given k the number of cut points set for categorising X in $k + 1$ intervals, let us denote $\mathbf{v}_k = (x_1, \dots, x_k)$ the vector of k cut points ordered from smaller to larger, and X_{cat_k} the corresponding categorised variable taking values from 0 to k . Then, what we propose is that the vector of k cut points $\mathbf{v}_k = (x_1, \dots, x_k)$, which maximises the AUC of the logistic regression model shown in equation (5.1) is thus the vector of the optimal k cut points.

$$\text{logit}(\pi(X_{cat_k})) = \beta_0 + \sum_{q=1}^k \beta_q 1_{\{X_{cat_k}=q\}}. \quad (5.1)$$

Suppose now that along with the predictor variable X we want to categorise, a set of other p predictors, $\mathbf{Z} = (Z_1, \dots, Z_p)$, are of interest. Then, what we propose is that the categorisation of X in a multivariate setting including the p predictors, will be that for which the AUC of the multivariate logistic regression model in equation (5.2) is maximised.

$$\text{logit}(\pi(\mathbf{Z}, X_{cat_k})) = \beta_0 + \sum_{r=1}^p \beta_r Z_r + \sum_{q=p+1}^{p+k} \beta_q 1_{\{X_{cat_k}=q-p\}}. \quad (5.2)$$

Given a random sample $\{(x_i, \mathbf{z}_i, y_i)\}_{i=1}^N$ drawn from (X, \mathbf{Z}, Y) the estimation of the parameters in equations (5.1) and (5.2) as well as of the associated AUCs can

be done as presented previously in Section 3.1. However, the problem now lies in looking for the vector of the cut points that maximise the AUC. To achieve this, we propose two alternative algorithms, respectively named *AddFor* and *Genetic*.

AddFor

Using this algorithm, one cut point is searched for at a time. In other words, this algorithm first seeks x_1 (in a grid of size M of equally spaced values in the range of X), such that the AUC of the logistic regression model shown in equation (5.2) for $k = 1$ will be maximised. Once x_1 has been selected, it is fixed and the algorithm proceeds to seek x_2 (in the grid of size M) ($x_2 \neq x_1$), so as to ensure that the AUC of the model in equation (5.2) for $k = 2$ will be maximised. The process is then repeated until the vector of k cut points, $\mathbf{v}_k = (x[1], \dots, x[k])$, has been obtained, with $x[o]$ denoting the o -th ordered cut point.

Genetic

Using genetic algorithms, the most widely known type of evolutionary algorithms (Eiben and Smith 2003), this method simultaneously finds the vector of k cut points, $\mathbf{v}_k = (x_1, \dots, x_k)$, which maximises the AUC of the logistic regression model in equation (5.2). Evolutionary algorithms are inspired by the concept of natural evolution. The underlying idea is that, given a population of individuals, environmental pressure leads to survival of the fittest, leading in turn to a rise in the overall fitness of the population. In a more mathematical context, given a function to be maximised (fitness function), a collection of heuristic rules are used to modify a population of possible solutions in such a way that each generation of potential solutions, tends to be, on average, better than its predecessor. The measure of whether one potential solution is better than another is the potential solution's fitness value. In our case, the AUC is the selected fitness function to be maximised, and the vector of optimal cut points would then be the best possible solution.

5.1.1 Optimism correction

As we presented in Section 3.4.3, the obtained AUC may be biased upward when the same data set is used to: a) fit the logistic regression model (involved in the cut point selection process); and b) compute the AUC (Copas and Corbett 2002). In our setting, the aim was to look for the vector of cut points \mathbf{v}_k that maximises the

AUC of the corresponding logistic model. Once the cut points have been selected, the AUC for the corresponding categorical variable may be biased. Hence, for a given set of cut points, we propose to correct the AUC for the logistic model of the corresponding categorical variable. The bias correction method is based on the bootstrap bias correction method initially proposed by Harrell (2001) and also recommended later in Steyerberg (2009) and described in Chapter 3, Section 3.4.3. Specifically, our proposal for the bootstrap bias correction method can be described as follows:

Step 1. Categorise the predictor variable on the basis of the original sample

$\{(x_i, z_i, y_i)\}_{i=1}^N$ (denote it as $x_{cat_k i}$) and compute the corresponding AUC as shown in equation (3.25). Let's denote this *apparent* AUC as \widehat{AUC}_{app} .

Step 2. For $b = 1, \dots, B$, generate the bootstrap resample $\{(x_{ib}^*, z_{ib}^*, y_{ib}^*)\}_{i=1}^N$ by drawing a random sample of size N with replacement from the original sample, and categorise the bootstrapped predictor $\{x_{ib}^*\}_{i=1}^N$ on the basis of the optimal cut points obtained in Step 1.

Step 3. Fit the logistic regression model to the bootstrap resample with the categorised version of the predictor. Let us denote $\widehat{\beta}^b$ as the vector of the estimated regression coefficients based on this bootstrap resample. Compute the corresponding AUC, \widehat{AUC}_{boot}^b for $b = 1, \dots, B$.

Step 4. Obtain the predicted probabilities for the original sample based on the fitted logistic regression model obtained in Step 3, i.e.,

$$\text{logit}^{-1}(\widehat{\beta}_0^b + \sum_{r=1}^p \widehat{\beta}_r^b z_{ri} + \sum_{q=p+1}^{p+k} \widehat{\beta}_q^b \mathbf{1}_{\{x_{cat_k i} = q-p\}}),$$

and compute the AUC. Let's denote this AUC as \widehat{AUC}_o^b for $b = 1, \dots, B$.

Once the above process has been completed, the optimism O of the original AUC is calculated as follows:

$$O = \frac{1}{B} \sum_{b=1}^B |\widehat{AUC}_{boot}^b - \widehat{AUC}_o^b|$$

and the bias-corrected AUC is then computed as $\widehat{AUC}_{app} - O$.

The difference between the original proposal for bootstrap bias correction proposed by Harrell (2001) and our slight modification lies on the calculation of the

optimism O . In our particular setting in which the AUC for a categorised predictor is estimated, we noted that \widehat{AUC}_{boot}^b was not always higher than \widehat{AUC}_o^b and hence in some circumstances the average of the difference tended to 0, so we propose to consider the absolute value of the difference.

Finally, we would like to point out that in order to mimic the study design, it is advisable that the resampling procedure described in Step 2 be done according to the design of the study. For instance, for a case-control study, data should be resampled separately within cases and controls. Moreover, if the data are clustered, the resampling units should be the clusters.

Up to now we have presented how to correct the bias of the estimated AUC for a given set of cut points to categorise the original predictor variable X . However, our proposal consists of searching the cut points in such a way that the corresponding AUC is maximised. Since this AUC may be biased and is still not corrected, we thought that perhaps it could have an impact on the selection of the optimal cut points. Hence, we propose to correct the AUC during the selection of the cut points, based on an iterative procedure of the bias correction method proposed above. Thus, the selection of the cut points would be done based on the maximisation of the bias-corrected AUC.

5.1.2 Selection of the optimal number of cut points

We are aware that in theory the optimal number of cut points for the categorisation of a continuous variable does not exist, since above all the possible number of cut points, the best option would be the continuous variable. However, in clinical practice categorical versions of the continuous variables are usually preferred without it always being clear which is the best number of categories to be used. So far, we have talked about how to estimate the optimal cut points for a given number k of cut points. In this section, we propose a naive approach to compare two given categorisations of the predictor variable X .

As we mentioned earlier in Chapter 3, Section 3.4, there are different approaches to comparing the AUC of two models, either when they are developed from the same data or not (DeLong et al. 1988, Venkatraman 2000, Venkatraman and Begg 1996). Nevertheless, the use of DeLong's test has been criticised when it is used to evaluate the incremental discriminative ability provided by the addition of a new predictor (Demler et al. 2012). In our approach, the aim is to compare the AUCs associated with two categorised variables, in order to decide the optimal number of cut points

needed in each situation. Our models are dependent since both categorical variables have been built from the same data set; however, the aim is to evaluate whether a significant increase in the discriminative ability is obtained by adding an extra cut point. Hence, as an alternative to what it is in the literature, we propose to compare the performance of two categorised versions of the continuous predictor X in terms of the bias-corrected AUC for the categorised variable.

This approach is based on the difference between the bias-corrected AUCs obtained for $k = l$ and $k = l + 1$ cut points. To determine the need for an extra optimal cut point, we propose to compute the confidence interval (CI) for this difference. Once the CI has been computed, an extra cut point is considered to be needed as long as the CI does not contain the zero. Specifically, in this work, bootstrap-based methods (Efron and Tibshirani 1993) are proposed for computing the percentiles and constructing the CIs. The procedure to compute the CI for the difference of the bias-corrected AUCs can be summarised as follows:

1. For $v = 1, \dots, V$, generate the bootstrap resample $\{(x_{iv}^*, z_{iv}^*, y_{iv}^*)\}_{i=1}^N$ by drawing a random sample of size N with replacement from the original sample.
2. Compute the bias-corrected AUC for the categorised variables for $k = l$ and $k = l + 1$ cut points and denote it as $\widehat{AUC}_{l,v}^*$ and $\widehat{AUC}_{l+1,v}^*$ respectively. The bias-corrected AUC is computed as explained above in Section 5.1.1, but for Step 1, using the optimal cut points obtained for $k = l$ and $k = l + 1$ on the basis of the original sample.
3. Compute the difference between the bias-corrected AUCs obtained for $k = l + 1$ and $k = l$

$$\widehat{AUC}_{Diff,v}^* = \widehat{AUC}_{l+1,v}^* - \widehat{AUC}_{l,v}^*.$$

Once the above process has been completed, the $(1 - \alpha)$ % limits for the CI for the difference are given by

$$\left(\widehat{AUC}_{Diff}^{\alpha/2}, \widehat{AUC}_{Diff}^{1-\alpha/2} \right)$$

where \widehat{AUC}_{Diff}^p represents the p -percentile of the estimated $\widehat{AUC}_{Diff,v}^*$ ($v = 1, \dots, V$).

Additionally, we considered a second criterion to evaluate the need for an extra optimal cut point. This was the integrated discrimination improvement (IDI) index, proposed by Pencina et al. (2008). IDI is a useful measure to compare and assess the improvement in terms of risk prediction of two predictive models. Accordingly, in our particular setting, the IDI can be a useful measure to evaluate the improvement

offered by adding an extra cut point. In particular, we propose the criterion that an extra cut point is needed as long as a statistically significant IDI is obtained when comparing the fitted logistic regression models obtained with $k = l$ and $k = l + 1$ cut points.

5.2 Empirical validation

In this section we present three simulation studies that we conducted to analyse the empirical performance of the methods described in Section 5.1 above, and report the results obtained.

1. The first simulation study is performed under known theoretical conditions that verify linear effects in the logistic regression model. We used this setting with three different purposes: a) study the need for the bias correction of the AUC and compare the correction during the selection of the cut points (first level) and at the end of the process that is, once the optimal cut points have been selected (second level); b) validate the estimated cut points and the performance of the algorithms *AddFor* and *Genetic*; and c) study the convergence of the bias-corrected AUC of the categorised variable to the AUC of the continuous predictor.
2. We conducted a second simulation study considering nonlinear effects. We compared the performance of the proposed algorithms in the estimation of the optimal cut points as well as in the bias-corrected AUCs.
3. In the third simulation study, we conducted a backward validation in which the cut points for a continuous variable were previously established. The aim was to validate the proposed method also when the cut points were scientifically pre-established based on clinical knowledge.

For all simulation studies, we begin by presenting the scenarios and set up and end up summarising the results obtained. All computations were performed in (64 bit) R 3.0.1 and a workstation equipped with 24GB of RAM, an Intel Xeon E5620 processor (2.40 Ghz), and Windows 7 operating system. Specifically, the `genoud` function of the `rgenoud` (Mebane and Sekhon 2011) package was used to compute the genetic algorithms, and the `glm` function of the `stats` package was used for the estimation of the logistic model.

5.2.1 Validation under known theoretical conditions

Scenarios and Set-Up

In the first setting, the predictor variable X was simulated from a normal distribution separately in each of the populations defined by the outcome ($Y = 0$ and $Y = 1$), i.e., $X|(Y = 0) \simeq N(\mu_0, \sigma_0)$ and $X|(Y = 1) \simeq N(\mu_1, \sigma_1)$. It should be noted that, when σ_0 and σ_1 are equal, X is linearly related to the log odds of the response (see Appendix A).

Moreover, for a fixed number of cut points, their theoretical location is known (Tsuruta and Bax 2006), as well as the AUC associated with the corresponding categorical covariate. Accordingly, the aims of this simulation study were fourfold:

- a) To compare the obtained cut points when the optimism correction of the AUC is performed during or after the selection process.
- b) To compare the obtained bias-corrected AUC and the theoretical one.
- c) To compare the estimated optimal cut points obtained with the proposed methodology and the theoretical optimal cut points.
- d) To study the convergence of the bias-corrected AUC of the categorised variable to the AUC of the continuous predictor.

Specifically, we considered $X|(Y = 0) \simeq N(0, 1)$ and $X|(Y = 1) \simeq N(1.5, 1)$. The simulations were done assuming the same number of individuals in $Y = 0$ and $Y = 1$ and total sample sizes of $N = 500$ and $N = 1000$. As far as the number of cut points is concerned, $k = 1, 2$ and 3 were considered. Finally, for the *AddFor* algorithm grid sizes of $M = 100$ and $M = 1000$ were used. In all cases, $R = 500$ replicates of simulated data were performed and $B = 50$ was considered for the AUC bias correction procedure.

Results

AUC bias correction at two levels: selection of cut points

Figure 5.1 shows the boxplot of the differences between the cut points obtained when the AUC overestimation was corrected during or after the selection of the cut points, using the *AddFor* algorithm for $k = 1$ cut points, a sample size of $N = 500$ and $R = 500$ replicates. As can be observed, the results suggest that correction during or after the cut point selection procedure has no impact on the obtained cut

points. For instance, when a grid of size $M = 100$ was used, the mean and median of the difference between estimated cut points were -0.021 and 0.000 , respectively. When a grid of size $M = 1000$ was used, -0.027 and -0.003 values were obtained for mean and median of the difference. For these simulations, only the *AddFor* algorithm was used for computational reasons.

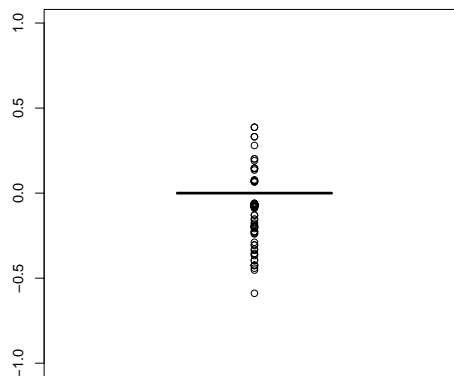
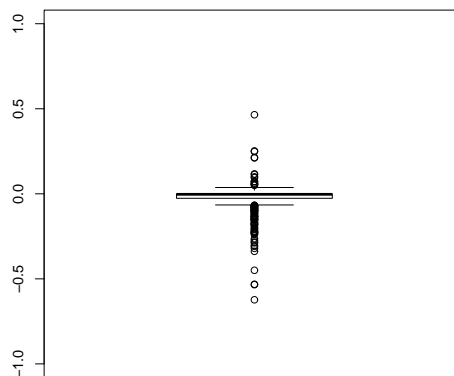
(a) *AddFor* $M = 100$ (b) *AddFor* $M = 1000$

Figure 5.1: Boxplot of the difference of the estimated optimal cut points when the AUC overestimation was corrected at first and second levels. Results are based on 500 simulated data and a sample size of $N = 500$, according to the simulation study under known theoretical conditions. From left to right: (a) *AddFor* algorithm with a grid of $M = 100$ and $k = 1$ number of cut points; (b) *AddFor* algorithm with a grid of $M = 1000$ and $k = 1$ number of cut points.

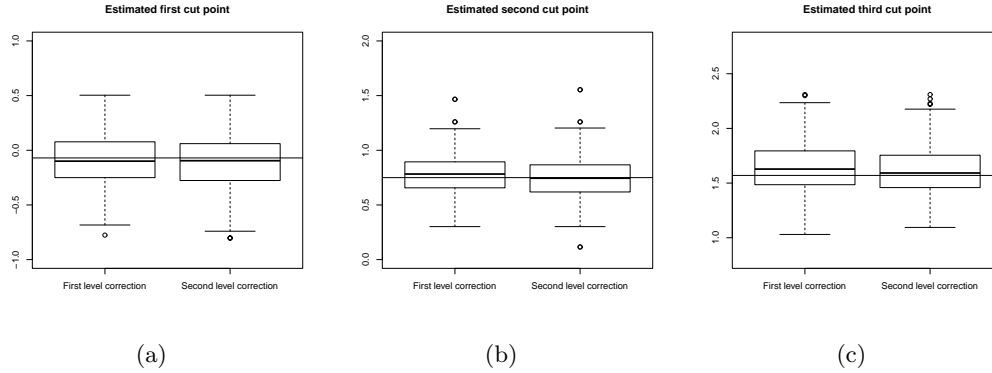


Figure 5.2: Boxplot of the estimated optimal cut points when the AUC overestimation was corrected at first or second level. Results are based on 500 simulated data, $N = 500$ and $k = 3$ number of cut points, according to the simulation study under known theoretical conditions, where the theoretical optimal cut points are $-0.07, 0.75$ and 1.57 , respectively, when $k = 3$. From left to right: (a) estimated first cut point; (b) estimated second cut point and (c) estimated third cut point with the *AddFor* algorithm with a grid of size $M = 100$.

Figure 5.2 shows the boxplot of the estimated optimal cut points when the AUC overestimation was corrected at first or second level, using the *AddFor* algorithm with a grid of size $M = 100$ for $k = 3$ cut points, a sample size of $N = 500$ and $R = 500$ replicates. Similarly to the case of a single cut point, the results suggest that correction during or after the cut point selection procedure has no impact on the obtained cut points.

In Table 5.1, numerical results of the estimated cut points at first or second level are shown. Mean, standard deviation, median, bias and the mean squared error (MSE) are reported for the estimated cut points. Note that for $k = 1$, the bias obtained is slightly smaller when the cut points have been estimated based on the maximisation of the corrected AUC, while for $k = 3$ the bias for the second and third cut point is slightly smaller when the AUC has been corrected at the end of the process. Nevertheless, the MSE of the estimated optimal cut points at first or second level is very similar.

AUC bias correction at two levels: estimation of the AUC

Additionally, we observed that we provided a bias-corrected AUC when the correction was performed at either the first or second level. Detailed results are shown in Table 5.2 where the mean, standard deviation, bias and MSE are reported for the AUC, and bias-corrected AUC at the first and second levels.

Table 5.1: Estimated optimal cut points obtained when the AUC overestimation was corrected at first or second level according to the simulation study under known theoretical conditions. Results are based on 500 simulated data and $N = 500$. The *AddFor* algorithm was used with a grid of size $M = 100$ and 1000 for $k = 1$ and a grid of size $M = 100$ for $k = 3$ number of cut points. Mean (sd), median, bias and MSE of the estimated cut points are reported.

No. of cut points	Method	Theoretical cut points	Mean (sd)	Median	Bias	MSE
Estimation at a first level						
$k = 1$	Addfor 100	0.773	0.764 (0.187)	0.771	-0.009	0.035
	Addfor 1000		0.748 (0.178)	0.739	-0.025	0.032
$k = 3$	Addfor 100	-0.068	-0.104 (0.238)	-0.099	-0.036	
		0.750	0.779 (0.182)	0.783	0.029	0.050
		1.568	1.641 (0.182)	1.628	0.073	
Estimation at a second level						
$k = 1$	Addfor 100	0.773	0.742 (0.188)	0.743	-0.031	0.036
	Addfor 1000		0.722 (0.173)	0.706	-0.051	0.033
$k = 3$	Addfor 100	-0.068	-0.111 (0.237)	-0.095	-0.043	
		0.750	0.750 (0.193)	0.745	0.000	0.049
		1.568	1.610 (0.224)	1.592	0.042	

Simulation results suggest that correcting the AUC bias at a first or second level had no impact in either the selection of cut points or the estimation of the bias-corrected AUC. However, correcting the AUC at a first level, i.e., during the search of the cut point, was computationally much more expensive and would not be feasible in practice. All these results suggest correcting the AUC bias at the end of the process. Accordingly, it is the approach followed in all simulations and real data analysis that we present henceforth in this chapter.

Selection of cut points

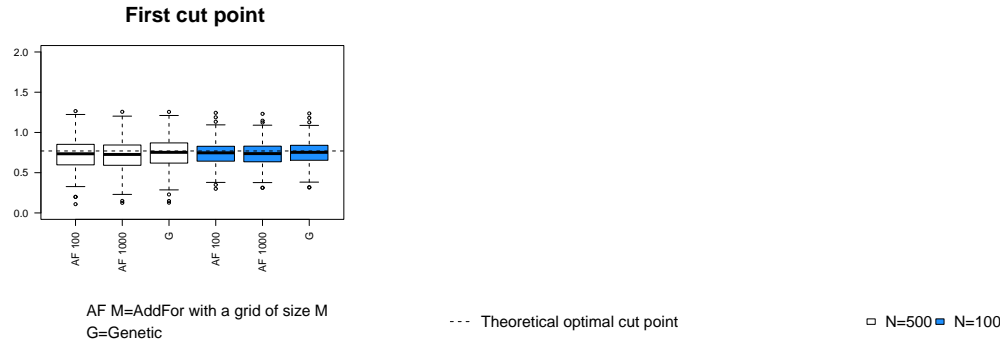
Figure 5.3 depicts the boxplot of the estimated optimal cut points over 500 simulated data sets, for each of the proposed algorithms, different sample sizes and number of cut points. As can be observed, the cut points obtained by the *Genetic* or *AddFor* algorithms were close to the theoretical optimal cut points, with both algorithms presenting a low bias. The corresponding detailed numerical results are shown in Table 5.3. Under this scenario, the theoretical optimal cut points are $\mathbf{v}_1 = (0.77)$, $\mathbf{v}_2 = (0.23, 1.27)$ and $\mathbf{v}_3 = (-0.07, 0.75, 1.57)$ for $k = 1, 2, 3$ number of cut points, respectively. Note that the average of the estimated cut

Table 5.2: Estimated AUC and bias-corrected AUC at first and second level according to the simulation study under known theoretical conditions. Results are based on 500 simulated data and $N = 500$. The *AddFor* algorithm was used with a grid of size $M = 100$ and $M = 1000$ for $k = 1$ and a grid of size $M = 100$ for $k = 3$ number of cut points. Mean (sd), median, bias and MSE of the estimated AUCs are reported.

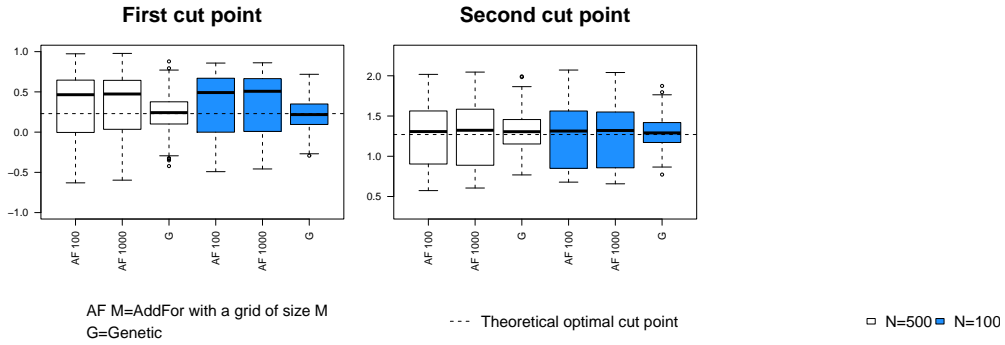
No. of cut points	Method	Theoretical AUC	Mean (sd)	Median	Bias	MSE
Estimated AUC						
$k = 1$	Addfor 100	0.750	0.781 (0.017)	0.782	0.031	0.0012
	Addfor 1000		0.784 (0.018)	0.784	0.034	0.0014
$k = 3$	Addfor 100	0.835	0.844 (0.016)	0.844	0.009	0.0003
First level bias-corrected AUC						
$k = 1$	Addfor 100	0.750	0.766 (0.017)	0.767	0.016	0.0006
	Addfor 1000		0.770 (0.018)	0.770	0.020	0.0007
$k = 3$	Addfor 100	0.835	0.831 (0.016)	0.831	-0.004	0.0003
Second level bias-corrected AUC						
$k = 1$	Addfor 100	0.750	0.766 (0.017)	0.767	0.016	0.0005
	Addfor 1000		0.768 (0.017)	0.769	0.018	0.0006
$k = 3$	Addfor 100	0.835	0.832 (0.016)	0.831	-0.004	0.0003

points across simulated data sets was very similar for both algorithms with these values being very close to the theoretical optimal cut points. As expected, the differences with respect to the theoretical optimal cut points were smaller when the sample size increased from 500 to 1000. For example, for $k = 3$ cut points, the average of the cut points obtained with the *Genetic* algorithm across all replicates were $\bar{\mathbf{v}}_3 = (-0.11, 0.76, 1.63)$ and $\bar{\mathbf{v}}_3 = (-0.09, 0.75, 1.61)$ for sample sizes of 500 and 1000, respectively, while with the *Addfor* algorithm and a grid of size 1000 they were $\bar{\mathbf{v}}_3 = (-0.11, 0.73, 1.60)$ and $\bar{\mathbf{v}}_3 = (-0.08, 0.74, 1.58)$. It should be noted that, when the number of cut points desired was 2, the *AddFor* did not perform as well as the *Genetic* algorithm. While the former closely located only one of the two optimal cut points, the latter managed to approximate both cut points. For instance, for a sample size of 500 and $k = 2$, the bias obtained for the estimated cut points were (0.004, 0.040) using the *Genetic* algorithm, and (0.117, -0.010) using the *AddFor* algorithm with a grid of size 1000.

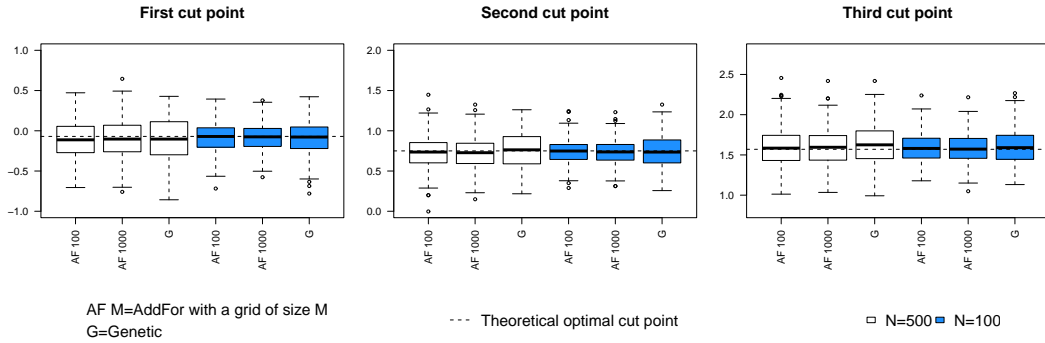
In Table 5.4, the average, bias and standard deviation of the bias-corrected AUC values over 500 simulated data sets are shown for each of the proposed algorithms, different sample sizes and number of cut points. Note that the AUC values obtained were almost unbiased, being the absolute value of the bias obtained for $k = 1$ and



(a) $k = 1$



(b) $k = 2$



(c) $k = 3$

Figure 5.3: Boxplot of the estimated optimal cut points based on 500 simulated data obtained according to the simulation study under known theoretical conditions and comparison with the theoretical optimal cut point. From top to bottom: (a) for $k = 1$ number of cut points; (b) for $k = 2$ number of cut points; and (c) for $k = 3$ number of cut points.

$k = 2$ less or equal to 0.02, and less or equal to 0.004 for $k = 3$. Additionally, the *Genetic* approach generally provided slightly higher AUC values than the *AddFor* algorithm. However, when the *AddFor* grid size was increased from 100 to 1000, the obtained results were almost the same as those obtained with the *Genetic* algorithm. For instance, for a sample size of 500 and $k = 3$ number of cut points, the average of bias-corrected AUCs were 0.831, 0.834 and 0.835 for the *AddFor* with grid sizes of 100 and 1000 and the *Genetic* algorithm, respectively, being the theoretical AUC 0.835.

Sample sizes of $N = 500$ and $N = 1000$ were selected to ensure a requirement commonly used for the specific framework of prediction models (Steyerberg 2009). Nevertheless, the performance of the proposed methodology was also verified for smaller sample sizes such as $N = 50, 100, 200$ and 300. Figures 5.4(a) to 5.4(c) depict the boxplots of the obtained optimal cut points over $R = 500$ simulated data sets, for each of the proposed algorithms (for the *AddFor* $M = 100$ and $M = 1000$ were considered), different sample sizes, and $k = 3$ number of cut points. As can be observed, the proposed algorithms performed satisfactorily even when small sample sizes were considered. As would be expected, the largest bias and variance were obtained with a sample size of 50. However, it should be pointed out that in the specific framework of the development of prediction models, small sample sizes are not used.

Convergence to the theoretical AUC

Finally, we studied the convergence to the theoretical AUC yield by the continuous predictor when the number of preselected cut points k is increased. Additionally, we studied the computing times obtained with each algorithm and for each number of cut points k selected. For this purpose, simulations were conducted for $k = 1$ to $k = 9$ cut points and a sample size of $N = 500$.

As shown in Figures 5.5(a) to 5.5(c) and Table 5.5, the convergence to the theoretical AUC was obtained with all the algorithms. However, as expected, the computing time increased considerably as the number of cut points to be selected increased, especially when the *Genetic* algorithm was used. For instance, the average time required by the *Genetic* method to compute three cut points for a sample size of 500 was 84.05 seconds versus 25.27 seconds required by the *AddFor* for a grid of size 1000.

Table 5.3: Validation under known theoretical conditions simulation study: average, standard deviation, median, bias and MSE of the estimated optimal cut points over 500 simulated data sets.

No. of cut points	Theoretical cut points	Method	Cut point Estimation			
			Mean (sd)	Median	Bias	MSE
Sample Size N = 500						
$k = 1$	0.773	Addfor 100	0.730 (0.183)	0.735	-0.043	0.035
		Addfor 1000	0.722 (0.183)	0.727	-0.051	0.036
		Genetic	0.747 (0.184)	0.753	-0.026	0.035
$k = 2$	0.227 1.274	Addfor 100	0.332 (0.373)	0.465	0.105	0.140
			1.255 (0.361)	1.307	-0.019	
		Addfor 1000	0.344 (0.358)	0.474	0.117	0.139
			1.264 (0.369)	1.323	-0.010	
		Genetic	0.231 (0.220)	0.242	0.004	0.050
			1.314 (0.222)	1.305	0.040	
$k = 3$	-0.068 0.750 1.568	Addfor 100	-0.121 (0.234)	-0.112	-0.053	0.048
			0.733 (0.187)	0.736	-0.017	
			1.599 (0.227)	1.584	0.031	
		Addfor 1000	-0.112 (0.235)	-0.103	-0.044	0.046
			0.726 (0.184)	0.728	-0.024	
			1.595 (0.215)	1.595	0.027	
		Genetic	-0.113 (0.266)	-0.103	-0.045	0.062
			0.758 (0.225)	0.763	0.008	
			1.631 (0.246)	1.625	0.063	
Sample Size N = 1000						
$k = 1$	0.773	Addfor 100	0.741 (0.146)	0.747	-0.032	0.022
		Addfor 1000	0.735 (0.144)	0.737	-0.038	0.022
		Genetic	0.750 (0.144)	0.753	-0.023	0.021
$k = 2$	0.227 1.274	Addfor 100	0.336 (0.361)	0.492	0.109	0.139
			1.227 (0.365)	1.314	-0.047	
		Addfor 1000	0.343 (0.356)	0.507	0.116	0.135
			1.227 (0.358)	1.320	-0.047	
		Genetic	0.221 (0.181)	0.218	-0.006	0.033
			1.297 (0.179)	1.290	0.023	
$k = 3$	-0.068 0.750 1.568	Addfor 100	-0.081 (0.170)	-0.071	-0.013	0.027
			0.742 (0.146)	0.749	-0.008	
			1.589 (0.177)	1.581	0.021	
		Addfor 1000	-0.084 (0.162)	-0.075	-0.016	0.026
			0.735 (0.144)	0.737	-0.015	
			1.582 (0.177)	1.572	0.014	
		Genetic	-0.090 (0.198)	-0.078	-0.022	0.041
			0.745 (0.195)	0.735	-0.005	
			1.609 (0.209)	1.590	0.041	

Table 5.4: Validation under known theoretical conditions simulation study: average, bias and standard deviation of the bias-corrected AUC values over 500 simulated data sets obtained, together with the theoretical AUC associated with the corresponding categorical covariate and the continuous predictor.

No. of cut points	Method	Theoretical AUC	Bias-corrected AUC			
			Mean (sd)	Median	Bias	MSE
Sample Size N = 500						
$k = 1$	Addfor 100	0.750	0.766 (0.017)	0.767	0.016	0.0005
	Addfor 1000		0.768 (0.017)	0.769	0.018	0.0006
	Genetic		0.769 (0.017)	0.770	0.019	0.0006
$k = 2$	Addfor 100	0.820	0.807 (0.017)	0.807	-0.013	0.0005
	Addfor 1000		0.810 (0.017)	0.810	-0.010	0.0004
	Genetic		0.818 (0.016)	0.820	-0.002	0.0004
$k = 3$	Addfor 100	0.835	0.831 (0.016)	0.832	-0.004	0.0003
	Addfor 1000		0.834 (0.016)	0.836	-0.001	0.0004
	Genetic		0.835 (0.016)	0.837	0.000	0.0004
Sample Size N = 1000						
$k = 1$	Addfor 100	0.750	0.768 (0.013)	0.768	0.018	0.0005
	Addfor 1000		0.770 (0.013)	0.770	0.020	0.0006
	Genetic		0.771 (0.013)	0.770	0.021	0.0006
$k = 2$	Addfor 100	0.820	0.807 (0.013)	0.807	-0.013	0.0003
	Addfor 1000		0.809 (0.013)	0.810	-0.011	0.0003
	Genetic		0.819 (0.012)	0.820	-0.001	0.0004
$k = 3$	Addfor 100	0.835	0.832 (0.012)	0.833	-0.003	0.0001
	Addfor 1000		0.835 (0.012)	0.835	0.000	0.0004
	Genetic		0.836 (0.012)	0.836	0.001	0.0004
Continuous predictor's theoretical AUC 0.855						

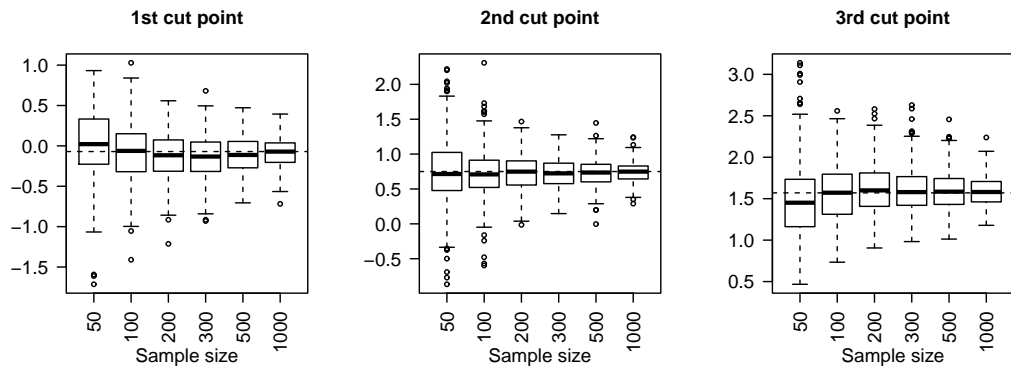
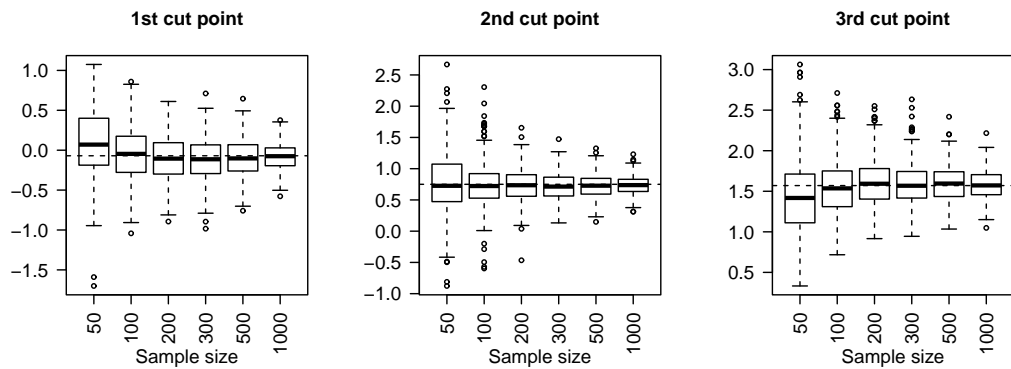
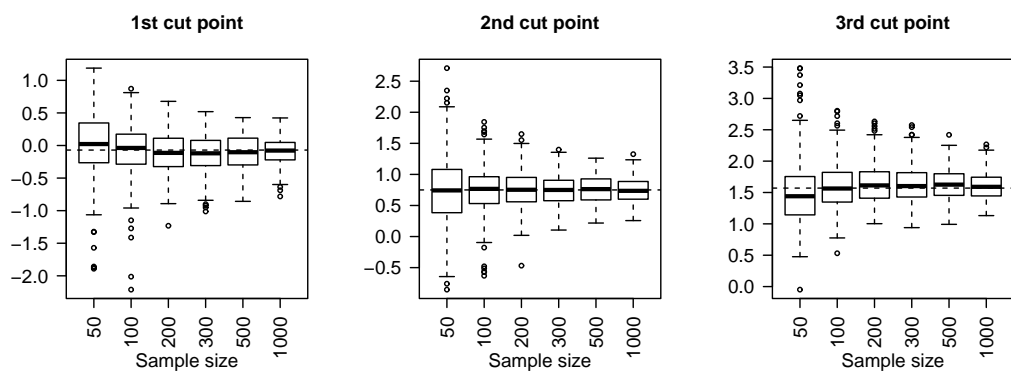
(a) *AddFor* $M = 100$ (b) *AddFor* $M = 1000$ (c) *Genetic*

Figure 5.4: Boxplot of the estimated optimal cut points based on 500 simulated data sets obtained according to the theoretical validation study. The figure shows the results for different sample sizes ($N = 50, 100, 200, 300, 500$ and 1000) and $k = 3$ number of cut points. From top to bottom: (a) *AddFor* $M = 100$; (b) *AddFor* $M = 1000$; and (c) *Genetic*. The theoretical cut point is represented by a dashed line.

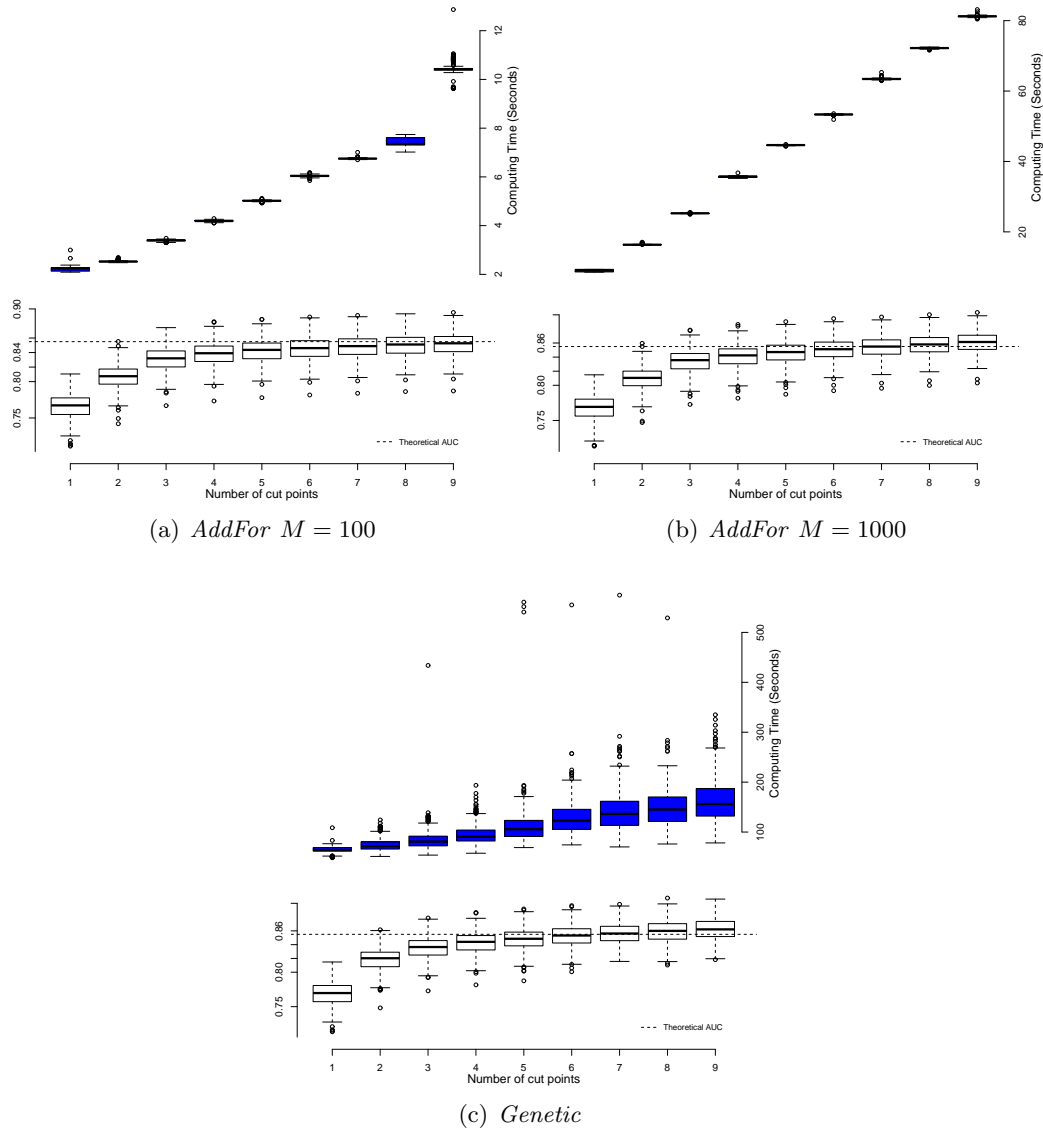


Figure 5.5: Convergence of the bias-corrected AUC values to the theoretical AUC. The boxplots plotted at the bottom of each graph correspond to the left axis and denote the AUC values obtained for each number of cut points k selected for a sample size of $N = 500$. The boxplots plotted at the top of each graph correspond to the right axis and denote the computing times (in seconds) obtained for each number of cut points k selected for a sample size of $N = 500$. (a) *AddFor* $M = 100$; (b) *AddFor* $M = 1000$; and (c) *Genetic*.

Table 5.5: Convergence of the bias-corrected AUC values to the theoretical AUC. Averaged bias-corrected AUC values for the *Addfor* ($M = 100$ and 1000) and *Genetic* algorithms together with the theoretical AUC for each categorical variable are reported.

k	Theoretical AUC	<i>AddFor</i> $M = 100$	<i>AddFor</i> $M = 1000$	<i>Genetic</i>
1	0.750	0.766	0.768	0.769
2	0.820	0.807	0.810	0.818
3	0.835	0.831	0.834	0.835
4	0.843	0.838	0.841	0.843
5	0.847	0.842	0.846	0.848
6	0.847	0.845	0.851	0.852
7	0.849	0.848	0.854	0.856
8	0.850	0.850	0.858	0.860
9	0.851	0.852	0.861	0.862

Continuous predictor's theoretical AUC 0.855

5.2.2 Comparison under nonlinear effects

Scenarios and Set-Up

This simulation study aimed to compare the performance of the two presented algorithms when the relationship between the covariate X , and the logit transformation of the response variable Y was nonlinear. On one hand, estimated cut points, and on the other, estimated bias-corrected AUCs were compared when algorithms *AddFor* or *Genetic* were used.

To this end, we defined the covariate X , according to a $U(0,1)$, and a binary outcome variable Y , where $Y \sim \text{Bernoulli}(\pi(X))$ and

$$\text{logit}(\pi(X)) = X - 10(X - 0.2)^3 + 110(X - 0.6)^3,$$

for which the empirical theoretical AUC value resulted in 0.824.

Simulations were conducted for a different number of cut points to be selected ($k=1, 2$ and 3), different grid sizes in which the cut points in the *AddFor* algorithm were to be sought ($M=100$ and 1000), and different sample sizes ($N=500$ and 1000), with $R = 500$ replicates being generated for each sample size. For each data set so generated and both algorithms, the bias-corrected AUC value obtained by the proposed categorisation was estimated. To evaluate the discriminative ability of the proposed categorised variable, the bias-corrected AUCs were compared to the

theoretical AUC value for the continuous variable, which was empirically calculated on the basis of the defined probabilities.

Results

Figure 5.6 depicts the boxplot of the estimated optimal cut points over 500 simulated data sets, for each of the proposed algorithms, different sample sizes and number of cut points. Numerical results are shown in Table 5.6. Simulation results suggest that both algorithms perform similarly for any number of cut points. Nevertheless, note that when three cut points are sought, the standard deviation for the second cut point is larger than for the first and third cut points, and that this is slightly lower when the *Genetic* algorithm is used. This cut point is located halfway between the other two. This can be seen in Figure 5.7 where the simulated effect and the location of the estimated optimal cut points with the *AddFor* with a grid of size $M = 100$ and the *Genetic* for $k = 2$ and $k = 3$ number of cut points are depicted. This suggests that the cut points obtained for $k = 2$ correspond to changes in risk, whereas the location of the third cut point obtained when sought for $k = 3$ divides the category with a higher number of individuals. Nevertheless, the estimated values for this cut point with both algorithms was almost the same.

In addition to the comparison of the estimated optimal cut points, the aim of this simulation study was to compare the estimated bias-corrected AUCs obtained with both algorithms. Figure 5.8 depicts the boxplot of the bias-corrected AUCs for the optimal categorisation over 500 simulated data sets, for each of the proposed algorithms, different sample sizes and number of cut points. The corresponding numerical results are shown in Table 5.7. Note that the *Genetic* approach generally provided slightly higher AUC values than the *Addfor* algorithm. However, when the *Addfor* grid size was increased from 100 to 1000, the obtained results were almost the same as those obtained with the *Genetic* algorithm. For instance, for a sample size of 500 and a desired number of three cut points, average bias-corrected AUC values of 0.810 and 0.809 were respectively obtained by the *Genetic* algorithm and the *AddFor* algorithm with a grid size of 1000, while the average bias-corrected AUC value obtained with the *AddFor* algorithm for a grid size of 100 was 0.805.

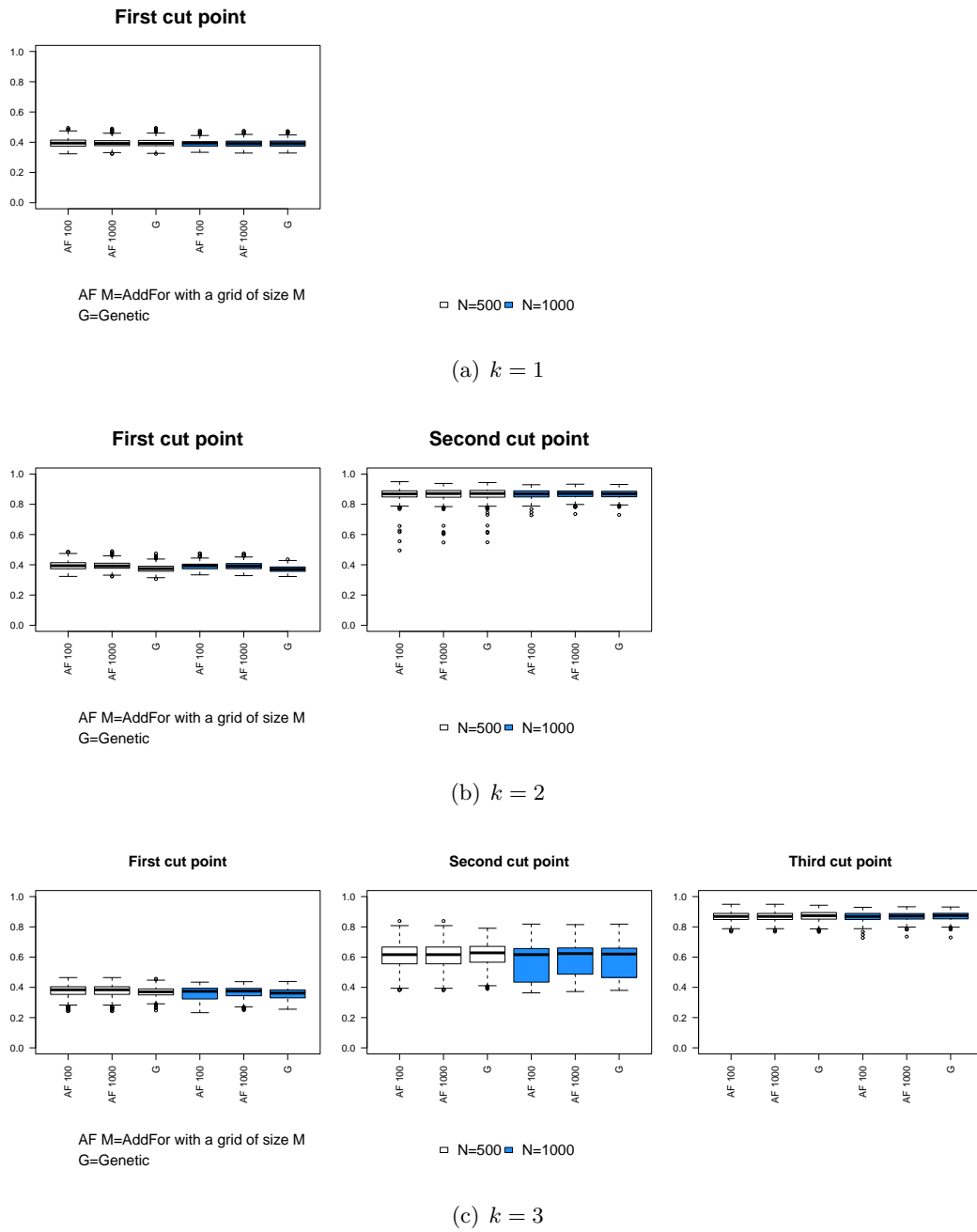


Figure 5.6: Boxplot of the estimated cut points based on 500 simulated data sets obtained according to the comparative study of the proposed algorithms. From top to bottom: a) for $k = 1$ number of cut points; (b) for $k = 2$ number of cut points; and (c) for $k = 3$ number of cut points.

Table 5.6: Mean, standard deviation and median values of the estimated optimal cut points over 500 simulated data sets obtained according to the comparison under a nonlinear effects simulation study.

No. of cut points	Method	Cut point Estimation			
		Mean (sd)	Median	Mean (sd)	Median
		Sample Size $N = 500$		Sample Size $N = 1000$	
k=1	Addfor 100	0.394 (0.029)	0.394	0.391 (0.024)	0.394
	Addfor 1000	0.394 (0.027)	0.392	0.392 (0.024)	0.391
	Genetic	0.395 (0.028)	0.393	0.392 (0.024)	0.391
k=2	Addfor 100	0.394 (0.029)	0.394	0.391 (0.024)	0.394
		0.866 (0.042)	0.869	0.868 (0.028)	0.869
	Addfor 1000	0.394 (0.027)	0.392	0.392 (0.024)	0.391
		0.865 (0.042)	0.871	0.868 (0.027)	0.871
	Genetic	0.376 (0.026)	0.374	0.372 (0.021)	0.370
		0.866 (0.041)	0.871	0.866 (0.028)	0.869
k=3	Addfor 100	0.373 (0.041)	0.384	0.358 (0.044)	0.374
		0.597 (0.099)	0.616	0.572 (0.112)	0.616
		0.869 (0.031)	0.869	0.868 (0.028)	0.869
	Addfor 1000	0.373 (0.041)	0.384	0.364 (0.041)	0.375
		0.598 (0.099)	0.616	0.584 (0.106)	0.623
		0.869 (0.031)	0.869	0.868 (0.027)	0.871
	Genetic	0.368 (0.034)	0.370	0.356 (0.036)	0.362
		0.609 (0.088)	0.628	0.583 (0.101)	0.620
		0.871 (0.032)	0.874	0.870 (0.028)	0.875

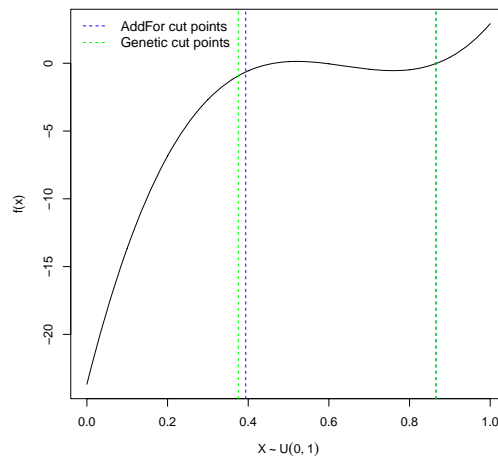
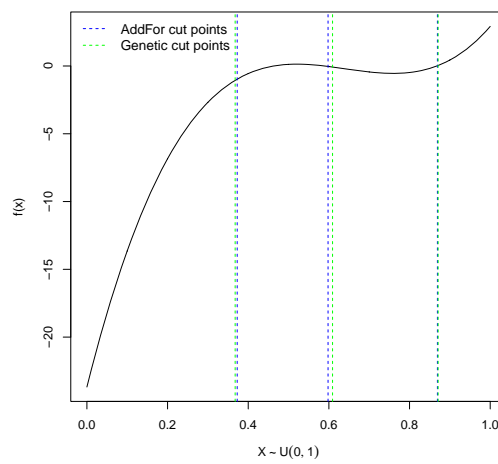
(a) $k = 2$ (b) $k = 3$

Figure 5.7: Simulated effect together with the location of the estimated optimal cut points with the *AddFor* ($M = 100$) and *Genetic* algorithms obtained over 500 simulated data and $N = 500$ sample size according to the comparison under a nonlinear effects simulation study. From left to right: (a) $k = 2$; and (b) $k = 3$.

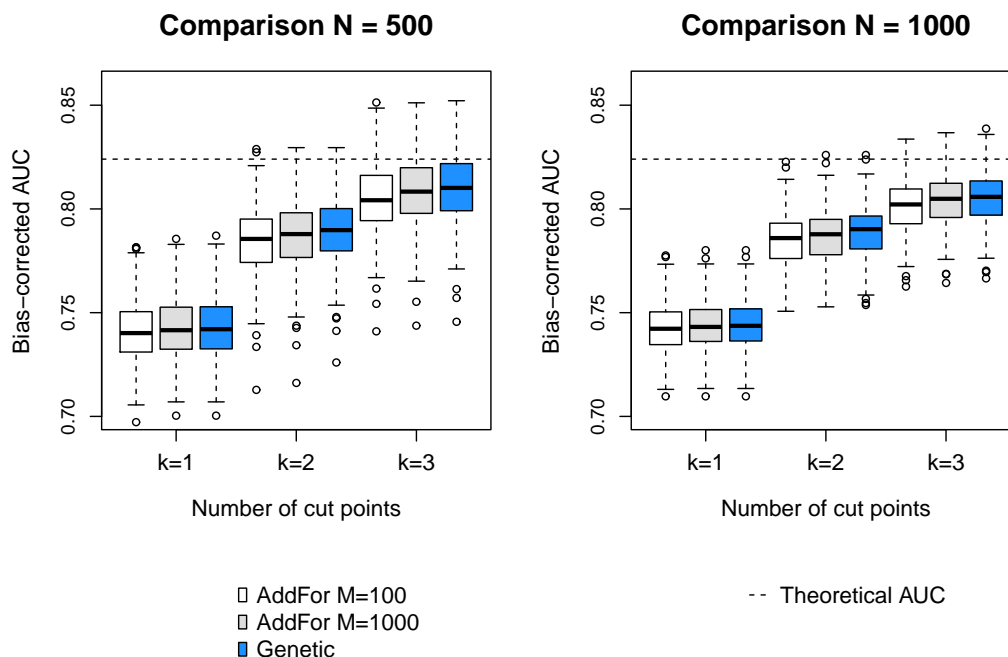


Figure 5.8: Boxplot of the bias-corrected AUCs for the optimal categorisation based on 500 simulated data sets obtained according to the comparison under a nonlinear effects simulation study.

5.2.3 Backward validation

Scenarios and Set-Up

In the third setting, we envisaged simulating a continuous variable X starting from a categorical variable whose cut points had been scientifically pre-established, and assuming that they represent an underlying continuum variable. The aim was to test whether the cut points obtained by applying the proposed methodology to the continuous variable were similar to the original cut-points. For this purpose, we considered the data set available at the IRYSS-COPD study (Quintana et al. 2011). In this data set, we selected the variable forced expiratory volume in 1 second in percentage ($FEV_{1\%}$) which was available both in a continuous scale and an ordinal scale. Although, there is no consensus about the best cut points to categorise this variable, in this study the cut points provided by the GOLD were used (Global Initiative for Chronic Obstructive Lung Disease 2013). Hence, the $FEV_{1\%}$ variable

Table 5.7: Average (standard deviation) of the bias-corrected AUC values over 500 simulated data sets obtained according to the comparison under nonlinear effects simulation study.

Method	No. of cut points	Sample size	
		$N = 500$	$N = 1000$
<i>AddFor</i> $M = 100$	1	0.741 (0.015)	0.742 (0.011)
	2	0.785 (0.016)	0.785 (0.012)
	3	0.805 (0.016)	0.802 (0.012)
<i>AddFor</i> $M = 1000$	1	0.743 (0.015)	0.744 (0.011)
	2	0.787 (0.016)	0.787 (0.012)
	3	0.809 (0.016)	0.804 (0.012)
<i>Genetic</i>	1	0.743 (0.015)	0.744 (0.011)
	2	0.790 (0.016)	0.789 (0.011)
	3	0.810 (0.016)	0.805 (0.012)
Continuous predictor's theoretical AUC			0.824

was categorised into four categories: mild ≥ 80 , moderate $[50 - 80)$, severe $[30 - 50)$ and very severe < 30 . This variable was available for a total number of $L = 2069$ patients.

To simulate the continuous covariate $FEV_{1\%}$ we propose a bootstrap method starting from the original categorical and continuous versions of $FEV_{1\%}$. Let us denote X the original continuous $FEV_{1\%}$ variable and X_{cat} the categorised variable taking values from 0 to 3, which correspond to mild, moderate, severe and very severe categories, respectively. For each $l = 0, \dots, 3$, consider d_{ls} as the s -th decile of X when $X_{cat} = l$. For each $u = 1, \dots, U$ and $l = 0, \dots, 3$, we generated the bootstrap sample $\{x_{iu}^*\}_{i=1}^{L_l}$ by drawing a sample of size L_l with replacement from the original sample $\{x_i\}_{i=1}^{L_l}$, where L_l denotes the number of individuals in the l -th category ($L = \sum_{l=0}^3 L_l$). We considered d_{ls}^* as the average of the U bootstrap deciles of each category, i.e., $d_{ls}^* = \frac{1}{U} \sum_{u=1}^U d_{ls}^u$. The continuous variable X_{sim} was simulated assuming a uniform distribution in the interval $(d_{l(s-1)}^*, d_{ls}^*)$, enclosed by the lower and upper limits of each category.

Additionally the dichotomous response variable Y was simulated according to the two scenarios that are shown in Table 5.8, trying to mimic two possible real situations. In Scenario I patients are distributed as 35%, 30%, 20% and 15% in mild, moderate, severe and very severe categories, respectively. In contrast, in Scenario

II, only 3% of patients belong to the mild category. Additionally, the percentage of individuals with $Y = 1$ (denoted as diseased), changes from Scenario I to Scenario II.

For each of the scenarios, $R = 500$ replicates were conducted for total sample sizes of $N = 500$ and $N = 1000$, and $B = 50$ and $U = 10,000$ bootstrap resamples were used. Optimal cut points were sought using the *Genetic* and *AddFor* algorithms, the latter with grid sizes of $M = 100$ and $M = 1000$.

Table 5.8: Backward validation study: total distribution of individuals in the four categories and distribution of diseased individuals in each category, under both scenarios.

$FEV_{1\%}$ [0,100]	Scenario I		Scenario II	
	Total	Diseased	Total	Diseased
Mild [80,100]	35%	5%	3%	0%
Moderate [50,80)	30%	20%	30%	4.5%
Severe [30,50]	20%	25%	47%	8.6%
Very severe [0,30)	15%	40%	20%	14.2%

Results

The backward-validation simulation study showed that whenever they were clinically significant in the sample, both the *AddFor* and *Genetic* algorithms were able to detect the original cut points. This can be observed in Figure 5.9 where the boxplots of the estimated optimal cut points based on 500 simulated data sets are depicted, for each of the proposed algorithms and different sample sizes. The corresponding numerical results are shown in Table 5.9 where the average of the optimal cut points together with the original cut points are shown. Note that the cut points obtained with the *Genetic* algorithm were slightly closer to the original cut points than the ones obtained with the *AddFor*. For instance, under Scenario I and a sample size of $N = 1000$, the averages of the estimated optimal cut points obtained by the *Genetic* algorithm were 32.03, 53.98 and 77.99, while the ones obtained with the *AddFor* method with a grid of size $M = 1000$ were 32.96, 56.91 and 77.17. It is worth remembering that the original three cut points were 30, 50 and 80. The results shown in Table 5.9 also show that under Scenario II, only two of the original three cut points were detected. The percentage of patients with a $FEV_{1\%}$ of over 80 was less than 3%, and none of them was diseased. Hence, having few individuals with

values above 80, the method was not able to detect that cut point. In this situation, the first two cut points were retained and the original cut points of 30 and 50 were detected.

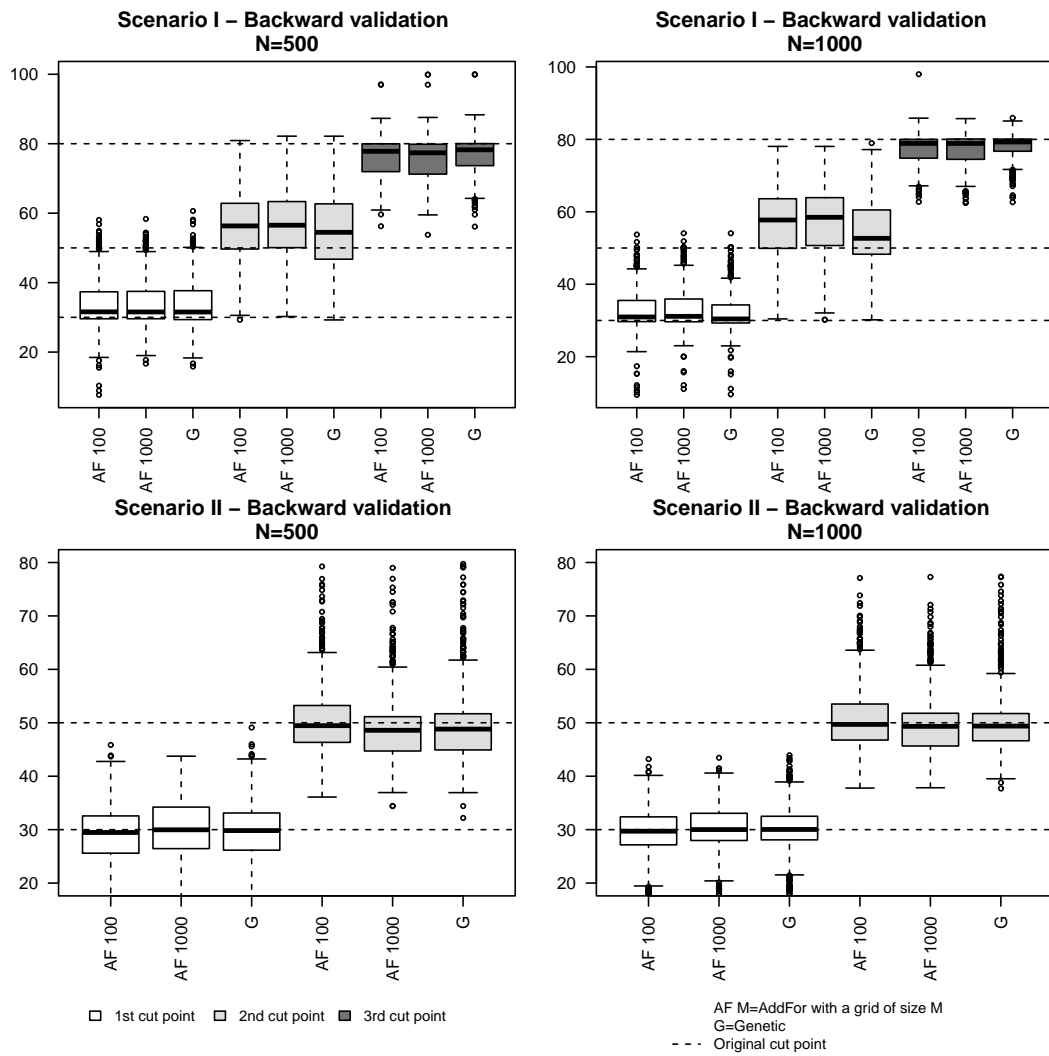


Figure 5.9: Boxplot of the estimated optimal cut points based on 500 simulated data sets obtained according to the backward validation study.

Table 5.9: Backward validation study: average of the estimated optimal cut points over 500 simulated data sets obtained.

Scenario	Sample size $N = 500$			Sample size $N = 1000$			Theoretical cut point
	<i>AddFor</i>	<i>AddFor</i>	<i>Genetic</i>	<i>AddFor</i>	<i>AddFor</i>	<i>Genetic</i>	
	$M = 100$	$M = 1000$		$M = 100$	$M = 1000$		
I							
1st cut point	33.72	33.99	33.93	32.67	32.96	32.03	30
2nd cut point	55.83	56.03	56.10	56.50	56.91	53.98	50
3rd cut point	75.98	75.55	79.03	77.29	77.17	77.99	80
II							
1st cut point	29.23	30.16	29.89	29.35	30.23	30.05	30
2nd cut point	50.61	49.21	49.89	51.02	50.02	50.39	50

5.3 Application to the IRYSS-COPD study

We applied the methodology proposed in this chapter to the IRYSS-COPD study presented in Chapter 2, Section 2.1. As pointed out before, preliminary analysis during the development of a prediction model for patients with eCOPD showed that clinical parameters related to short-term very severe evolution were the Glasgow comma scale (0: altered, 1:normal), the heart rate and the PCO_2 . Moreover, this preliminary analysis also suggested that the relationship between the heart rate and the response variable short-term very severe evolution appeared to be linear, while the relationship between the PCO_2 and the response variable did not. This can be seen in Figure 5.10 where the estimated effects of both heart rate and PCO_2 based on a logistic GAM (Wood 2006) are depicted. For this reason clinical researchers decided to introduce a categorised version of the PCO_2 variable into the prediction model.

As a first step, we considered categorising the PCO_2 variable into two, three and four categories in a univariate setting. To determine the optimal cut points we applied the two algorithms presented in Section 5.1.

Table 5.10 shows the results obtained in the categorisation of the predictor PCO_2 with the *AddFor* and the *Genetic* algorithms. For each number of cut points ($k = 1, 2$ and 3), the obtained optimal cut points together with the bias-corrected AUC are reported. Additionally, the difference in the bias-corrected AUCs, as well as the IDI indexes when compared models with 1 and 2 and 2 and 3 cut points are shown.

As can be observed, the cut points obtained with the *Genetic* and *AddFor* al-

Table 5.10: Results obtained in the categorisation of the predictor variable PCO_2 of the IRYSS-COPD study in a univariate setting. Estimated optimal cut points, bias-corrected AUC, difference of the bias-corrected AUC and confidence interval for this difference together with the IDI and its confidence interval are reported.

Method	k	Estimated cut points	Bias-corrected AUC	AUC difference (95% CI^*)	IDI (95% CI)
<i>Addfor</i> $M = 100$	1	50.87	0.674	0.022 (0.011, 0.036)	0.016 (0.008, 0.024)
	2	50.87; 62.67	0.696	0.014 (-0.003, 0.042)	0.001 (-0.0003, 0.002)
	3	47.92; 50.87; 62.67	0.709		
<i>Addfor</i> $M = 1000$	1	50.1	0.674	0.022 (0.011, 0.036)	0.016 (0.008, 0.024)
	2	50.1; 62.08	0.696	0.016 (-0.002, 0.045)	0.001 (-0.0003, 0.002)
	3	45.86; 50.1; 62.08	0.712		
<i>Genetic</i>	1	50.87	0.674	0.032 (0.010, 0.065)	0.016 (0.008, 0.025)
	2	47.74; 62.64	0.706	0.006 (-0.002, 0.025)	0.0002 (-0.0003, 0.001)
	3	34.06; 47.52; 62.58	0.713		

* 95% bootstrap confidence interval based on the percentile method for $V = 100$ bootstrap replicates; k : number of cut points sought.

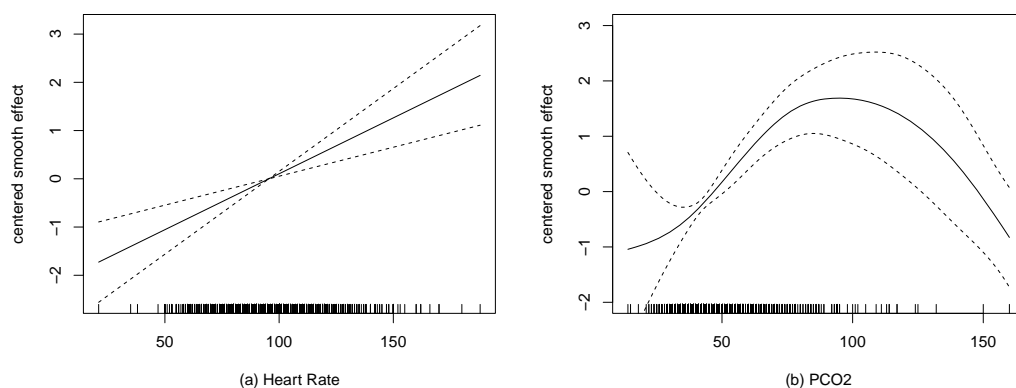


Figure 5.10: From left to right: (a) Relationship of the predictor variable Heart Rate with short-term very severe evolution adjusted by Glasgow and PCO₂ covariates. (b) Relationship of the predictor variable PCO₂ with short-term very severe evolution adjusted by Glasgow and heart rate covariates.

gorithms were quite similar, with those obtained for $k = 1$ being 50.10 and 50.87, those obtained for $k = 2$ being (50.10, 62.08) and (47.74, 62.64) and those obtained for $k = 3$ being (45.86, 50.10, 62.08) and (34.06, 47.52, 62.58), using the *AddFor* with a grid of size 1000 and the *Genetic* algorithms respectively. Note that values for the PCO₂ in the IRYSS-COPD study are recorded as integer numbers.

In the case of the *Genetic* algorithm, bias-corrected AUCs of 0.674, 0.706 and 0.713 were obtained for $k = 1, 2$ and 3, respectively. A difference (95% bootstrap CI) of 0.032 (0.010, 0.065) was obtained between AUCs for $k = 2$ and $k = 1$ cut points and a difference of 0.006 (-0.002, 0.025) between AUCs for $k = 3$ and $k = 2$ cut points. The IDI obtained when passed from $k = 1$ to $k = 2$ cut points was 0.016 (p-value = 0.0002). However, when passed from $k = 2$ to $k = 3$ cut points the IDI was 0.0002 (p-value = 0.385).

In the case of the *AddFor* algorithm with a grid of size 1000, bias-corrected AUCs of 0.674, 0.696 and 0.712 were obtained for $k = 1, 2$ and 3 respectively. A difference of 0.022 (0.011, 0.036) was obtained between AUCs for $k = 2$ and $k = 1$ cut points and a difference of 0.016 (-0.002, 0.045) between AUCs for $k = 3$ and $k = 2$ cut points. The IDI obtained when passed from $k = 1$ to $k = 2$ cut points was 0.016 (p-value = 0.0002). However, when passed from $k = 2$ to $k = 3$ cut points the IDI

was 0.0007 (p-value = 0.147).

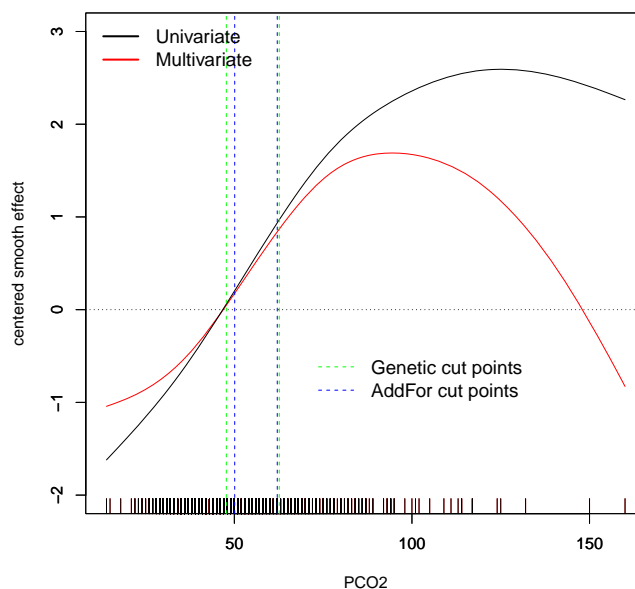


Figure 5.11: Estimated smooth relationship of the predictor variable PCO_2 with the response variable short-term very severe evolution in a univariate setting and in a multivariate setting adjusted by Glasgow and heart rate covariates, jointly with the cut points obtained with the *AddFor* and *Genetic* methods.

Summarising, the results suggest that the optimal number of cut points is two, being the vector of optimal cut points $\hat{\mathbf{v}}_2 = (47.74, 62.64)$ or $\hat{\mathbf{v}}_2 = (50.1, 62.08)$ if the *Genetic* or *AddFor* algorithm is chosen.

In addition, we obtained a 92% agreement between the categorical variables achieved with the *AddFor* and *Genetic* algorithms, measured by Cohen's weighted kappa (Cohen 1968), with a 95% CI of (0.91, 0.93). These results were face-validated by the clinicians involved in the IRYSS-COPD study.

Finally, we considered categorising the predictor variable PCO_2 in a multivariate setting adjusted by the other predictor variables considered by clinicians, which were Glasgow comma scale and heart rate. The cut points obtained for $k = 2$ were (47.03, 62.08) and (47.33, 62.54) with the *AddFor* and *Genetic* algorithms, respectively. In this case, the effect of the other covariates in the multivariate model did not change the optimal cut points obtained for the PCO_2 covariate. This result could

be explained by looking at the estimated effects shown in Figure 5.11, where it can be seen that the shape of the relationship between the continuous predictor and the response variable does not change from the univariate to the multivariate setting.

5.4 Conclusions

The main objective of the work presented in this chapter was to develop a valid method to obtain optimal cut points to categorise continuous predictor variables in a univariate or multivariate logistic regression model by the maximisation of the AUC. To do so, we first proposed two algorithms, namely *AddFor* and *Genetic*, to search for the optimal cut points. Additionally, we recommended a bias correction for the AUC together with a proposal for the selection of the optimal number of cut points.

The advantages that this proposal presents with respect to previously published proposals for the selection of more than one cut point are: a) it requires no distributional assumptions and can be used in any situation regardless of the distribution of the original continuous predictor (Tsuruta and Bax 2006); and b) it provides the objectivity afforded by an automatic method as opposed to the subjectivity of relying on a graphical display, as happens in the methodology proposed in Chapter 4. Furthermore, our approach has been developed so that a continuous predictor variable can be categorised both in a univariate or a multivariate context, depending on what the underlying setting is for each data set (univariate or multivariate) as proposed by Mazumdar et al. (2003). Although in the application to the IRYSS-COPD study the cut points obtained for the PCO₂ covariate in the univariate and multivariate settings were almost the same, this does not always need to be so. The cut points obtained in the multivariate setting may differ from those obtained in the univariate model. For example, if the relationship between the continuous predictor and the response variable is different in a univariate or a multivariate setting, the optimal cut points may be different. This could happen, for example, when confusion predictors are present in the multivariate model. Hence, in contrast to other categorisation methods, the proposed methodology thus enables a continuous variable to be categorised before or during the development of a prediction model, thereby allowing for the incorporation of potential cofounders.

Although we have not mentioned it explicitly, the number of individuals in each category is a relevant issue. In fact, if there are few individuals in one of the categories, or too many categories are considered, the estimates may be unstable


(O'Brien 2004). If this happens the maximum likelihood may not exist and hence the logistic regression would not be feasible. The algorithms we developed to select the optimal cut points control for the number of individuals in each category. If the convergence of maximum likelihood is not obtained or if the discriminative ability can not be estimated for a vector of cut points, that vector is not considered as eligible.


The simulation study shows that under the theoretical hypothesis, our approach yields the optimal location of the cut points. Additionally, the results obtained suggest that the cut points obtained correspond to the change at risk of having the outcome of interest. Indeed and according to clinicians' criteria, in the application of the method to the IRYSS-COPD study, the cut points obtained for the clinical parameter PCO_2 , classified patients into low, moderate and high risk of short-term very severe evolution. The proposed methods thus provide a classification of patients in terms of risk, which is precisely what is desirable in the development of prediction models to be used in clinical practice for decision-making.

Nevertheless, this proposal has some limitations that should be taken into account. Despite the fact that the results obtained with both algorithms are similar, one must bear in mind that the *AddFor* algorithm seeks the second cut point once the first has been fixed. Consequently, the selection of the first cut point has an influence on the consecutive cut points, which at times may lead to a non-optimal selection of cut points. This was observed in the simulation study performed under known theoretical conditions. As we have seen, when two cut points were sought, the cut points obtained with the *AddFor* algorithm had greater bias than the ones obtained with the *Genetic* algorithm. However, in some circumstances the *Genetic* algorithm may be not feasible due to its computational cost, especially if very large data sets are considered or many cut points are sought. Nevertheless, in general, as long as it is computationally feasible, we recommend the use of the *Genetic* algorithm.

Categorisation methods in a survival model

The work in this chapter has been previously presented in an international conference and is being prepared to be sent for publication to an international journal.

 *Barrio, I., Rodríguez-Álvarez, M.X., Meira-Machado, L., Esteban, C., and Arostegui, I. Comparison of the c-index and CPE indexes in the polycotomisation of continuous predictors in a Cox Proportional Hazards Model. (In preparation)*

 *XXVII International Biometric Conference. Optimal cut points to categorize continuous predictor variables in a Cox Proportional Hazards Model. Barrio, I., Rodríguez-Álvarez, M.X., Meira-Machado, L., Quintana, J.M., and Arostegui, I. Poster contribution. Florence July 2014.*

In Chapters 4 and 5 we presented methodologies to categorise continuous variables in the context of logistic regression. As pointed out before, in many circumstances the interest lies in studying the time until the event of interest occurs. Therefore, in this chapter we extend the methodology proposed in Chapter 5 for categorisation of continuous variables in logistic regression to the Cox PH regression model.

The rest of this chapter is organised as follows. In Section 6.1 we present the proposed methodology to categorise a continuous predictor variable in a Cox PH model by maximising the concordance probability index. Specifically, two different estimators were studied: c-index and CPE. Section 6.2 is devoted to an empirical validation of the proposed methodology where we present the scenarios of the simulation study conducted to validate the proposed methodology together with the results obtained. In Section 6.3 we implement the proposed methodology to the Stable-COPD study. Finally, in Section 6.4, we end this chapter with some conclusions and limitations.

6.1 Proposed Methodology

Let T be a non-negative random variable representing the time until the event of interest, and let X denote a continuous covariate that we want to categorise. We propose to categorise X in such a way that the best predictive survival model is obtained, considering the maximal concordance probability achieved. The concordance probability was estimated by two alternative estimators: the c-index proposed by Harrell et al. (1982) and the CPE proposed by Gönen and Heller (2005), as we explained in Chapter 3, Section 3.4.

Specifically, given k the number of cut points set for categorising X in $k + 1$ intervals, we propose that the vector of k cut points $\mathbf{v}_k = (x_1, \dots, x_k)$ which maximises the discriminative ability of the Cox PH model shown in equation (6.1), is thus the vector of the optimal k cut points:

$$h(t|X_{cat_k}) = h_0(t)e^{\beta_0 + \sum_{q=1}^k \beta_q 1_{\{X_{cat_k}=q\}}}. \quad (6.1)$$

Suppose now that along with the predictor variable X we want to categorise, a set of other p predictors, Z_1, \dots, Z_p , are of interest. Then, what we propose is that the categorisation of X in a multivariate Cox PH model including the p predictors, will be that for which the concordance probability of the model (6.2) is maximised.

$$h(t|(Z_1, \dots, Z_p, X_{cat_k})) = h_0(t)e^{\beta_0 + \sum_{r=1}^p \beta_r Z_r + \sum_{q=p+1}^{p+k} \beta_q 1_{\{X_{cat_k}=q-p\}}}. \quad (6.2)$$

Let $\{x_i, \mathbf{z}_i, y_i, \delta_i\}_{i=1}^N$ be a sample of size N , where x_i represents the observed value of the predictor variable we want to categorise; \mathbf{z}_i is the observed value of the set of other p predictors, y_i represents the observed follow-up time for subject i and δ_i is the censoring indicator. Estimation of the models in equations (6.1) and (6.2) as well as of the associated concordance probability, can be done as presented previously in Section 3.3 and Section 3.4 in Chapter 3, above. To estimate the vector of the cut points of X that maximises the c-index and the CPE, we propose the use of the above presented algorithms, namely *AddFor* and *Genetic*.

As happened when we searched for optimal cut points in logistic regression, although the *AddFor* method searches for each cut point at a time, the *Genetic* method simultaneously looks for all cut points that maximise the discriminative ability of the Cox PH model.

6.1.1 Optimism correction

As we presented in Chapter 3, Section 3.4.3, a discriminative ability measure estimator may be biased upward when the same data set is used to fit the model and estimate the model's discriminative ability, even more if the data is censored. The CPE was proposed as an unbiased alternative to Harrel's c-index by Gönen and Heller (2005) when the aim was to estimate the concordance probability and discriminatory power in a Cox PH regression model. Nevertheless, we proposed to correct the bias of both indexes since both were estimated using the same data that was previously used to estimate the vector of optimal cut points. The bootstrap bias correction approach proposed for the concordance probability estimator in a Cox PH model can be summarised as follows:

Let us denote \hat{c} the concordance probability estimator, which can be either the c-index or the CPE.

Step 1. Categorise the predictor variable on the basis of the original sample

$\{(x_i, z_i, y_i, \delta_i)\}_{i=1}^N$ and compute the corresponding concordance probability (see equations (3.27) and (3.28)). Let us denote this *apparent* concordance probability estimator as \hat{c}_{app} .

Step 2. For $b = 1, \dots, B$, generate the bootstrap resample $\{(x_{ib}^*, z_{ib}^*, y_{ib}^*, \delta_{ib}^*)\}_{i=1}^N$ by drawing a random sample of size N with replacement from the original sample, and categorise the bootstrapped predictor $\{x_{ib}^*\}_{i=1}^N$ on the basis of the optimal cut points obtained in Step 1.

Step 3. Fit the Cox PH model to the bootstrap resample with the categorised version of the predictor. Let us denote as $\hat{\beta}^b$ the vector of the estimated regression coefficients based on this bootstrap resample. Compute the corresponding concordance probability, \hat{c}_{boot}^b for $b = 1, \dots, B$.

Step 4. Obtain the linear predictor for the original sample based on the fitted Cox PH regression model obtained in Step 3, i.e,

$$\hat{\beta}_0^b + \sum_{r=1}^p \hat{\beta}_r^b z_{ri} + \sum_{q=p+1}^{p+k} \hat{\beta}_q^b 1_{\{x_{cat_k i}=q\}}$$

and compute the concordance probability. Let's denote this estimator as \hat{c}_o^b for $b = 1, \dots, B$.

Once the above process has been completed, the optimism O of the original concordance probability estimator is calculated as follows:

$$O = \frac{1}{B} \sum_{b=1}^B |\hat{\mathbf{c}}_{boot}^b - \hat{\mathbf{c}}_o^b|$$

and the bias-corrected concordance probability estimator is then computed as $\hat{\mathbf{c}}_{app} - O$.

In the same way as we did for the AUC, we also considered correcting the CPE and the c-index during the selection of the cut points. We computed several simulations similar to what we did in Section 5.2, Chapter 5, and saw that it had no influence on the selection of the optimal cut points. Consequently, for all the simulations and analysis presented in this chapter the c-index or CPE are corrected after the selection of the optimal cut points.

6.1.2 Selection of the optimal number of cut points

Similar to what we proposed in Section 5.1.2 for the logistic regression setting, we propose a bootstrap CI for the difference between the bias-corrected discrimination index of the two categorisation proposals in the Cox PH model in order to determine if an extra category is needed. This methodology is proposed when the maximisation index considered is either the c-index or the CPE.

The procedure to compute the CI for the difference of the bias-corrected discriminative ability index estimator can be summarised as follows. For ease of notation, let us denote $\hat{\mathbf{c}}$ as the discrimination index estimator, which in our specific framework may be either the c-index proposed by Harrell et al. (1982) or the CPE proposed by Gönen and Heller (2005).

Step 1. For $v = 1, \dots, V$, generate the bootstrap resample $\{(x_{iv}^*, z_{iv}^*, y_{iv}^*, \delta_{iv}^*)\}_{i=1}^N$ by drawing a random sample of size N with replacement from the original sample.

Step 2. Compute the bias-corrected discrimination index for the categorised variable for $k = l$ and $k = l + 1$ and denote it as $\hat{\mathbf{c}}_{l,v}^*$ and $\hat{\mathbf{c}}_{l+1,v}^*$ respectively. The bias-corrected discrimination index is computed as explained above in Section 6.1.1, now using for Step 1 the optimal cut points obtained for $k = l$ and $k = l + 1$ on the basis of the original sample.

Step 3. Compute the difference between the bias-corrected discrimination indexes obtained for $k = l + 1$ and $k = l$

$$\widehat{\mathbf{c}}_{Diff,v}^* = \widehat{\mathbf{c}}_{l+1,v}^* - \widehat{\mathbf{c}}_{l,v}^*.$$

Once the above process has been completed, the $(1 - \alpha)$ % limits for the CI for the difference are given by

$$\left(\widehat{\mathbf{c}}_{Diff}^{\alpha/2}, \widehat{\mathbf{c}}_{Diff}^{1-\alpha/2} \right)$$

where $\widehat{\mathbf{c}}_{Diff}^p$ represents the p-percentile of the estimated $\widehat{\mathbf{c}}_{Diff,v}^*$ ($v = 1, \dots, V$).

We propose to determine whether an extra optimal cut point is needed if the CI does not contain the zero.

6.2 Empirical validation

In this section we present a simulation study conducted to analyse the empirical performance of the methodology proposed in Section 6.1, above. In this case, the simulation study was performed in such a way that the theoretical cut points are known. The aims of this simulation study are threefold: a) to compare which of the concordance probability estimators, c-index or CPE, performs better in the selection of optimal cut points; b) to compare the estimated optimal cut points with the theoretical optimal cut points; and c) to compare the bias-corrected c-index and CPE to the theoretical discriminative ability index.

All computations were performed in (64 bit) R 3.1.2 and a workstation equipped with 24GB of RAM, an Intel Xeon E5620 processor (2.40 Ghz), and Windows 7 operating system. Specifically, the `genoud` function of the `rgenoud` (Mebane and Sekhon 2011) package was used to compute the genetic algorithms, the `cph` function of the `rms` package (Harrell 2015) was used for the estimation of the Cox PH model, and the c-index and the `phcpe2` function of the package `CPE` (Mo et al. 2012) was used to estimate the CPE. Finally, the `coxph` function of the `survival` package (Therneau 2014) together with the `termplot` function of the `stats` package were used to plot the effects of the Cox PH model.

Scenarios and set-up

To simulate the data we assumed that X is a continuous predictor variable normally distributed with mean $\mu = 0$ and variance $\sigma = 2$. Considering the theoretically

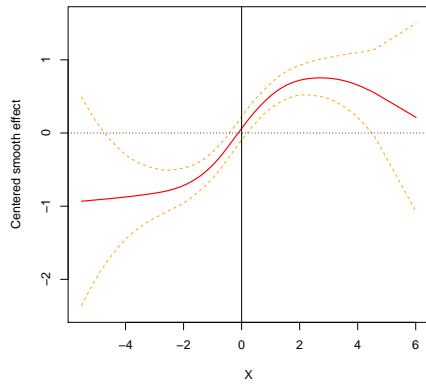
optimal cut points, c_1, c_2, \dots, c_k , we built a categorical variable, X_{cat_k} , such that $X_{cat_k} = 0$ if $X \leq c_1$, $X_{cat_k} = 1$ if $c_1 < X \leq c_2$, \dots , and $X_{cat_k} = k$ if $X > c_k$. Survival times T corresponding to each X_{cat_k} value were generated from different independent Weibull distributions, with shape and scale parameters given by (γ_i, λ_i) for $X_{cat_k} = i$, $0 \leq i \leq k$. The follow-up time was subjected to right censoring, C , according to the uniform model $U(0, \tau)$, and the event indicator δ was defined as $I(T \leq C)$. Simulations were performed for total sample sizes of $N = 500$ and $N = 1000$. As far as the number of cut points is concerned, $k = 1, 2$ and 3 were considered. Finally, for the *AddFor* algorithm, grid sizes of $M = 100$ and $M = 1000$ were used. In all cases, $R = 500$ replicates of simulated data were performed and $B = 50$ was used for CPE and c-index bias correction procedure.

Several settings were considered in this simulation study, which are summarised in Table 6.1. First of all, we considered $k = 1, 2$ and 3 number of cut points. For $k = 1$ we considered a) an increasing risk relationship between the continuous predictor X and survival time T (Scenarios I, II and III); and b) a decreasing risk relationship between the continuous predictor X and survival time T (Scenarios IV, V and VI). Additionally, we considered different positions for the theoretical cut points: a) centred in the predictor's distribution (Scenarios I and IV); b) shifted to high risk area (Scenarios II and VI); and c) shifted to low risk area (Scenarios III and V). For $k = 2$ and $k = 3$ we considered a linear (Scenarios VII and IX) and a nonlinear (Scenarios VIII and X) relationship between the continuous predictor X and survival time T . The relationship between the continuous predictor and the response variable in each scenario is shown in Figure 6.1 and Figure 6.2, and is computed with a smooth function using the `termplot` function of the `stats` package.

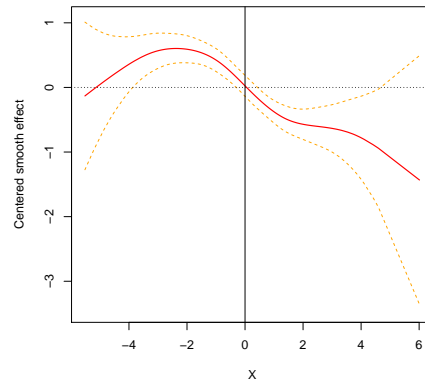
Finally, we also considered a scenario in which the proportional hazards assumption is violated. Specifically, $k = 2$ number of cut points ($c_1 = -1$ and $c_2 = 1$) and the Weibull distribution parameters, $(\gamma_0, \lambda_0) = (1, 0.5)$, $(\gamma_1, \lambda_1) = (3, 3)$ and $(\gamma_2, \lambda_2) = (10, 1)$, were considered to simulate survival times in each category, in such a way that a nonlinear relationship was simulated between the continuous predictor X and survival time T . It should be noted that since the γ parameter of the Weibull distribution is not 1 in all the categories of X_{cat_k} , the proportional hazards assumption does not hold.

Table 6.1: Description of the different scenarios considered for the simulation study. γ and λ are the shape and scale parameters of the Weibull distribution respectively and the censoring indicator δ is defined as $I(T \leq C)$ where $C = U(0, \tau)$.

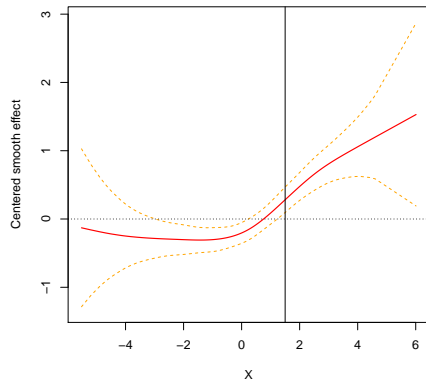
Scenario	Theoretical cut points	Weibull parameters	Censorship (τ)		
			20%	50%	70%
I	0	$\gamma_0 = 1, \lambda_0 = 3$ $\gamma_1 = 1, \lambda_1 = 1$	10	2.8	1.25
II	1.5	$\gamma_0 = 1, \lambda_0 = 3$ $\gamma_1 = 1, \lambda_1 = 1$	12	3.8	1.7
III	-1.5	$\gamma_0 = 1, \lambda_0 = 3$ $\gamma_1 = 1, \lambda_1 = 1$	7	2	1
IV	0	$\gamma_0 = 1, \lambda_0 = 1$ $\gamma_1 = 1, \lambda_1 = 3$	10	2.8	1.25
V	1.5	$\gamma_0 = 1, \lambda_0 = 1$ $\gamma_1 = 1, \lambda_1 = 3$	7	2	1
VI	-1.5	$\gamma_0 = 1, \lambda_0 = 1$ $\gamma_1 = 1, \lambda_1 = 3$	12	3.8	1.7
VII	-1 & 1	$\gamma_0 = 1, \lambda_0 = 0.5$ $\gamma_1 = 1, \lambda_1 = 1$ $\gamma_2 = 1, \lambda_2 = 2$	6	1.5	0.7
VIII	-1 & 1	$\gamma_0 = 1, \lambda_0 = 0.5$ $\gamma_1 = 1, \lambda_1 = 3$ $\gamma_2 = 1, \lambda_2 = 1$	8	2	0.8
IX	-1.5 & 0 & 1.5	$\gamma_0 = 1, \lambda_0 = 0.5$ $\gamma_1 = 1, \lambda_1 = 1$ $\gamma_2 = 1, \lambda_2 = 2$ $\gamma_3 = 1, \lambda_3 = 3$	8	2	1
X	-1.5 & 0 & 1.5	$\gamma_0 = 1, \lambda_0 = 0.5$ $\gamma_1 = 1, \lambda_1 = 3$ $\gamma_2 = 1, \lambda_2 = 1$ $\gamma_3 = 1, \lambda_3 = 0.5$	6	1.5	0.65



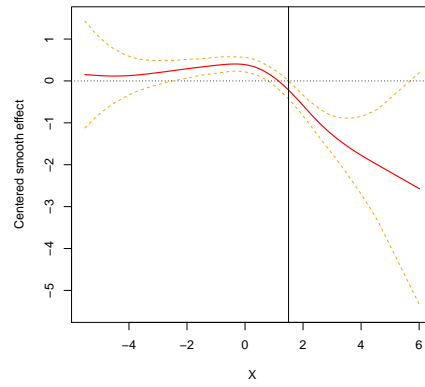
(a) Scenario I



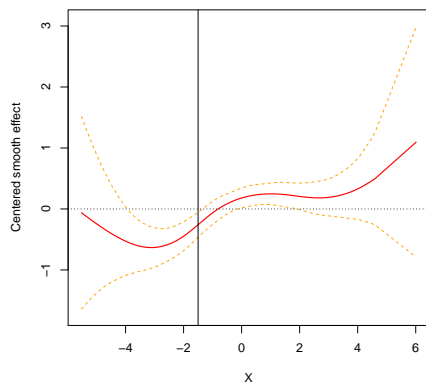
(b) Scenario IV



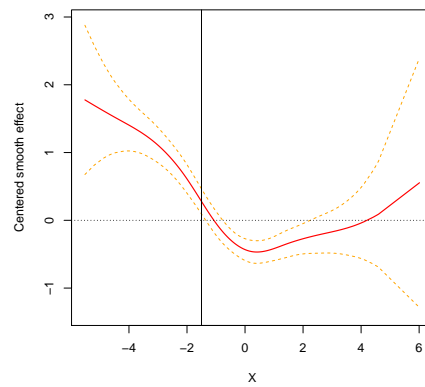
(c) Scenario II



(d) Scenario V



(e) Scenario III



(f) Scenario VI

Figure 6.1: Simulated data for sample size of $N = 500$ and censoring rate of 50% in scenarios I to VI, where one theoretical cut point was considered. In all cases, data from the first replicate is plotted.

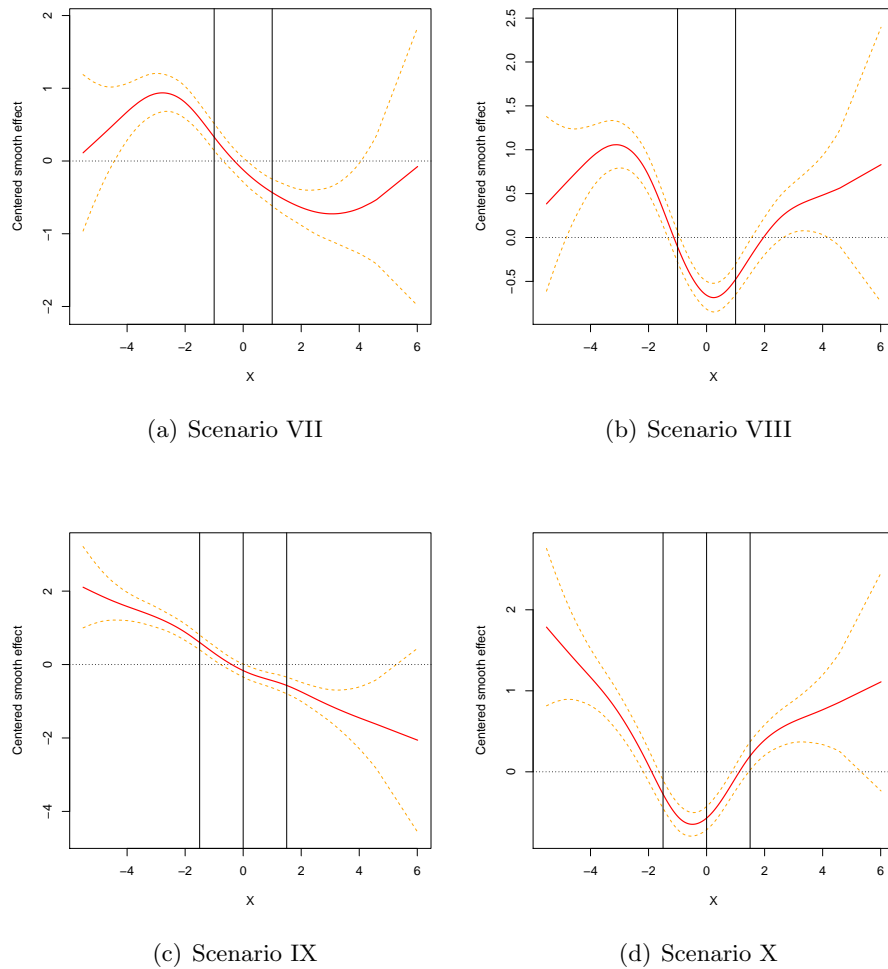


Figure 6.2: Simulated data for sample size of $N = 500$ and censoring rate of 50% in scenarios VII and VIII, where two theoretical cut points were considered (Figures (a) and (b)) and scenarios IX and X where three theoretical cut points were considered (Figures (c) and (d)). In all cases, data from the first replicate is plotted.

Results

Given the large number of proposed scenarios and different conclusions obtained, we begin by summarising the main findings.

The simulations results suggest that in general the CPE works better when it comes to low censoring rates (20%), while the c-index works better when it comes to high censoring rates (70%). Although this trend is true in general, it is not met

in all cases, especially when the goal is to find a single cut point. The method is successful when it comes to searching for two or three cut points. However, when the aim is to search for a unique cut point, the method's performance depends largely on the location of the theoretical optimal cut point. Additionally, the results become worse as the censoring rate increases, regardless of the discrimination index used. However, when the algorithm used is the *Genetic*, the MSE is smaller than with the *AddFor*, obtaining good results even for high censoring rates. Finally, smaller bias and MSE for discrimination indexes are obtained when a sample size of 1000 is used compared to a 500 sample size.

Figure 6.3 depicts the boxplot of the estimated optimal cut points over 500 simulated data sets, for each of the proposed algorithm, c-index and CPE estimator and a sample size of $N = 500$ for Scenarios I, II and III, where a single optimal cut point is searched for an increasing risk relationship between the continuous predictor X and the outcome. Numerical results for these scenarios are given in Table 6.2 (Scenario I), Table 6.3 (Scenario II) and Table 6.4 (Scenario III). The obtained results show that when the theoretical optimal cut point is centred, i.e., $c_1 = 0$, the proposed method performs satisfactorily regardless of the discrimination index used and censorship rate. This can be observed in Figure 6.3(a) and Table 6.2. However, when the theoretical cut is offset, the method is not able to find it, particularly when the censoring rate is high. At this point we must clarify the fact that depending on whether the cut point is shifted to the area of high risk ($c_1 = 1.5$) or low risk ($c_1 = -1.5$), differences between using the CPE or the c-index are considerable. Differences can be observed when comparing Figures 6.3(b) and 6.3(c).

Figure 6.4 depicts the boxplot of the estimated optimal cut points over 500 simulated data sets, for each of the proposed algorithm, c-index and CPE estimator, 20% and 70% censoring rates and a sample size of $N = 500$ for scenarios IV, V and VI, where a single optimal cut point is searched for a decreasing risk relationship between the continuous predictor X and the outcome. Numerical results for these scenarios are given in Table 6.5 (Scenario VI), Table 6.6 (Scenario V) and Table 6.7 (Scenario VI). Similar results to the ones obtained for Scenarios I to III are obtained in Scenarios IV to VI. As can be observed in Figure 6.4(a), when the theoretical cut point is centred, in this case $c_1 = 0$, the method performs satisfactorily and no differences are observed between the performance of the CPE and c-index. However, when the theoretical cut point is $c_1 = 1.5$, that is, it is shifted to the area of low-risk (Scenario V), the estimation of the optimal cut points obtained with the c-index

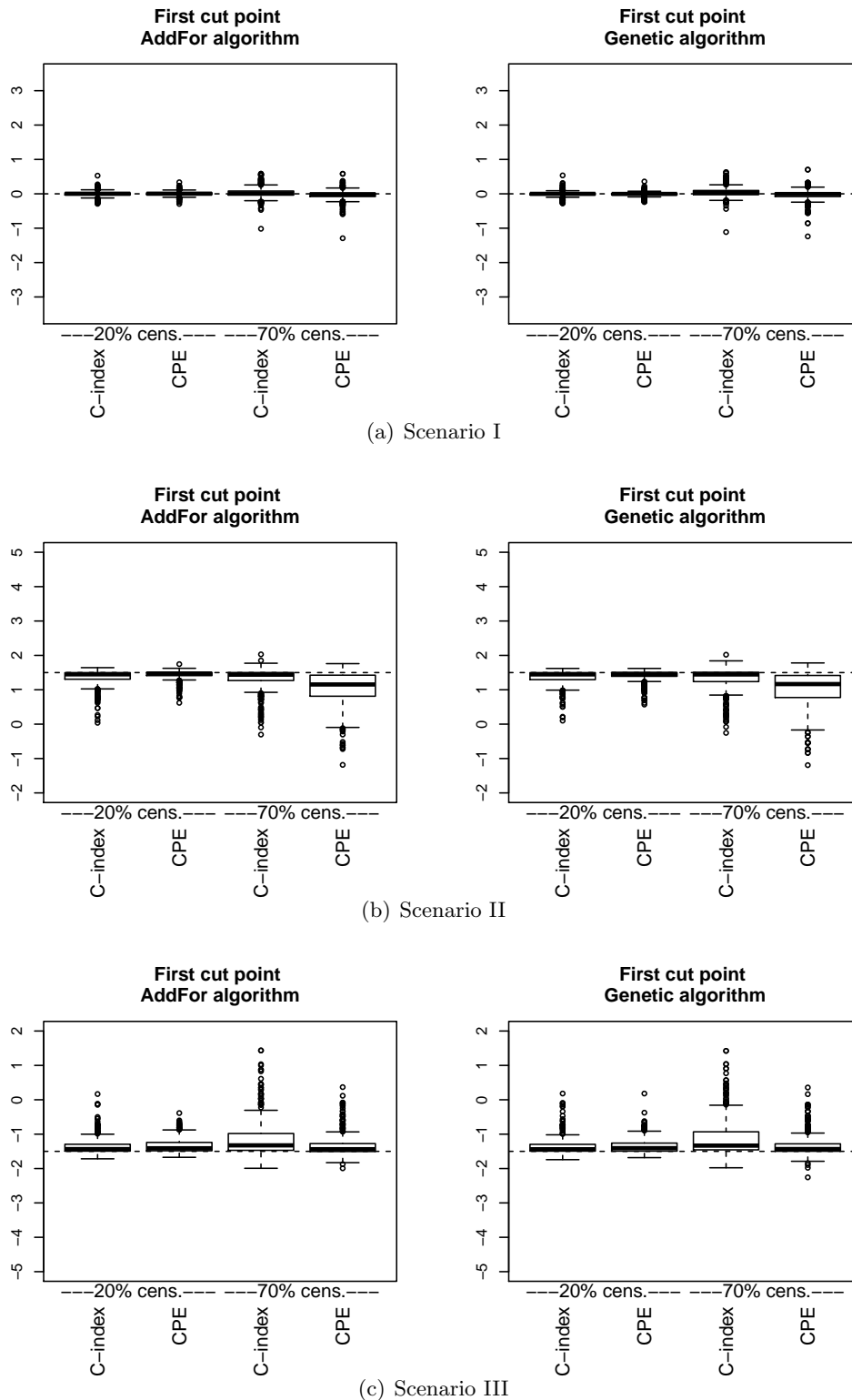


Figure 6.3: Boxplot of the estimated optimal cut points based on 500 simulated data sets, $N = 500$ sample size, one theoretical cut point and increasing risk relationship with the outcome. Results are shown for *AddFor* ($M = 100$) and *Genetic* algorithms, censoring rates of 20% and 70% and c-index and CPE discriminative ability estimators. From top to bottom: (a) theoretical cut points (0); (b) theoretical cut points (1.5); and (c) theoretical cut points (-1.5).

Table 6.2: Simulation results when one theoretical optimal cut point 0 was chosen with an increasing risk relationship with the outcome $(\gamma_0, \lambda_0) = (1, 3)$ and $(\gamma_1, \lambda_1) = (1, 1)$ and censorship of 20%, 50% and 70% (Scenario I). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	0	0.002 (0.060)	0.004	0.002	0.004	-0.002 (0.077)	0.002	-0.002	0.006
	50%	0	-0.025 (0.118)	-0.002	-0.025	0.015	-0.002 (0.100)	0.003	-0.002	0.010
	70%	0	-0.034 (0.142)	-0.014	-0.034	0.021	0.033 (0.160)	0.012	0.033	0.027
Addfor 1000	20%	0	-0.011 (0.056)	-0.007	-0.011	0.003	-0.009 (0.078)	-0.005	-0.009	0.006
	50%	0	-0.032 (0.118)	-0.010	-0.032	0.015	-0.003 (0.098)	-0.001	-0.003	0.010
	70%	0	-0.038 (0.154)	-0.011	-0.038	0.025	0.037 (0.154)	0.014	0.037	0.025
Genetic	20%	0	-0.008 (0.055)	-0.003	-0.008	0.003	-0.006 (0.078)	-0.001	-0.006	0.006
	50%	0	-0.030 (0.116)	-0.007	-0.030	0.014	-0.002 (0.096)	0.001	-0.002	0.009
	70%	0	-0.032 (0.155)	-0.006	-0.032	0.025	0.041 (0.152)	0.021	0.041	0.025
Sample Size N = 1000										
Addfor 100	20%	0	0.009 (0.030)	0.008	0.009	0.001	-0.002 (0.043)	0.001	-0.002	0.002
	50%	0	-0.005 (0.048)	-0.001	-0.005	0.002	0.008 (0.062)	0.004	0.008	0.004
	70%	0	-0.013 (0.070)	-0.006	-0.013	0.005	0.024 (0.090)	0.008	0.024	0.009
Addfor 1000	20%	0	-0.003 (0.025)	-0.002	-0.003	0.001	-0.001 (0.031)	0.000	-0.001	0.001
	50%	0	-0.007 (0.042)	-0.003	-0.007	0.002	0.009 (0.055)	0.002	0.009	0.003
	70%	0	-0.014 (0.084)	-0.002	-0.014	0.007	0.025 (0.084)	0.010	0.025	0.008
Genetic	20%	0	-0.002 (0.026)	-0.001	-0.002	0.001	-0.001 (0.034)	0.000	-0.001	0.001
	50%	0	-0.005 (0.043)	-0.001	-0.005	0.002	0.010 (0.058)	0.003	0.010	0.003
	70%	0	-0.013 (0.084)	-0.001	-0.013	0.007	0.027 (0.085)	0.011	0.027	0.008

Table 6.3: Simulation results when one theoretical optimal cut point 1.5 was chosen with an increasing risk relationship with the outcome $(\gamma_0, \lambda_0) = (1, 3)$ and $(\gamma_1, \lambda_1) = (1, 1)$ and censorship of 20%, 50% and 70% (Scenario II). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	1.5	1.426 (0.140)	1.473	-0.074	0.025	1.364 (0.221)	1.449	-0.136	0.067
	50%	1.5	1.287 (0.299)	1.398	-0.213	0.135	1.352 (0.258)	1.441	-0.148	0.088
	70%	1.5	1.032 (0.485)	1.154	-0.468	0.454	1.334 (0.308)	1.439	-0.166	0.122
Addfor 1000	20%	1.5	1.411 (0.136)	1.456	-0.089	0.026	1.356 (0.211)	1.436	-0.144	0.065
	50%	1.5	1.284 (0.282)	1.386	-0.216	0.126	1.349 (0.261)	1.442	-0.151	0.091
	70%	1.5	1.015 (0.513)	1.162	-0.485	0.497	1.320 (0.320)	1.442	-0.180	0.135
Genetic	20%	1.5	1.414 (0.145)	1.463	-0.086	0.028	1.364 (0.212)	1.447	-0.136	0.063
	50%	1.5	1.294 (0.279)	1.391	-0.206	0.120	1.355 (0.264)	1.450	-0.145	0.090
	70%	1.5	1.023 (0.515)	1.166	-0.477	0.492	1.325 (0.320)	1.446	-0.175	0.133
Sample Size N = 1000										
Addfor 100	20%	1.5	1.471 (0.069)	1.488	-0.029	0.006	1.425 (0.124)	1.467	-0.075	0.021
	50%	1.5	1.358 (0.195)	1.446	-0.142	0.058	1.420 (0.133)	1.466	-0.080	0.024
	70%	1.5	1.160 (0.323)	1.248	-0.340	0.219	1.404 (0.164)	1.460	-0.096	0.036
Addfor 1000	20%	1.5	1.460 (0.065)	1.481	-0.040	0.006	1.424 (0.120)	1.471	-0.076	0.020
	50%	1.5	1.357 (0.196)	1.436	-0.143	0.059	1.427 (0.123)	1.474	-0.073	0.020
	70%	1.5	1.152 (0.336)	1.258	-0.348	0.234	1.410 (0.152)	1.469	-0.090	0.031
Genetic	20%	1.5	1.463 (0.064)	1.482	-0.037	0.005	1.426 (0.120)	1.474	-0.074	0.020
	50%	1.5	1.361 (0.196)	1.440	-0.139	0.058	1.427 (0.125)	1.476	-0.073	0.021
	70%	1.5	1.150 (0.335)	1.207	-0.350	0.235	1.414 (0.154)	1.473	-0.086	0.031

Table 6.4: Simulation results when one theoretical optimal cut point -1.5 was chosen with an increasing risk relationship with the outcome $(\gamma_0, \lambda_0) = (1, 3)$ and $(\gamma_1, \lambda_1) = (1, 1)$ and censorship of 20%, 50% and 70% (Scenario III). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	-1.5	-1.339 (0.201)	-1.411	0.161	0.066	-1.355 (0.228)	-1.435	0.145	0.073
	50%	-1.5	-1.375 (0.223)	-1.439	0.125	0.065	-1.269 (0.354)	-1.401	0.231	0.179
	70%	-1.5	-1.336 (0.318)	-1.438	0.164	0.128	-1.128 (0.519)	-1.326	0.372	0.407
Addfor 1000	20%	-1.5	-1.354 (0.209)	-1.418	0.146	0.065	-1.362 (0.235)	-1.443	0.138	0.074
	50%	-1.5	-1.362 (0.243)	-1.447	0.138	0.078	-1.261 (0.358)	-1.403	0.239	0.185
	70%	-1.5	-1.339 (0.314)	-1.442	0.161	0.124	-1.114 (0.539)	-1.345	0.386	0.439
Genetic	20%	-1.5	-1.347 (0.209)	-1.408	0.153	0.067	-1.355 (0.240)	-1.439	0.145	0.078
	50%	-1.5	-1.357 (0.242)	-1.439	0.143	0.079	-1.255 (0.355)	-1.383	0.245	0.186
	70%	-1.5	-1.333 (0.308)	-1.433	0.167	0.123	-1.107 (0.530)	-1.333	0.393	0.435
Sample Size N = 1000										
Addfor 100	20%	-1.5	-1.379 (0.161)	-1.444	0.121	0.040	-1.412 (0.159)	-1.465	0.088	0.033
	50%	-1.5	-1.422 (0.150)	-1.469	0.078	0.028	-1.339 (0.258)	-1.429	0.161	0.092
	70%	-1.5	-1.410 (0.207)	-1.470	0.090	0.051	-1.251 (0.339)	-1.375	0.249	0.177
Addfor 1000	20%	-1.5	-1.384 (0.170)	-1.454	0.116	0.043	-1.405 (0.171)	-1.468	0.095	0.038
	50%	-1.5	-1.427 (0.153)	-1.482	0.073	0.029	-1.340 (0.262)	-1.434	0.160	0.094
	70%	-1.5	-1.411 (0.200)	-1.474	0.089	0.048	-1.252 (0.339)	-1.372	0.248	0.176
Genetic	20%	-1.5	-1.382 (0.171)	-1.453	0.118	0.043	-1.407 (0.169)	-1.471	0.093	0.037
	50%	-1.5	-1.425 (0.153)	-1.479	0.075	0.029	-1.334 (0.264)	-1.427	0.166	0.097
	70%	-1.5	-1.408 (0.201)	-1.473	0.092	0.049	-1.248 (0.340)	-1.368	0.252	0.179

have a big bias and MSE, especially for censoring rates over 50%. Detailed results are shown in Table 6.6 and Figure 6.4(b). On the other hand, when the theoretical cut point is $c_1 = -1.5$, this is, it is shifted to the area of high-risk (Scenario VI), the c-index performs better than the CPE (see Table 6.7 and Figure 6.4(c)).

Figure 6.5 and Figure 6.6 depict the boxplots of the estimated optimal cut points over 500 simulated data sets, for each of the proposed algorithm, c-index and CPE estimator, 20% and 70% censoring rates and a sample size of $N = 500$ for Scenarios VII and VIII, where two optimal cut points are sought for a linear and nonlinear risk relationship between the continuous predictor X and the outcome respectively. Numerical results are reported in Table 6.8 for Scenario VII and Table 6.9 for Scenario VIII, respectively.

Simulation results for $k = 2$ cut points showed that the theoretical optimal cut points were estimated more accurately when the relationship between the continuous predictor X and the response variable was not linear (Scenario VIII) than when it was linear (Scenario VII) (see Figures 6.5 and 6.6). When the *AddFor* algorithm was used, smaller bias and MSE values were obtained with the CPE index for low censoring rates (20%), while the c-index performed better for high censoring rates (70%). For censoring rates around 50% no differences were observed between both discrimination indexes. However, when the *Genetic* method was used, estimated cut points had smaller bias and MSE than the ones obtained with the *AddFor* algorithm, especially in Scenario VII where the relationship between the continuous predictor and the response variable was linear. In fact, in this scenario and for a sample size of 500, 70% censoring rate and CPE index, MSE of 0.164 and 0.32 were obtained when the *Genetic* and *AddFor* algorithms were used respectively.

Figure 6.7 and Figure 6.8 depict the boxplots of the estimated optimal cut points over 500 simulated data sets, for each of the proposed algorithms, c-index and CPE estimator, 20% and 70% censoring rates and a sample size of $N = 500$ for Scenarios IX and X, where three optimal cut points are sought for a linear and nonlinear risk relationship between the continuous predictor X and the outcome, respectively. Numerical results are reported in Table 6.10 for Scenario IX and Table 6.11 for Scenario X, respectively.

Similar to what we observed for $k = 2$, simulation results for $k = 3$ suggested that the estimation of the optimal cut points was not very accurate when the relationship between the continuous predictor variable X and the response variable was linear (Scenario IX), particularly when the censoring rate was around 50% or above. As can be observed in Table 6.10, for high censoring rates, estimation of the second and

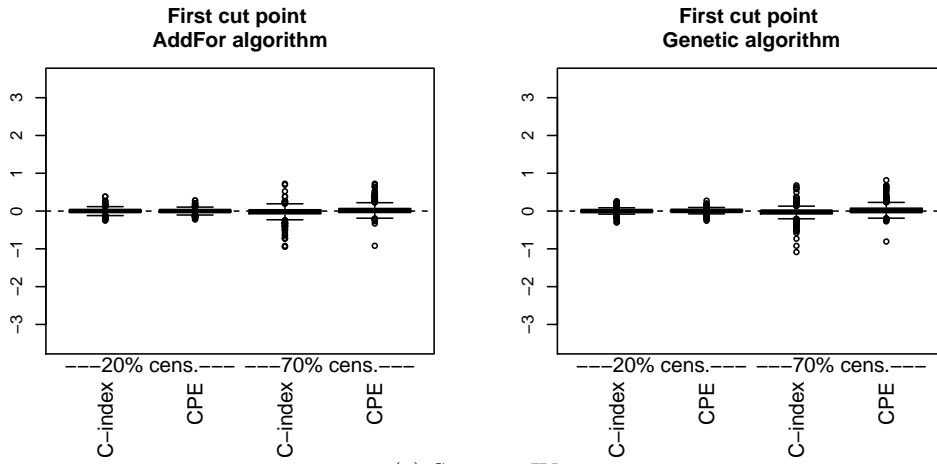
third cut points presents a high bias. However, in Scenario X where the relationship is not linear, the method performs satisfactorily (see Figure 6.8). Additionally, for high censoring rates, the third cut point bias is smaller when the c-index is used rather than when the CPE is used, for both algorithms *AddFor* or *Genetic* (see Table 6.10).

Table 6.5: Simulation results when one theoretical optimal cut point 0 was chosen with a decreasing risk relationship with the outcome $(\gamma_0, \lambda_0) = (1, 1)$ and $(\gamma_1, \lambda_1) = (1, 3)$ and censorship of 20%, 50% and 70% (Scenario IV). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

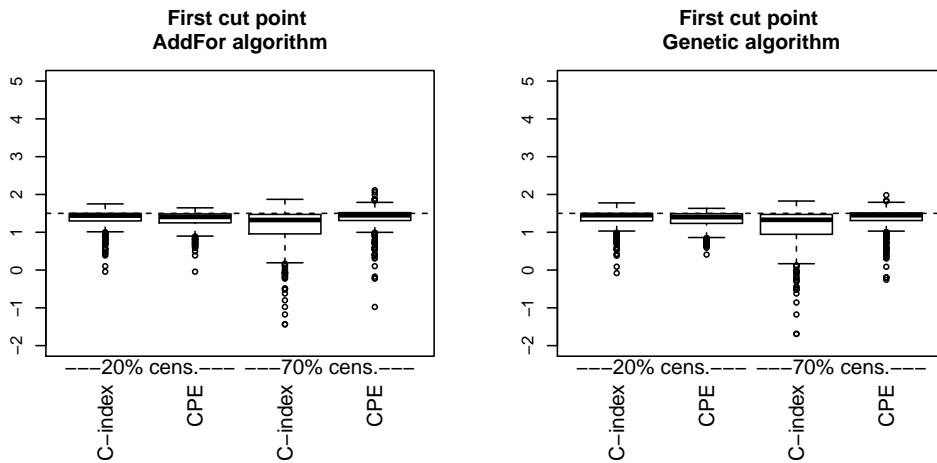
Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	0	-0.001 (0.060)	0.000	-0.001	0.004	0.000 (0.072)	0.001	0.000	0.005
	50%	0	0.008 (0.085)	0.001	0.008	0.007	-0.011 (0.100)	-0.005	-0.011	0.010
	70%	0	0.032 (0.141)	0.008	0.032	0.021	-0.037 (0.161)	-0.015	-0.037	0.027
Addfor 1000	20%	0	0.005 (0.051)	0.000	0.005	0.003	-0.005 (0.062)	-0.004	-0.005	0.004
	50%	0	0.013 (0.092)	0.000	0.013	0.009	-0.017 (0.095)	-0.011	-0.017	0.009
	70%	0	0.029 (0.143)	0.000	0.029	0.021	-0.037 (0.150)	-0.021	-0.037	0.024
Genetic	20%	0	0.011 (0.052)	0.005	0.011	0.003	0.001 (0.063)	0.000	0.001	0.004
	50%	0	0.017 (0.092)	0.004	0.017	0.009	-0.014 (0.095)	-0.008	-0.014	0.009
	70%	0	0.033 (0.141)	0.000	0.033	0.021	-0.036 (0.151)	-0.016	-0.036	0.024
Sample Size N = 1000										
Addfor 100	20%	0	-0.005 (0.033)	-0.005	-0.005	0.001	0.003 (0.041)	0.002	0.003	0.002
	50%	0	0.009 (0.056)	0.003	0.009	0.003	-0.005 (0.057)	-0.004	-0.005	0.003
	70%	0	0.021 (0.076)	0.008	0.021	0.006	-0.011 (0.077)	-0.009	-0.011	0.006
Addfor 1000	20%	0	0.002 (0.024)	0.000	0.002	0.001	0.000 (0.041)	-0.002	0.000	0.002
	50%	0	0.006 (0.050)	0.000	0.006	0.003	-0.010 (0.051)	-0.007	-0.010	0.003
	70%	0	0.013 (0.074)	-0.002	0.013	0.006	-0.022 (0.073)	-0.013	-0.022	0.006
Genetic	20%	0	0.003 (0.027)	0.002	0.003	0.001	0.002 (0.042)	-0.001	0.002	0.002
	50%	0	0.008 (0.050)	0.001	0.008	0.003	-0.009 (0.052)	-0.006	-0.009	0.003
	70%	0	0.012 (0.076)	0.000	0.012	0.006	-0.021 (0.080)	-0.013	-0.021	0.007

Table 6.12 shows the results obtained with the *AddFor* algorithm ($M = 100$), a sample size of $N = 1000$ and 500 replicates for the bias correction of the CPE and c-index concordance probability estimators. The theoretical concordance probability has been calculated empirically for the theoretical categorical variable in each scenario over 1000 replicates for a sample size of $N = 10,000$.

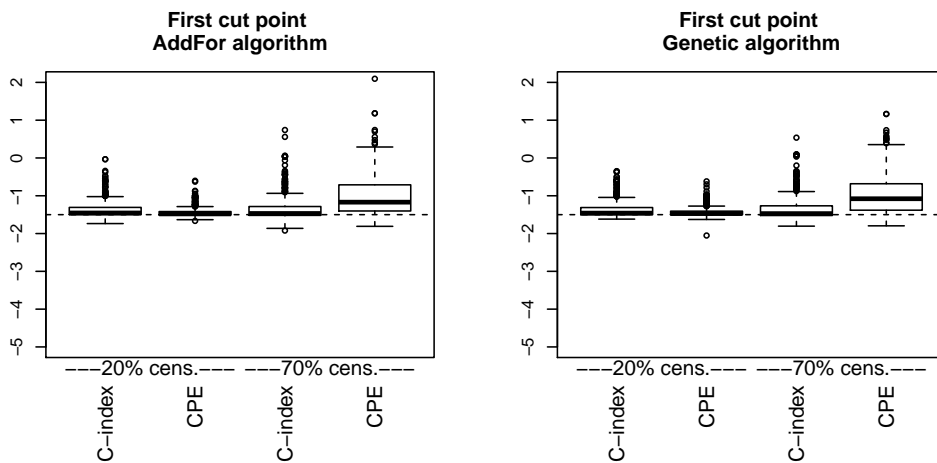
The results corroborate the fact that the CPE is an unbiased estimator, where slight differences can be observed between the estimated CPE and bias-corrected CPE. However, for the c-index, different results are obtained depending on the



(a) Scenario IV



(b) Scenario V



(c) Scenario VI

Figure 6.4: Boxplot of the estimated optimal cut points based on 500 simulated data sets, $N = 500$ sample size, one theoretical cut point and decreasing risk relationship with the outcome. Results are shown for *AddFor* ($M = 100$) and *Genetic* algorithms, censoring rates of 20% and 70% and c-index and CPE discriminative ability estimators. From top to bottom: (a) theoretical cut points (0); (b) theoretical cut points (1.5); and (c) theoretical cut points (-1.5).

Table 6.6: Simulation results when one theoretical optimal cut point 1.5 was chosen with a decreasing risk relationship with the outcome $(\gamma_0, \lambda_0) = (1, 1)$ and $(\gamma_1, \lambda_1) = (1, 3)$ and censorship of 20%, 50% and 70% (Scenario V). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	1.5	1.327 (0.227)	1.407	-0.173	0.081	1.352 (0.246)	1.441	-0.148	0.082
	50%	1.5	1.366 (0.262)	1.455	-0.134	0.087	1.267 (0.368)	1.403	-0.233	0.189
	70%	1.5	1.368 (0.314)	1.449	-0.132	0.116	1.143 (0.504)	1.325	-0.357	0.381
Addfor 1000	20%	1.5	1.316 (0.224)	1.394	-0.184	0.084	1.348 (0.238)	1.442	-0.152	0.079
	50%	1.5	1.356 (0.253)	1.442	-0.144	0.085	1.251 (0.364)	1.384	-0.249	0.195
	70%	1.5	1.355 (0.302)	1.449	-0.145	0.112	1.129 (0.510)	1.323	-0.371	0.397
Genetic	20%	1.5	1.322 (0.228)	1.401	-0.178	0.084	1.360 (0.237)	1.451	-0.140	0.076
	50%	1.5	1.361 (0.258)	1.448	-0.139	0.086	1.260 (0.357)	1.390	-0.240	0.185
	70%	1.5	1.358 (0.302)	1.451	-0.142	0.111	1.132 (0.513)	1.329	-0.368	0.398
Sample Size N = 1000										
Addfor 100	20%	1.5	1.383 (0.153)	1.442	-0.117	0.037	1.419 (0.133)	1.465	-0.081	0.024
	50%	1.5	1.422 (0.151)	1.470	-0.078	0.029	1.344 (0.253)	1.438	-0.156	0.088
	70%	1.5	1.424 (0.182)	1.475	-0.076	0.039	1.283 (0.322)	1.399	-0.217	0.150
Addfor 1000	20%	1.5	1.381 (0.163)	1.438	-0.119	0.041	1.415 (0.145)	1.475	-0.085	0.028
	50%	1.5	1.423 (0.145)	1.478	-0.077	0.027	1.344 (0.248)	1.437	-0.156	0.086
	70%	1.5	1.426 (0.169)	1.476	-0.074	0.034	1.266 (0.327)	1.390	-0.234	0.161
Genetic	20%	1.5	1.387 (0.158)	1.445	-0.113	0.038	1.417 (0.147)	1.479	-0.083	0.028
	50%	1.5	1.428 (0.143)	1.480	-0.072	0.026	1.339 (0.256)	1.432	-0.161	0.092
	70%	1.5	1.423 (0.180)	1.479	-0.077	0.038	1.266 (0.327)	1.391	-0.234	0.162

Table 6.7: Simulation results when one theoretical optimal cut point -1.5 was chosen with a decreasing risk relationship with the outcome $(\gamma_0, \lambda_0) = (1, 1)$ and $(\gamma_1, \lambda_1) = (1, 3)$ and censorship of 20%, 50% and 70% (Scenario VI). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	-1.5	-1.436 (0.135)	-1.480	0.064	0.022	-1.367 (0.232)	-1.455	0.133	0.071
	50%	-1.5	-1.282 (0.298)	-1.411	0.218	0.136	-1.368 (0.241)	-1.466	0.132	0.075
	70%	-1.5	-0.997 (0.524)	-1.170	0.503	0.526	-1.340 (0.339)	-1.469	0.160	0.140
Addfor 1000	20%	-1.5	-1.437 (0.134)	-1.480	0.063	0.022	-1.364 (0.255)	-1.467	0.136	0.083
	50%	-1.5	-1.271 (0.311)	-1.395	0.229	0.149	-1.366 (0.254)	-1.472	0.134	0.082
	70%	-1.5	-0.976 (0.512)	-1.087	0.524	0.536	-1.348 (0.334)	-1.480	0.152	0.134
Genetic	20%	-1.5	-1.429 (0.134)	-1.472	0.071	0.023	-1.361 (0.238)	-1.459	0.139	0.076
	50%	-1.5	-1.261 (0.315)	-1.397	0.239	0.156	-1.360 (0.253)	-1.468	0.140	0.083
	70%	-1.5	-0.968 (0.518)	-1.079	0.532	0.551	-1.343 (0.324)	-1.472	0.157	0.129
Sample Size N = 1000										
Addfor 100	20%	-1.5	-1.472 (0.068)	-1.487	0.028	0.005	-1.430 (0.128)	-1.467	0.070	0.021
	50%	-1.5	-1.391 (0.174)	-1.456	0.109	0.042	-1.422 (0.143)	-1.467	0.078	0.026
	70%	-1.5	-1.177 (0.332)	-1.278	0.323	0.214	-1.408 (0.174)	-1.465	0.092	0.039
Addfor 1000	20%	-1.5	-1.465 (0.063)	-1.487	0.035	0.005	-1.432 (0.126)	-1.478	0.068	0.021
	50%	-1.5	-1.384 (0.179)	-1.456	0.116	0.045	-1.426 (0.143)	-1.478	0.074	0.026
	70%	-1.5	-1.175 (0.330)	-1.276	0.325	0.215	-1.415 (0.173)	-1.478	0.085	0.037
Genetic	20%	-1.5	-1.463 (0.064)	-1.485	0.037	0.005	-1.423 (0.137)	-1.475	0.077	0.025
	50%	-1.5	-1.385 (0.176)	-1.457	0.115	0.044	-1.422 (0.150)	-1.476	0.078	0.028
	70%	-1.5	-1.172 (0.330)	-1.277	0.328	0.216	-1.417 (0.174)	-1.478	0.083	0.037

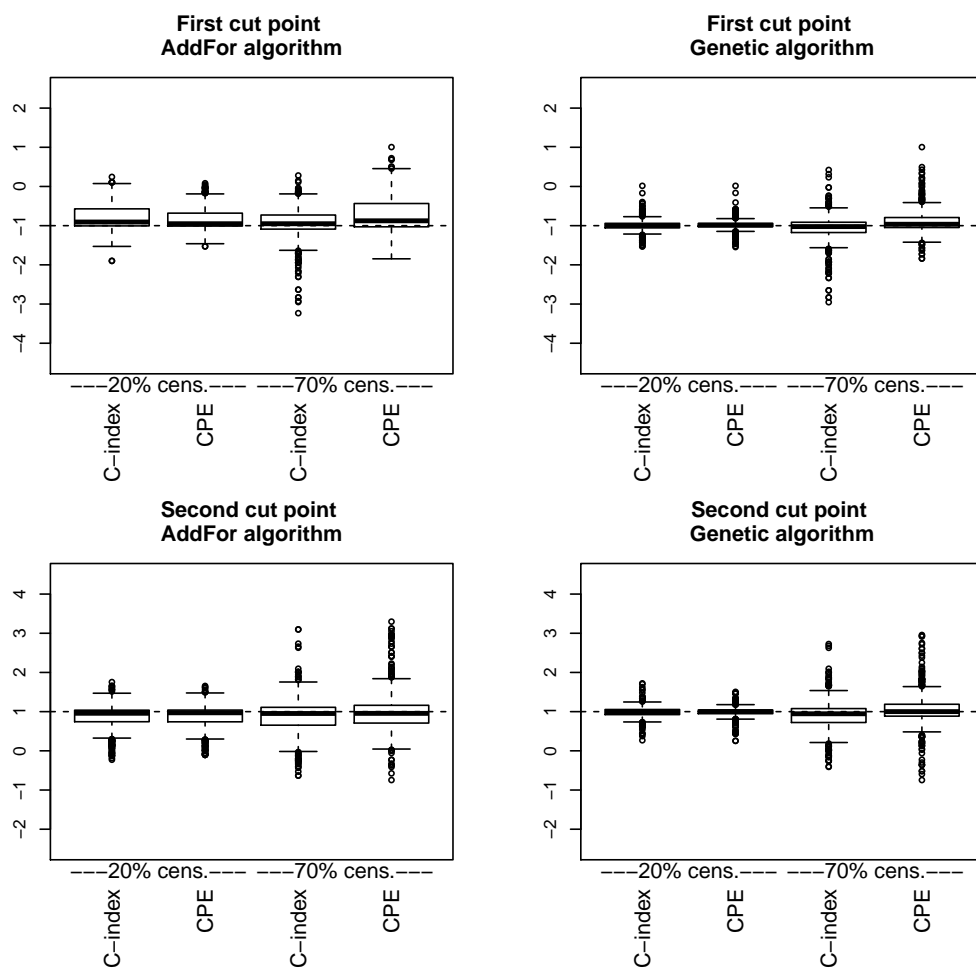


Figure 6.5: Boxplot of the estimated optimal cut points based on 500 simulated data sets, sample size $N = 500$ and Scenario VII - two theoretical optimal cut points (-1 and 1) and a linear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 1)$ and $(\gamma_2, \lambda_2) = (1, 2)$. Results are shown for *AddFor* ($M = 100$) and *Genetic* algorithms, censoring rates of 20% and 70% and c-index and CPE discriminative ability estimators.

Table 6.8: Simulation results when two theoretical optimal cut points -1 and 1 were chosen with a linear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 1)$ and $(\gamma_2, \lambda_2) = (1, 2)$ and censorship of 20%, 50% and 70% (Scenario VII). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	-1	-0.826 (0.303)	-0.951	0.174	0.127	-0.789 (0.344)	-0.907	0.211	0.152
		1	0.857 (0.335)	0.970	-0.143		0.856 (0.347)	0.958	-0.144	
	50%	-1	-0.782 (0.375)	-0.928	0.218	0.175	-0.834 (0.351)	-0.907	0.166	0.167
		1	0.894 (0.389)	0.960	-0.106		0.862 (0.407)	0.971	-0.138	
	70%	-1	-0.726 (0.467)	-0.876	0.274	0.327	-0.954 (0.451)	-0.951	0.046	0.251
		1	1.001 (0.602)	0.956	0.001		0.861 (0.527)	0.953	-0.139	
Addfor 1000	20%	-1	-0.827 (0.291)	-0.944	0.173	0.128	-0.772 (0.339)	-0.892	0.228	0.151
		1	0.848 (0.345)	0.979	-0.152		0.859 (0.339)	0.969	-0.141	
	50%	-1	-0.792 (0.372)	-0.945	0.208	0.172	-0.825 (0.355)	-0.890	0.175	0.178
		1	0.884 (0.386)	0.951	-0.116		0.868 (0.429)	0.964	-0.132	
	70%	-1	-0.737 (0.464)	-0.908	0.263	0.323	-0.945 (0.452)	-0.951	0.055	0.252
		1	0.958 (0.600)	0.941	-0.042		0.817 (0.514)	0.931	-0.183	
Genetic	20%	-1	-0.983 (0.146)	-0.987	0.017	0.021	-0.990 (0.169)	-1.000	0.010	0.027
		1	0.991 (0.147)	1.002	-0.009		0.992 (0.157)	1.000	-0.008	
	50%	-1	-0.953 (0.202)	-0.980	0.047	0.054	-1.012 (0.252)	-1.000	-0.012	0.073
		1	1.022 (0.255)	0.999	0.022		0.968 (0.286)	0.984	-0.032	
	70%	-1	-0.901 (0.318)	-0.964	0.099	0.164	-1.076 (0.381)	-1.024	-0.076	0.167
		1	1.055 (0.463)	1.000	0.055		0.902 (0.418)	0.946	-0.098	
Sample Size N = 1000										
Addfor 100	20%	-1	-0.855 (0.256)	-0.965	0.145	0.119	-0.800 (0.308)	-0.950	0.200	0.127
		1	0.814 (0.342)	0.972	-0.186		0.853 (0.312)	0.974	-0.147	
	50%	-1	-0.809 (0.323)	-0.958	0.191	0.150	-0.831 (0.273)	-0.923	0.169	0.122
		1	0.827 (0.359)	0.956	-0.173		0.862 (0.349)	0.982	-0.138	
	70%	-1	-0.759 (0.391)	-0.932	0.241	0.203	-0.896 (0.319)	-0.945	0.104	0.161
		1	0.931 (0.437)	0.972	-0.069		0.828 (0.424)	0.971	-0.172	
Addfor 1000	20%	-1	-0.839 (0.268)	-0.957	0.161	0.123	-0.804 (0.301)	-0.941	0.196	0.122
		1	0.827 (0.344)	0.986	-0.173		0.851 (0.306)	0.982	-0.149	
	50%	-1	-0.805 (0.320)	-0.947	0.195	0.144	-0.822 (0.288)	-0.921	0.178	0.123
		1	0.822 (0.342)	0.959	-0.178		0.863 (0.335)	0.979	-0.137	
	70%	-1	-0.762 (0.376)	-0.943	0.238	0.189	-0.895 (0.293)	-0.952	0.105	0.156
		1	0.909 (0.415)	0.968	-0.091		0.814 (0.425)	0.962	-0.186	
Genetic	20%	-1	-0.987 (0.064)	-0.993	0.013	0.006	-0.991 (0.081)	-0.997	0.009	0.008
		1	0.995 (0.084)	1.004	-0.005		0.995 (0.099)	1.001	-0.005	
	50%	-1	-0.981 (0.094)	-0.992	0.019	0.015	-1.001 (0.107)	-1.001	-0.001	0.022
		1	1.013 (0.141)	1.007	0.013		0.966 (0.180)	0.991	-0.034	
	70%	-1	-0.958 (0.157)	-0.983	0.042	0.054	-1.021 (0.166)	-1.007	-0.021	0.056
		1	1.035 (0.283)	1.011	0.035		0.922 (0.281)	0.981	-0.078	

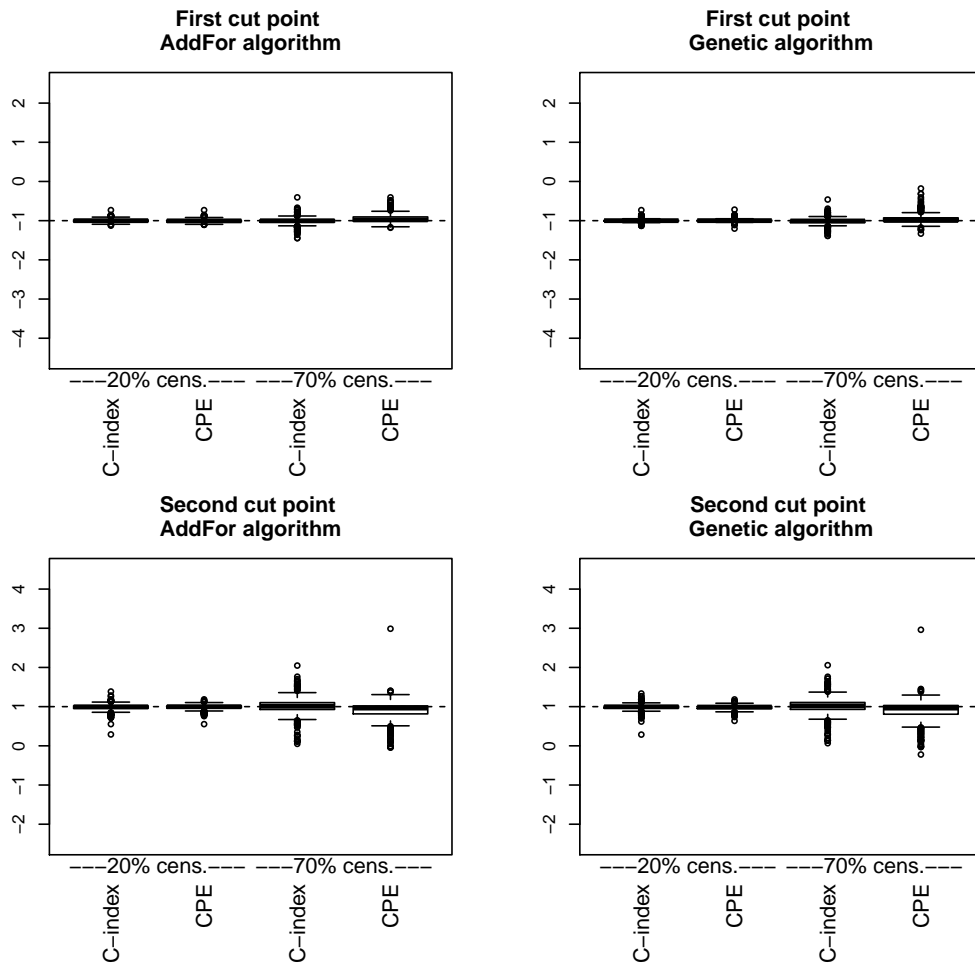


Figure 6.6: Boxplot of the estimated optimal cut points based on 500 simulated data sets, sample size $N = 500$ and Scenario VIII - two theoretical optimal cut points (-1 and 1) and a nonlinear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 3)$. Results are shown for *AddFor* ($M = 100$) and *Genetic* algorithms, censoring rates of 20% and 70% and c-index and CPE discriminative ability estimators.

Table 6.9: Simulation results when two theoretical optimal cut points -1 and 1 were chosen with a nonlinear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 3)$ and $(\gamma_2, \lambda_2) = (1, 1)$ and censorship of 20%, 50% and 70% (Scenario VIII). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	-1	-1.005 (0.039)	-1.006	-0.005	0.003	-1.000 (0.039)	-1.000	0.000	0.005
		1	0.988 (0.062)	0.998	-0.012		0.983 (0.085)	0.995	-0.017	
	50%	-1	-0.991 (0.051)	-0.994	0.009	0.014	-1.002 (0.049)	-1.000	-0.002	0.010
		1	0.941 (0.146)	0.982	-0.059		1.001 (0.134)	1.001	0.001	
	70%	-1	-0.947 (0.107)	-0.981	0.053	0.049	-1.011 (0.093)	-1.005	-0.011	0.034
		1	0.889 (0.267)	0.963	-0.111		1.009 (0.242)	1.011	0.009	
Addfor 1000	20%	-1	-1.001 (0.032)	-1.004	-0.001	0.003	-1.003 (0.039)	-1.004	-0.003	0.005
		1	0.972 (0.06)	0.986	-0.028		0.980 (0.085)	0.991	-0.020	
	50%	-1	-0.989 (0.056)	-0.999	0.011	0.017	-1.009 (0.051)	-1.007	-0.009	0.011
		1	0.924 (0.158)	0.972	-0.076		0.996 (0.137)	0.999	-0.004	
	70%	-1	-0.956 (0.113)	-0.994	0.044	0.052	-1.024 (0.090)	-1.012	-0.024	0.034
		1	0.880 (0.274)	0.957	-0.120		1.008 (0.242)	1.015	0.008	
Genetic	20%	-1	-0.999 (0.033)	-1.000	0.001	0.003	-0.999 (0.037)	-1.000	0.001	0.004
		1	0.975 (0.062)	0.989	-0.025		0.985 (0.080)	0.995	-0.015	
	50%	-1	-0.990 (0.052)	-0.996	0.010	0.017	-1.005 (0.053)	-1.002	-0.005	0.011
		1	0.928 (0.161)	0.979	-0.072		1.000 (0.139)	1.005	0.000	
	70%	-1	-0.958 (0.118)	-0.994	0.042	0.053	-1.015 (0.090)	-1.008	-0.015	0.034
		1	0.887 (0.277)	0.963	-0.113		1.013 (0.243)	1.019	0.013	
Sample Size N = 1000										
Addfor 100	20%	-1	-1.007 (0.025)	-1.007	-0.007	0.001	-0.998 (0.025)	-0.999	0.002	0.002
		1	0.997 (0.043)	1.001	-0.003		0.991 (0.051)	0.996	-0.009	
	50%	-1	-0.995 (0.031)	-0.998	0.005	0.003	-0.999 (0.032)	-0.999	0.001	0.003
		1	0.972 (0.070)	0.986	-0.028		0.995 (0.072)	0.996	-0.005	
	70%	-1	-0.977 (0.045)	-0.983	0.023	0.015	-1.005 (0.052)	-1.000	-0.005	0.010
		1	0.935 (0.150)	0.969	-0.065		1.008 (0.129)	0.997	0.008	
Addfor 1000	20%	-1	-1.001 (0.013)	-1.001	-0.001	0.001	-1.002 (0.017)	-1.002	-0.002	0.001
		1	0.988 (0.036)	0.994	-0.012		0.991 (0.042)	0.997	-0.009	
	50%	-1	-0.996 (0.024)	-1.000	0.004	0.003	-1.003 (0.026)	-1.003	-0.003	0.003
		1	0.970 (0.073)	0.992	-0.030		1.001 (0.066)	1.000	0.001	
	70%	-1	-0.984 (0.046)	-0.997	0.016	0.016	-1.014 (0.050)	-1.006	-0.014	0.009
		1	0.927 (0.159)	0.976	-0.073		1.014 (0.126)	1.004	0.014	
Genetic	20%	-1	-0.999 (0.013)	-1.000	0.001	0.001	-1.001 (0.018)	-1.000	-0.001	0.001
		1	0.988 (0.037)	0.994	-0.012		0.993 (0.044)	0.999	-0.007	
	50%	-1	-0.997 (0.023)	-0.999	0.003	0.003	-1.002 (0.030)	-1.001	-0.002	0.003
		1	0.972 (0.074)	0.994	-0.028		1.002 (0.071)	1.002	0.002	
	70%	-1	-0.989 (0.041)	-0.999	0.011	0.016	-1.013 (0.051)	-1.005	-0.013	0.009
		1	0.930 (0.157)	0.980	-0.070		1.016 (0.125)	1.007	0.016	

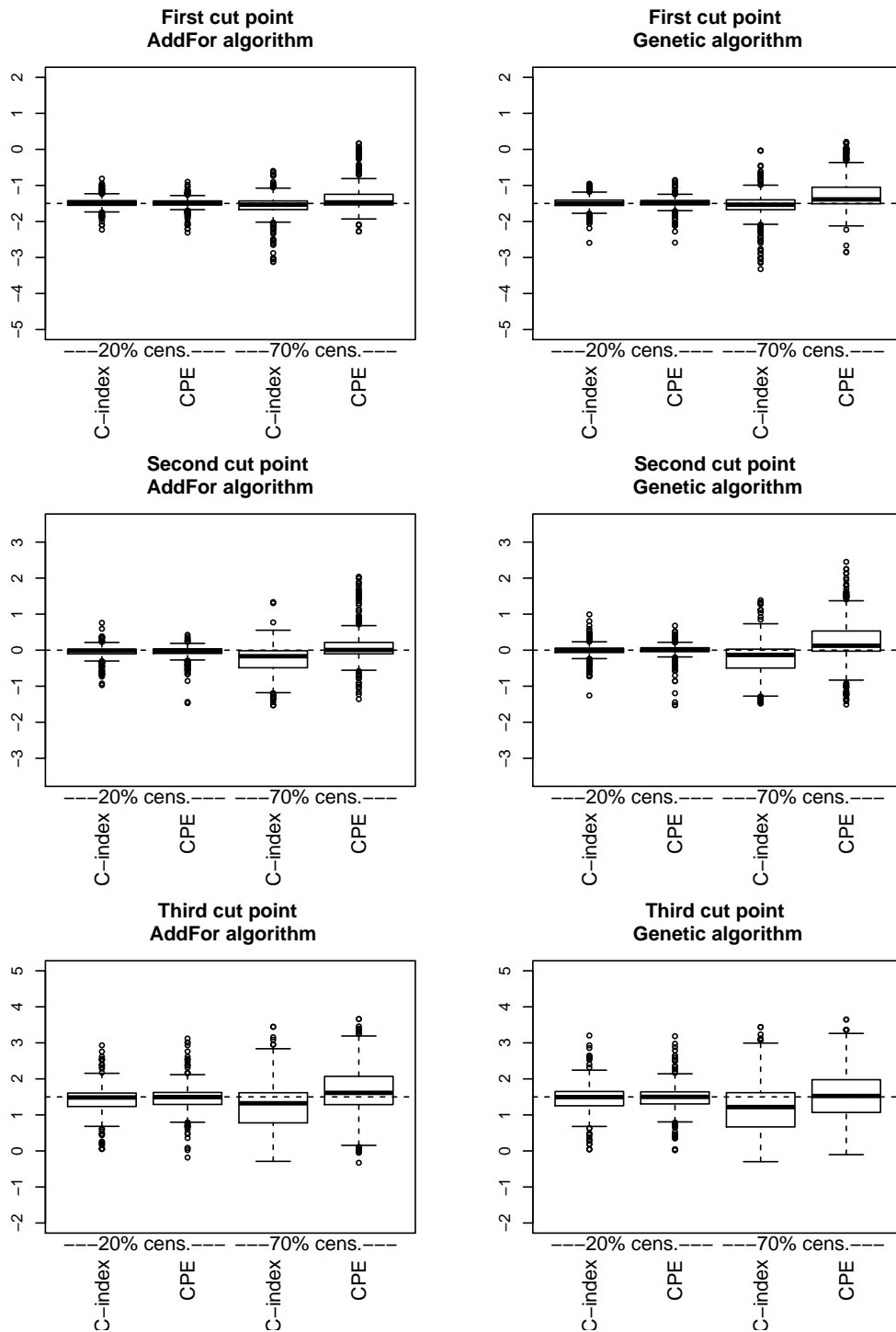


Figure 6.7: Boxplot of the estimated optimal cut points based on 500 simulated data sets, sample size $N = 500$ and Scenario IX - three theoretical optimal cut points $(-1.5, 0$ and $1)$ and a linear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 1), (\gamma_2, \lambda_2) = (1, 2)$ and $(\gamma_3, \lambda_3) = (1, 3)$. Results are shown for *AddFor* ($M = 100$) and *Genetic* algorithms, censoring rates of 20% and 70% and c-index and CPE discriminative ability estimators.

Table 6.10: Simulation results when three theoretical optimal cut points $-1.5, 0$ and 1 were chosen with a linear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 1), (\gamma_2, \lambda_2) = (1, 2)$ and $(\gamma_3, \lambda_3) = (1, 3)$ and censorship of 20%, 50% and 70% (Scenario IX). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	-1.5	-1.496 (0.151)	-1.498	0.004	0.068	-1.490 (0.164)	-1.500	0.010	0.076
		0	-0.053 (0.187)	-0.014	-0.053		-0.049 (0.182)	-0.018	-0.049	
		1.5	1.458 (0.377)	1.494	-0.042		1.415 (0.400)	1.484	-0.085	
	50%	-1.5	-1.477 (0.185)	-1.494	0.023	0.120	-1.533 (0.264)	-1.510	-0.033	0.198
		0	-0.013 (0.213)	-0.003	-0.013		-0.169 (0.297)	-0.098	-0.169	
		1.5	1.598 (0.521)	1.539	0.098		1.332 (0.615)	1.398	-0.168	
	70%	-1.5	-1.323 (0.448)	-1.472	0.177	0.323	-1.576 (0.337)	-1.533	-0.076	0.306
		0	0.109 (0.479)	0.005	0.109		-0.281 (0.404)	-0.168	-0.281	
		1.5	1.657 (0.688)	1.612	0.157		1.212 (0.690)	1.323	-0.288	
Addfor 1000	20%	-1.5	-1.480 (0.141)	-1.481	0.020	0.062	-1.477 (0.168)	-1.492	0.023	0.076
		0	-0.043 (0.171)	-0.010	-0.043		-0.047 (0.185)	-0.018	-0.047	
		1.5	1.439 (0.362)	1.481	-0.061		1.436 (0.399)	1.478	-0.064	
	50%	-1.5	-1.448 (0.229)	-1.476	0.052	0.142	-1.534 (0.294)	-1.510	-0.034	0.218
		0	-0.004 (0.266)	-0.002	-0.004		-0.189 (0.33)	-0.098	-0.189	
		1.5	1.546 (0.548)	1.510	0.046		1.282 (0.612)	1.376	-0.218	
	70%	-1.5	-1.288 (0.472)	-1.456	0.212	0.341	-1.563 (0.339)	-1.534	-0.063	0.353
		0	0.058 (0.474)	0.003	0.058		-0.300 (0.423)	-0.195	-0.300	
		1.5	1.566 (0.725)	1.554	0.066		1.096 (0.714)	1.199	-0.404	
Genetic	20%	-1.5	-1.487 (0.160)	-1.490	0.013	0.071	-1.491 (0.170)	-1.499	0.009	0.077
		0	-0.009 (0.204)	0.008	-0.009		-0.009 (0.193)	-0.001	-0.009	
		1.5	1.465 (0.380)	1.501	-0.035		1.454 (0.403)	1.494	-0.046	
	50%	-1.5	-1.459 (0.224)	-1.486	0.041	0.159	-1.522 (0.279)	-1.507	-0.022	0.190
		0	0.071 (0.287)	0.024	0.071		-0.101 (0.326)	-0.019	-0.101	
		1.5	1.604 (0.572)	1.553	0.104		1.384 (0.603)	1.460	-0.116	
	70%	-1.5	-1.226 (0.488)	-1.388	0.274	0.444	-1.570 (0.395)	-1.535	-0.070	0.366
		0	0.246 (0.625)	0.123	0.246		-0.230 (0.502)	-0.133	-0.230	
		1.5	1.548 (0.753)	1.526	0.048		1.165 (0.722)	1.218	-0.335	
Sample Size N = 1000										
Addfor 100	20%	-1.5	-1.488 (0.079)	-1.494	0.012	0.028	-1.479 (0.100)	-1.488	0.021	0.030
		0	-0.038 (0.133)	-0.020	-0.038		-0.038 (0.110)	-0.015	-0.038	
		1.5	1.489 (0.243)	1.497	-0.011		1.460 (0.252)	1.485	-0.040	
	50%	-1.5	-1.477 (0.095)	-1.482	0.023	0.052	-1.500 (0.119)	-1.501	0.000	0.100
		0	-0.008 (0.114)	-0.002	-0.008		-0.116 (0.238)	-0.038	-0.116	
		1.5	1.494 (0.364)	1.490	-0.006		1.384 (0.449)	1.457	-0.116	
	70%	-1.5	-1.468 (0.177)	-1.480	0.032	0.155	-1.551 (0.241)	-1.512	-0.051	0.208
		0	0.006 (0.209)	0.004	0.006		-0.219 (0.347)	-0.100	-0.219	
		1.5	1.579 (0.619)	1.502	0.079		1.232 (0.569)	1.327	-0.268	
Addfor 1000	20%	-1.5	-1.481 (0.082)	-1.489	0.019	0.019	-1.482 (0.101)	-1.494	0.018	0.029
		0	-0.024 (0.086)	-0.008	-0.024		-0.032 (0.110)	-0.011	-0.032	
		1.5	1.499 (0.204)	1.499	-0.001		1.455 (0.246)	1.477	-0.045	
	50%	-1.5	-1.476 (0.105)	-1.487	0.024	0.067	-1.510 (0.154)	-1.498	-0.010	0.113
		0	-0.020 (0.159)	-0.004	-0.020		-0.129 (0.240)	-0.043	-0.129	
		1.5	1.494 (0.404)	1.482	-0.006		1.322 (0.459)	1.418	-0.178	
	70%	-1.5	-1.464 (0.185)	-1.486	0.036	0.144	-1.539 (0.214)	-1.505	-0.039	0.210
		0	-0.013 (0.203)	-0.003	-0.013		-0.233 (0.338)	-0.131	-0.233	
		1.5	1.479 (0.595)	1.458	-0.021		1.212 (0.576)	1.331	-0.288	
Genetic	20%	-1.5	-1.482 (0.083)	-1.492	0.018	0.023	-1.483 (0.102)	-1.494	0.017	0.03
		0	-0.003 (0.120)	0.001	-0.003		0.000 (0.083)	-0.001	0.000	
		1.5	1.503 (0.220)	1.502	0.003		1.479 (0.267)	1.492	-0.021	
	50%	-1.5	-1.478 (0.103)	-1.488	0.022	0.071	-1.509 (0.172)	-1.497	-0.009	0.092
		0	0.010 (0.117)	0.002	0.010		-0.037 (0.213)	-0.003	-0.037	
		1.5	1.497 (0.434)	1.489	-0.003		1.435 (0.444)	1.470	-0.065	
	70%	-1.5	-1.463 (0.205)	-1.484	0.037	0.168	-1.562 (0.248)	-1.510	-0.062	0.224
		0	0.042 (0.303)	0.009	0.042		-0.154 (0.374)	-0.022	-0.154	
		1.5	1.476 (0.607)	1.462	-0.024		1.325 (0.642)	1.384	-0.175	

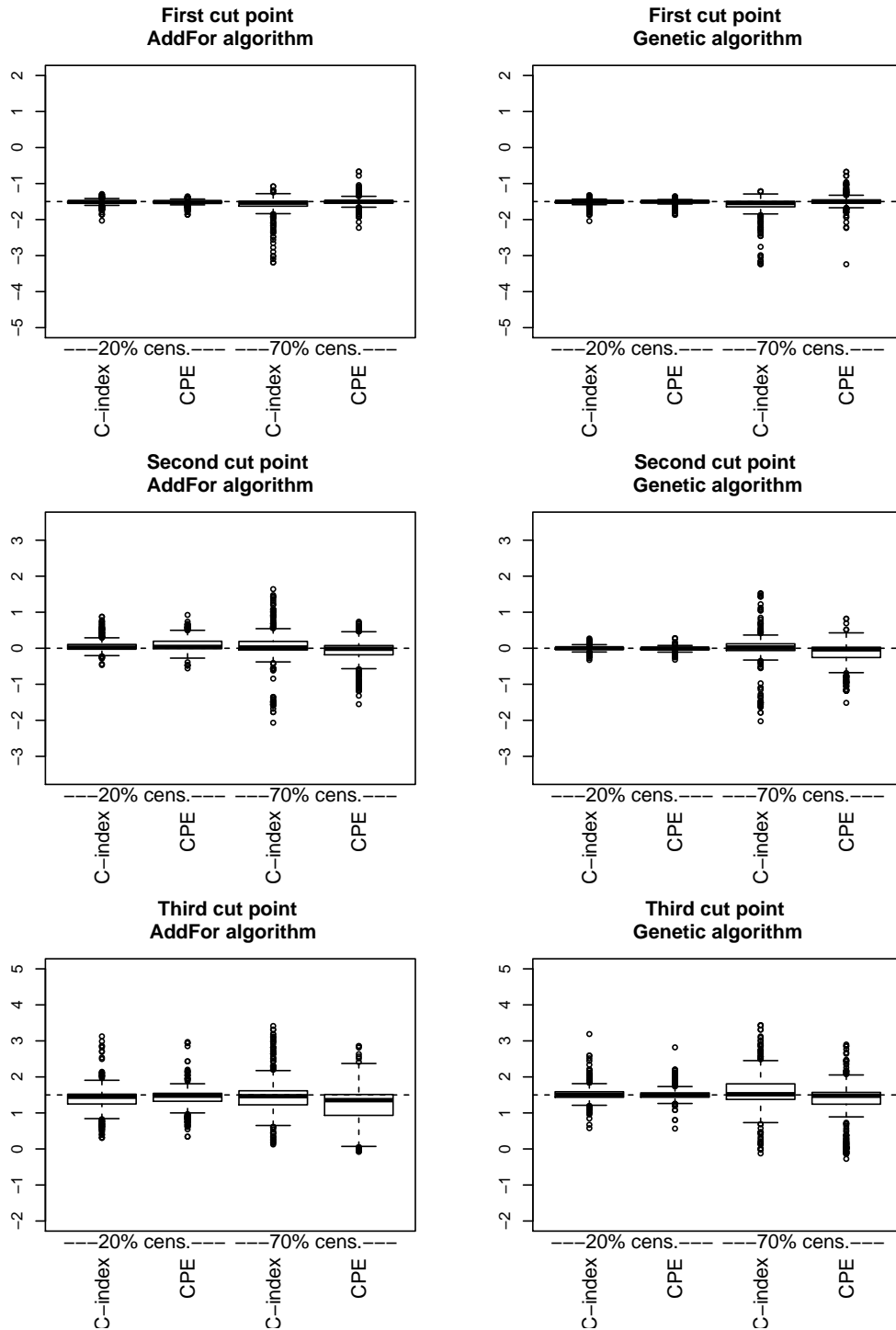


Figure 6.8: Boxplot of the estimated optimal cut points based on 500 simulated data sets, sample size $N = 500$ and Scenario IX - three theoretical optimal cut points $(-1.5, 0$ and $1)$ and a nonlinear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 3), (\gamma_2, \lambda_2) = (1, 1)$ and $(\gamma_3, \lambda_3) = (1, 0.5)$. Results are shown for *AddFor* ($M = 100$) and *Genetic* algorithms, censoring rates of 20% and 70% and C-index and CPE discriminative ability estimators.

Table 6.11: Simulation results when three theoretical optimal cut points $-1.5, 0$ and 1 were chosen with a nonlinear relationship with the outcome $(\gamma_0, \lambda_0) = (1, 0.5), (\gamma_1, \lambda_1) = (1, 3), (\gamma_2, \lambda_2) = (1, 1)$ and $(\gamma_3, \lambda_3) = (1, 0.5)$ and censorship of 20%, 50% and 70% (Scenario X). Mean, standard deviation (sd), median (Me.), bias and mean squared error (MSE) for the estimated cut points are reported when CPE or c-index discriminative ability estimators are used as the maximisation criteria.

Method	Cens.	t.cut point	Cut point CPE Estimation				Cut point C-index Estimation			
			Mean (sd)	Me.	Bias	MSE	Mean (sd)	Me.	Bias	MSE
Sample Size N = 500										
Addfor 100	20%	-1.5	-1.520 (0.060)	-1.512	-0.020	0.051	-1.517 (0.073)	-1.509	-0.017	0.063
		0	0.094 (0.187)	0.036	0.094		0.071 (0.183)	0.015	0.071	
		1.5	1.423 (0.318)	1.480	-0.077		1.383 (0.364)	1.450	-0.117	
	50%	-1.5	-1.513 (0.07)	-1.507	-0.013	0.071	-1.536 (0.114)	-1.516	-0.036	0.077
		0	0.027 (0.212)	0.009	0.027		0.082 (0.258)	0.021	0.082	
		1.5	1.385 (0.385)	1.441	-0.115		1.425 (0.373)	1.445	-0.075	
	70%	-1.5	-1.503 (0.132)	-1.502	-0.003	0.173	-1.597 (0.257)	-1.530	-0.097	0.211
		0	-0.097 (0.360)	-0.014	-0.097		0.033 (0.527)	0.019	0.033	
		1.5	1.212 (0.528)	1.355	-0.288		1.488 (0.529)	1.465	-0.012	
Addfor 1000	20%	-1.5	-1.511 (0.060)	-1.500	-0.011	0.048	-1.517 (0.067)	-1.506	-0.017	0.065
		0	0.086 (0.186)	0.016	0.086		0.068 (0.189)	0.010	0.068	
		1.5	1.428 (0.308)	1.466	-0.072		1.357 (0.362)	1.437	-0.143	
	50%	-1.5	-1.507 (0.065)	-1.503	-0.007	0.082	-1.537 (0.118)	-1.514	-0.037	0.088
		0	0.009 (0.227)	0.005	0.009		0.087 (0.264)	0.028	0.087	
		1.5	1.343 (0.409)	1.433	-0.157		1.411 (0.407)	1.442	-0.089	
	70%	-1.5	-1.506 (0.133)	-1.507	-0.006	0.203	-1.615 (0.265)	-1.538	-0.115	0.226
		0	-0.119 (0.385)	-0.012	-0.119		0.049 (0.535)	0.045	0.049	
		1.5	1.165 (0.563)	1.301	-0.335		1.467 (0.551)	1.461	-0.033	
Genetic	20%	-1.5	-1.513 (0.059)	-1.503	-0.013	0.014	-1.516 (0.068)	-1.505	-0.016	0.021
		0	-0.016 (0.059)	-0.007	-0.016		-0.004 (0.068)	0.000	-0.004	
		1.5	1.517 (0.185)	1.496	0.017		1.529 (0.231)	1.507	0.029	
	50%	-1.5	-1.505 (0.069)	-1.503	-0.005	0.038	-1.536 (0.102)	-1.515	-0.036	0.053
		0	-0.036 (0.139)	-0.004	-0.036		0.021 (0.167)	0.008	0.021	
		1.5	1.535 (0.296)	1.504	0.035		1.574 (0.337)	1.514	0.074	
	70%	-1.5	-1.493 (0.17)	-1.503	0.007	0.172	-1.617 (0.286)	-1.538	-0.117	0.209
		0	-0.147 (0.327)	-0.030	-0.147		-0.031 (0.480)	0.020	-0.031	
		1.5	1.318 (0.571)	1.477	-0.182		1.606 (0.539)	1.519	0.106	
Sample Size N = 1000										
Addfor 100	20%	-1.5	-1.518 (0.032)	-1.517	-0.018	0.043	-1.507 (0.039)	-1.510	-0.007	0.041
		0	0.118 (0.186)	0.040	0.118		0.073 (0.159)	0.017	0.073	
		1.5	1.447 (0.276)	1.504	-0.053		1.381 (0.277)	1.479	-0.119	
	50%	-1.5	-1.509 (0.037)	-1.509	-0.009	0.059	-1.516 (0.085)	-1.509	-0.016	0.056
		0	0.025 (0.173)	0.007	0.025		0.076 (0.207)	0.015	0.076	
		1.5	1.338 (0.346)	1.468	-0.162		1.404 (0.320)	1.463	-0.096	
	70%	-1.5	-1.513 (0.124)	-1.504	-0.013	0.078	-1.539 (0.165)	-1.511	-0.039	0.098
		0	0.002 (0.245)	0.006	0.002		0.090 (0.329)	0.016	0.090	
		1.5	1.317 (0.354)	1.432	-0.183		1.456 (0.382)	1.464	-0.044	
Addfor 1000	20%	-1.5	-1.505 (0.027)	-1.501	-0.005	0.041	-1.506 (0.030)	-1.502	-0.006	0.04
		0	0.104 (0.177)	0.021	0.104		0.070 (0.158)	0.009	0.070	
		1.5	1.429 (0.275)	1.491	-0.071		1.390 (0.277)	1.477	-0.110	
	50%	-1.5	-1.506 (0.032)	-1.502	-0.006	0.056	-1.518 (0.100)	-1.504	-0.018	0.059
		0	0.022 (0.173)	0.008	0.022		0.076 (0.234)	0.019	0.076	
		1.5	1.352 (0.337)	1.476	-0.148		1.416 (0.316)	1.469	-0.084	
	70%	-1.5	-1.518 (0.119)	-1.504	-0.018	0.084	-1.548 (0.153)	-1.512	-0.048	0.092
		0	-0.011 (0.249)	0.005	-0.011		0.101 (0.324)	0.029	0.101	
		1.5	1.313 (0.376)	1.425	-0.187		1.465 (0.368)	1.477	-0.035	
Genetic	20%	-1.5	-1.505 (0.026)	-1.501	-0.005	0.006	-1.505 (0.025)	-1.501	-0.005	0.007
		0	-0.004 (0.033)	-0.002	-0.004		0.002 (0.037)	0.001	0.002	
		1.5	1.517 (0.122)	1.501	0.017		1.523 (0.138)	1.504	0.023	
	50%	-1.5	-1.504 (0.031)	-1.501	-0.004	0.011	-1.512 (0.045)	-1.506	-0.012	0.015
		0	-0.018 (0.071)	-0.003	-0.018		0.016 (0.071)	0.007	0.016	
		1.5	1.519 (0.159)	1.503	0.019		1.549 (0.188)	1.510	0.049	
	70%	-1.5	-1.500 (0.053)	-1.502	0.000	0.028	1.541 (0.140)	-1.512	-0.041	0.050
		0	-0.048 (0.144)	-0.010	-0.048		0.026 (0.200)	0.018	0.026	
		1.5	1.510 (0.240)	1.501	0.010		1.593 (0.285)	1.529	0.093	

simulated scenarios. For example, in Scenario II and Scenario VI where the c-index performs successfully when it comes to search the optimal cut point (see Table 6.3 and Table 6.7), we observe that the c-index bias increases as censorship increases, but the bias correction procedure corrects this bias, although for a 70% censoring rate it is still 0.014 and 0.013 in Scenarios II and VI, respectively. However, in Scenarios III and IV where we are not able to find the location of the optimal cut point by the maximisation of the c-index, the bias correction procedure is not correcting this bias; on the contrary, the bias gets larger. In our opinion, the reason for this might be that, since we are not able to select the optimal cut points, the categorisation for which we are estimating the c-index is not optimal, and hence, this estimation cannot be compared with the theoretical concordance probability. Similar conclusions are obtained for $k = 2$ and $k = 3$ cut points.

Finally, the *AddFor* algorithm was applied in the scenario in which the proportional hazards assumption does not hold for $k = 2$ number of cut points. No differences were observed in estimated optimal cut points with regard to the comparable scenario for proportional hazards. Data for these simulation results are not shown because no differences were observed with the proportional hazards scenario and the non-proportional hazard was out of the scope of the work presented in this chapter.

Table 6.12: Simulation results obtained when the *AddFor* algorithm ($M = 100$) and $R = 500$ replicates were performed for a sample size of $N = 1000$. Obtained mean (standard deviation) and bias of the estimated c-index and CPE are reported together with the mean (standard deviation) and bias for the bias-corrected c-index and CPE estimators.

Scenario	Cens.	Theoretical concordance probability	Estimated C-index			Bias-corrected C-index			Estimated CPE			Bias-corrected CPE		
			mean (sd)	bias	mean (sd)	mean (sd)	bias	mean (sd)	mean (sd)	bias	mean (sd)	mean (sd)	bias	
I	20%		0.627 (0.009)	0.002	0.621 (0.009)	-0.004	0.624 (0.007)	-0.001	0.624 (0.007)	-0.001	0.624 (0.007)	-0.001		
	50%	0.625	0.630 (0.011)	0.005	0.621 (0.012)	-0.004	0.625 (0.009)	0.000	0.625 (0.009)	0.000	0.625 (0.009)	0.000		
	70%		0.629 (0.014)	0.004	0.618 (0.014)	-0.007	0.625 (0.011)	0.000	0.625 (0.011)	0.000	0.625 (0.011)	0.000		
II	20%		0.595 (0.008)	0.007	0.588 (0.008)	0.001	0.588 (0.007)	0.000	0.585 (0.007)	0.000	0.585 (0.007)	-0.002		
	50%	0.588	0.605 (0.010)	0.017	0.596 (0.010)	0.009	0.589 (0.007)	0.001	0.586 (0.007)	-0.001	0.586 (0.007)	-0.001		
	70%		0.614 (0.014)	0.026	0.602 (0.014)	0.014	0.591 (0.010)	0.004	0.589 (0.010)	0.004	0.589 (0.010)	0.002		
III	20%		0.585 (0.009)	-0.002	0.579 (0.009)	-0.009	0.588 (0.008)	0.000	0.585 (0.008)	0.000	0.585 (0.008)	-0.002		
	50%	0.588	0.580 (0.010)	-0.007	0.573 (0.010)	-0.015	0.588 (0.010)	0.001	0.586 (0.010)	0.001	0.586 (0.010)	-0.002		
	70%		0.578 (0.011)	-0.010	0.569 (0.012)	-0.019	0.589 (0.013)	0.001	0.586 (0.013)	0.001	0.586 (0.013)	-0.002		
IV	20%		0.627 (0.009)	0.002	0.620 (0.009)	-0.004	0.625 (0.007)	0.000	0.624 (0.007)	0.000	0.624 (0.007)	0.000		
	50%	0.625	0.629 (0.011)	0.004	0.621 (0.012)	-0.004	0.625 (0.009)	0.000	0.624 (0.009)	0.000	0.624 (0.009)	0.000		
	70%		0.629 (0.014)	0.004	0.618 (0.015)	-0.007	0.625 (0.012)	0.000	0.625 (0.012)	0.000	0.625 (0.012)	0.000		
V	20%		0.585 (0.008)	-0.003	0.578 (0.008)	-0.009	0.589 (0.007)	0.001	0.586 (0.007)	0.001	0.586 (0.007)	-0.001		
	50%	0.587	0.580 (0.009)	-0.008	0.572 (0.010)	-0.015	0.588 (0.010)	0.001	0.585 (0.010)	0.001	0.585 (0.010)	-0.002		
	70%		0.577 (0.011)	-0.010	0.568 (0.011)	-0.019	0.588 (0.012)	0.000	0.585 (0.012)	0.000	0.585 (0.012)	-0.002		
VI	20%		0.594 (0.009)	0.007	0.588 (0.009)	0.001	0.588 (0.007)	0.001	0.585 (0.007)	0.001	0.585 (0.007)	-0.002		
	50%	0.587	0.604 (0.011)	0.017	0.596 (0.011)	0.009	0.589 (0.008)	0.001	0.586 (0.008)	0.001	0.586 (0.008)	-0.001		
	70%		0.612 (0.015)	0.025	0.601 (0.015)	0.013	0.591 (0.01)	0.003	0.589 (0.010)	0.003	0.589 (0.010)	0.001		
VII	20%		0.634 (0.010)	-0.002	0.626 (0.010)	-0.010	0.632 (0.009)	-0.004	0.631 (0.009)	-0.004	0.631 (0.009)	-0.005		
	50%	0.636	0.638 (0.012)	0.002	0.628 (0.013)	-0.007	0.633 (0.010)	-0.003	0.631 (0.010)	-0.003	0.631 (0.010)	-0.004		
	70%		0.640 (0.015)	0.005	0.628 (0.015)	-0.008	0.635 (0.013)	-0.001	0.633 (0.013)	-0.001	0.633 (0.013)	-0.002		
VIII	20%		0.678 (0.009)	0.003	0.671 (0.009)	-0.004	0.674 (0.008)	-0.001	0.673 (0.008)	-0.001	0.673 (0.008)	-0.002		
	50%	0.675	0.682 (0.012)	0.007	0.673 (0.012)	-0.002	0.674 (0.009)	-0.001	0.673 (0.009)	-0.001	0.673 (0.009)	-0.002		
	70%		0.682 (0.015)	0.006	0.670 (0.016)	-0.005	0.676 (0.013)	0.000	0.674 (0.013)	0.000	0.674 (0.013)	-0.001		
IX	20%		0.668 (0.010)	0.005	0.660 (0.010)	-0.002	0.664 (0.008)	0.001	0.662 (0.008)	0.001	0.662 (0.008)	0.000		
	50%	0.663	0.675 (0.013)	0.012	0.665 (0.013)	0.003	0.665 (0.011)	0.002	0.663 (0.011)	0.002	0.663 (0.011)	0.000		
	70%		0.678 (0.015)	0.016	0.666 (0.015)	0.004	0.666 (0.013)	0.004	0.665 (0.013)	0.004	0.665 (0.013)	0.002		
X	20%		0.667 (0.011)	-0.001	0.660 (0.011)	-0.008	0.666 (0.009)	-0.001	0.665 (0.009)	-0.001	0.665 (0.009)	-0.003		
	50%	0.668	0.665 (0.012)	-0.003	0.656 (0.012)	-0.012	0.667 (0.011)	-0.001	0.665 (0.011)	-0.001	0.665 (0.011)	-0.003		
	70%		0.663 (0.014)	-0.005	0.651 (0.015)	-0.017	0.669 (0.014)	0.001	0.667 (0.014)	0.001	0.667 (0.014)	-0.001		

6.3 Application to the Stable-COPD study

We applied the methodology proposed in Section 6.1 to the Stable-COPD study presented in Chapter 2, Section 2.2.

As we mentioned before, COPD is the third leading cause of death worldwide (Murray et al. 2013). Patients suffering from COPD have difficulties breathing, and hence, they have airflow limitation; therefore, spirometry is an important test to evaluate the disease. Classically, the severity of the disease has been graded by $FEV_{1\%}$ which represents the proportion of air that a person is able to expire in the first second of expiration. Although this is a continuous variable, in practice it is commonly categorised and used to classify patients in distinct severity groups. Recently, Almagro et al. (2014) proposed new thresholds to categorise $FEV_{1\%}$ into mild ($\geq 70\%$), moderate (56 – 69%), severe (36 – 55%) and very severe ($\leq 35\%$) categories and predict survival at 5 years in COPD patients. They compared this categorisation to the most common nowadays, such as, the Global Obstructive Lung Disease (GOLD) and ATS/ERS guidelines (Global Initiative for Chronic Obstructive Lung Disease 2013), and the old ATS standards proposal. The latter is the one used by the BODE index (Celli et al. 2004).

Clinical researchers involved in the Stable-COPD study presented us with two goals. First, the aim was to categorise the predictor variable $FEV_{1\%}$ into four categories (mild, moderate, severe and very severe), i.e., $k = 3$, in a univariate setting in order to compare the results obtained with previous categorisation proposals such as COCOMICS or GOLD. The second goal was to look for the best categorisation (location and number of cut points) in a multivariate setting, taking into account the effect of age and dyspnoea, which are seen as important predictors for the severity of sCOPD patients (Bestall et al. 1999).

Table 6.13 shows the results obtained in the univariate setting for $k = 3$ cut points with the CPE and c-index estimators. The same results were obtained with the *Genetic* and *AddFor* algorithms. However, the results obtained when the CPE estimator was used differed from the ones obtained with the c-index. The censoring rate in our data set was 66.6% and the relationship between the predictor $FEV_{1\%}$ and the response variable time until death in a 5 year follow-up was approximately linear, as can be seen in Figure 6.9. In view of the results obtained in the simulation study, we considered focusing on the results obtained with the c-index, since this appeared to perform better under this scenario.

The categorisation proposal obtained in the univariate setting with the *AddFor*

Table 6.13: Results obtained in the categorisation of the predictor variable $FEV_{1\%}$ of the Stable-COPD study in a univariate setting for $k = 3$ number of cut points.

Method	Concordance probability estimator	Estimated cut points	Concordance probability	
			Estimated	Bias-corrected
Addfor 100	c-index	36.45 ; 50.52 ; 64.58	0.620	0.605
Addfor 1000		36.98 ; 50.05 ; 64.07	0.620	0.604
Genetic		36.29 ; 50.75 ; 64.38	0.620	0.602
Addfor 100	CPE	52.27 ; 56.67 ; 64.58	0.611	0.609
Addfor 1000		52.05 ; 56.93 ; 64.07	0.611	0.609
Genetic		52.26 ; 56.25 ; 64.57	0.611	0.609

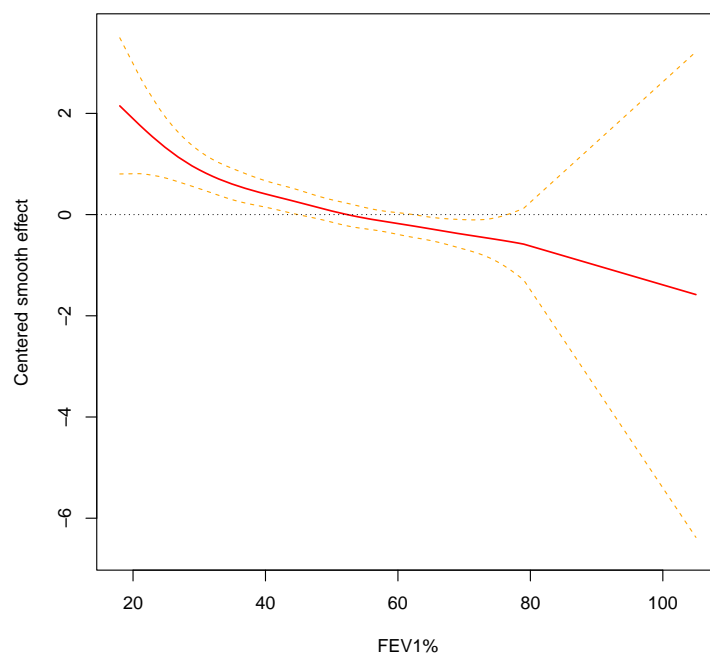


Figure 6.9: Estimated smooth relationship of the predictor variable $FEV_{1\%}$ with the response variable time until death in a univariate setting.

and *Genetic* algorithms, together with the GOLD, COCOMICS and BODE categorisation proposals is shown in Table 6.14. In addition, when applied to this data set,

the results we obtained improved the ones obtained with other categorisation proposals in terms of maximal estimated c-index. Nevertheless, the cut points obtained are very similar to the ones used in the BODE index.

Table 6.14: Results obtained in the categorisation of the predictor variable $FEV_{1\%}$ in a Cox proportional hazards univariate model, together with the cut points proposal available in the literature and the corresponding c-index when applied to the Stable-COPD study.

	Mild	Moderate	Severe	Very Severe	c-index	
					Estimated	Bias corrected
COCOMICS	≥ 70	(55 – 70)	(35 – 55]	≤ 35	0.6003	0.587
GOLD	≥ 80	[50 – 80)	[30 – 50)	< 30	0.5879	0.565
BODE	≥ 65	[50 – 65)	(35 – 50)	≤ 35	0.6003	0.591
<i>Genetic</i> c-index	> 64	(50 – 64]	(36 – 50]	≤ 36	0.6200	0.605
<i>AddFor</i> c-index	> 64	(50 – 64]	(36 – 50]	≤ 36	0.6200	0.602

Additionally, we considered categorising the predictor variable $FEV_{1\%}$ in a multivariate Cox PH model in which the effect of the covariates age and dyspnoea was taken into account. In fact, these variables together with a categorisation of $FEV_{1\%}$ are the ones used in the ADO index (Puhan et al. 2009), which turned out to be the best multivariate index to predict 5-year mortality based on the c-index (Marin et al. 2013).

Table 6.15 shows the results obtained in the multivariate setting for $k = 2$ and $k = 3$ cut points with the CPE and c-index estimators. In a first stage we looked for $k = 3$ cut points and compared them with $k = 2$ cut points, which are also the number of cut points used in the categorisation of $FEV_{1\%}$ in the ADO index. When we compared $k = 2$ versus $k = 3$ number of cut points, we obtained a bootstrap 95% CI for the bias-corrected c-index of $(-0.005, 0.015)$ with the *AddFor* algorithm with a grid of size $M = 1000$ and the c-index as the maximisation criteria. Almost same results were obtained with the *AddFor* with a grid of size $M = 100$ and the *Genetic* algorithm. Consequently, the optimal number of cut points considering the multivariate setting would be $k = 2$. The same optimal cut points were obtained with the *AddFor* and the *Genetic* algorithms, resulting in mild-moderate ($> 50\%$), severe (30 – 50%) and very severe ($< 30\%$) categories. An estimated c-index of 0.734 was obtained, which was higher than the c-index obtained with the ADO categorisation proposal, which was 0.719. The multivariate Cox model with the optimal categorisation $FEV_{1\%}$ adjusted by age and dyspnoea is summarised in Table 6.16.

Table 6.15: Results obtained in the categorisation of the predictor variable $FEV_{1\%}$ of the Stable-COPD study in an multivariate setting with the predictors age and dyspnoea.

Method	Estimator	k	cut points	Concordance probability		
				Estimated	Bias corrected	Difference (95% CI*)
Addfor 100	c-index	2	29.42 ; 50.52	0.734	0.716	-0.002 (-0.005,0.016)
		3	29.42 ; 49.64 ; 50.52	0.737	0.715	
Addfor 1000	c-index	2	29.93 ; 50.05	0.734	0.714	0.008 (-0.005,0.015)
		3	29.93 ; 49.00 ; 50.05	0.737	0.722	
Genetic	c-index	2	29.32 ; 50.69	0.734	0.717	0.006 (-0.004,0.013)
		3	29.90 ; 49.95 ; 50.54	0.737	0.723	
Addfor 100	CPE	2	29.42 ; 50.52	0.709	0.704	0.007 (8e-04,0.013)
		3	29.42 ; 49.64 ; 50.52	0.715	0.712	
Addfor 1000	CPE	2	29.93 ; 50.05	0.709	0.705	0.005 (4e-04,0.012)
		3	29.93 ; 49.00 ; 50.05	0.715	0.710	
Genetic	CPE	2	29.79 ; 50.63	0.709	0.705	0.006 (9e-04,0.013)
		3	29.69 ; 49.37 ; 50.82	0.715	0.711	

Note that results obtained for $k = 3$ cut points when the CPE was used did not correspond with the ones obtained with the c-index. As we mentioned for the univariate setting, the c-index outperformed the CPE in this setting in the simulation results, and hence, we focused on the results obtained when the c-index was maximised.

Table 6.16: Results of the Cox proportional hazards model with the optimal categorisation obtained for the predictor variable $FEV_{1\%}$ adjusted by age and dyspnoea, considering the c-index as the index to be maximised and *Genetic* and *AddFor* algorithms. Hazard ratio (HR) and coefficient estimates (β) are reported together with the p-values of their significance.

Variables	β	HR	p-value
FEV			
≤ 29	1.575	4.832	< 0.0001
(29 – 50]	0.522	1.685	0.003
> 50	-	-	-
Age	0.091	1.095	< 0.0001
Dyspnoea			
0-1	-	-	-
2	0.332	1.394	0.065
3	0.830	2.293	0.004
4	0.825	2.282	0.010
c-index 0.734 and bias-corrected c-index 0.719			

6.4 Conclusions

As an extension to the methodology proposed in Chapter 5, the goal in this chapter was to obtain optimal cut points to categorise continuous predictor variables in a Cox PH model. Hence, in this chapter we present the two algorithms proposed in Chapter 5 but adjusted to the Cox PH approach. We have proposed two alternative estimators of the concordance probability for the maximisation of the discriminative ability of the Cox PH model. These are the c-index proposed by Harrell et al. (1982) and the CPE proposed by Gönen and Heller (2005). Additionally, we have developed a proposal to select the optimal number of cut points based on the bias-corrected concordance probability estimator.

To the best of our knowledge, all previous proposals for categorisation of continuous predictors in a survival model have been done to select a unique cut point; this is to dichotomise the continuous variable. However, we consider that more than two categories may be needed. In fact, the most common multivariate prediction

models used to predict mortality in COPD patients, such as BODE or ADO, use categorised versions (with more than two categories) of continuous predictors (Celli et al. 2004, Puhan et al. 2009).

Sima and Gönen (2013) proposed the maximal discrimination as a method to dichotomise a continuous predictor when the outcome is right censored. They compared the maximisation of the discrimination indexes CPE and c-index together with the maximisation of the long-rank, Wald and partial likelihood ratio statistics for the location of one optimal cut point. What our proposal adds to the proposal of Sima and Gönen (2013) is first, the selection of more than one cut point and second, the ability to categorise a continuous predictor in a multivariate setting.

Both in the application to a real data set of patients with sCOPD and in the simulation study, we see that we have proposed a valid methodology when the aim is to categorise a predictor variable in more than two categories. When we applied the proposed methodology to the Stable-COPD study, we improved the existing categorisation proposals in terms of discriminative ability, obtaining clinically valid optimal cut points. We showed that the optimal cut points proposed in the CO-COMICS study can be improved by applying the methodology presented above.

Additionally, the simulation results for $k = 2$ and $k = 3$ showed that the proposed method performed satisfactorily, especially when the relationship between the predictor and the outcome was not linear. This result makes sense since the change on risk of decreasing time to event might be more pronounced when the relation is not linear. Furthermore, we showed that the two algorithms that we proposed to seek the optimal cut points performed well, although the *Genetic* performed slightly better, as it happened in the proposal for the logistic regression.

Nevertheless, in our simulation study we also considered the particular case of $k = 1$. In fact, Scenarios I and II are similar to the ones considered in Sima and Gönen (2013) (they differ in sample size). When the true cut point lies in the centre of the continuous predictors distribution (Scenario I) we obtained results similar to those obtained by Sima and Gonen; that is, the CPE and the c-index have comparable performance. When the true cut point migrates from the centre of the distribution, Sima and Gonen found that the CPE performed better than the c-index. However, we saw that for a unique and not centred cut point, neither CPE nor c-index performed satisfactorily. Depending on whether the location of the theoretical cut point was shifted to the area of high or low risk, smaller bias and MSE values were obtained for CPE or c-index. Consequently, a limitation of this proposal is that in practice, it should not be used for dichotomisation. Nevertheless,


in case it is used, we recommend to plot the relationship between the predictor continuous variable and the outcome to determine the location of the continuous predictor's value for which the average risk is obtained, and use the best criterion based on its location.


Another limitation of this approach is that we have focused exclusively on the c-index and the CPE as estimators of the discriminative ability of the model. Other estimators have been proposed and compared in the literature; however, none has been proven to be the best (Schmid and Potapov 2012). In our case, we first considered the c-index proposed by Harrell et al. (1982) because it is the most widely used estimator. Since the bias of this estimator increases with increasing censorship, we considered an unbiased estimator that had been specifically developed for the Cox PH model, as it is the CPE proposed by Gönen and Heller (2005). On the other hand, we have not considered time-dependent discriminative ability measures as a parameter for selecting optimal cut points. Thus, we have assumed that the optimal cut points to categorise a continuous predictor variable in a Cox PH model are the same for any given time of interest which may not be true.

Chapter 7

Software development

The work presented in the second section of this chapter is being prepared for publication in the Journal of Statistical Software and has been partially presented in an international conference.

 Barrio, I., Rodríguez-Álvarez, M.X., and Arostegui, I. *CatPredi: An R package for optimal categorisation of continuous predictors*. Journal of Statistical Software; (In preparation)

 ERCIM WG on Computing and Statistics 2012. *Development and implementation of a methodology to select optimal cut-points to categorize continuous covariates in prediction models*. Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Invited contribution. Oviedo, December 2012. This talk was awarded, “Best oral presentation”.

In previous chapters we have proposed several methods to categorise continuous predictor variables in different settings. We now focus on how these methods can be applied in practice using the software R (R Core Team 2014).

In the first section of this chapter, we present the R function we developed to calculate the average-risk category presented in Chapter 4.

Additionally, we have developed a user-friendly R package, named `CatPredi`, to obtain optimal cut points to categorise continuous predictor variables. This package includes the functions needed to apply the methods presented in Chapter 5 and Chapter 6, respectively. Thus, the `CatPredi` package allows for categorising any predictor variable when the response variable is binary or time to event, in either a univariate or a multivariate setting. This package is presented in Section 7.2.

7.1 An R function to calculate the average-risk category

We have developed an R function called `average.risk` to calculate the average-risk category proposed in Chapter 4, above. This function depends on the libraries `mgcv` (Wood 2006) and `BB` (Varadhan and Gilbert 2009) that need to be loaded. The former is needed to fit the GAM model while the latter is used to compute the average-risk value x_0 using the `dfsane` function. This function solves a “Large-Scale Nonlinear System of Equations” (Varadhan and Gilbert 2009).

Before going into the main `average-risk` function, let us briefly introduce the functions `f.smooth` and `f.smooth2`. For each value x of the predictor variable X , a given fitted logistic GAM model and a predicted probability $p_0 = \pi_0$, the function `f.smooth` returns the value $\beta_0 + f(x) - \text{logit}(p_0)$ according to equation (4.2) in Chapter 4. In a similar way, the function `f.smooth2` returns the values of $f(x)$ for each x in X . The specific syntax for these functions is given below.

```
f.smooth <- function (x, model, p0) {

df <- data.frame(x)
names(df) <- attributes(terms(model))$term.labels
res <- as.numeric(predict(model, newdata=df, type="terms")[1]
- family(model)$linkfun(p0) + coef(model)[1])
res
}

f.smooth2 <- function (x, model) {

df <- data.frame(x)
names(df) <- attributes(terms(model))$term.labels
res <- as.numeric(predict(model, newdata=df, type="terms"))
res
}
```

The `average.risk` function returns the average-risk interval, together with the average-risk value x_0 . The arguments needed in this function are the following:

- **Y**: the response variable
- **X**: the predictor variable we want to categorise

- `data`: the data set with all needed variables
- `point`: a real value argument, indicating the initial guess for x_0 .

The code developed for this function is as follows:

```
average.risk <- function(Y, X , data , point ) {

#Fit the GAM model
fitted.model <- gam(Y~s(X, bs="ps"), method="REML",
family = binomial, data = data)

#Look for x0 such that f(x0)=0
x0 <- dfsane(par=point, fn=f.smooth2, model= fitted.model,
control=list(trace=FALSE), quiet=TRUE)

#New data
data.new <- data.frame(X=x0$par)

l.pred <- predict(fitted.model, se=TRUE, newdata=data.new)

u.ci <- exp(l.pred$fit + 1.96*l.pred$se.fit)/(1+exp(l.pred$fit
+ 1.96*l.pred$se.fit))
l.ci <- exp(l.pred$fit - 1.96*l.pred$se.fit)/(1+exp(l.pred$fit
- 1.96*l.pred$se.fit))

p.max <- predict(fitted.model, newdata = data.frame(X=max(X)),
type="response")
p.arisk <- predict(fitted.model, newdata=data.new, type="response")

pos <- p.max - p.arisk
#Upper and lower limits for the average-risk category
if(pos>0){
  inf_x0<-dfsane(par=x0$par-1, fn=f.smooth,
model= fitted.model, p0=l.ci, control=list(trace=FALSE),
quiet=TRUE)
  sup_x0<-dfsane(par=x0$par+1, fn=f.smooth,
```

```

    model= fitted.model, p0=u.ci, control=list(trace=FALSE),
    quiet=TRUE)
} else if(pos <= 0) {
  inf_x0<-dfsane(par=x0$par-1, fn=f.smooth,
  model= fitted.model, p0=u.ci, control=list(trace=FALSE),
  quiet=TRUE)
sup_x0<-dfsane(par=x0$par+1, fn=f.smooth,
  model= fitted.model, p0=l.ci, control=list(trace=FALSE),
  quiet=TRUE)
}
average.risk.cat <- c(inf_x0$par , sup_x0$par)
average.point <- x0$par
res <- list(average.risk.cat = average.risk.cat,
average.point = average.point)
res
}

```

The specific syntax for the categorisation of the predictor variable RR of the IRYSS-COPD study is shown below. Note that after reading the data set, we will select those individuals for which there is no missing value in the predictor variable. Then, before computing the average risk category, we will fit the GAM model and plot the graph in order to visualise the relationship between the predictor variable RR and the poor evolution outcome. Finally, the function `average.risk` will return the average-risk category together with the average-risk point.

```

R> data.ecopd <- read.table("ecopd.txt")
R> no_miss <- which(is.na(data.ecopd$RR)==FALSE)
R> data.ecopd<-data.ecopd[no_miss,]
R> fit <- gam(PoorEvolution~s(RR, bs="ps"), method="REML",
+ family=binomial, data=data.ecopd)
R> plot(fit,shade=T,scale=0, xlab="Respiratory rate",
+ ylab="f(Respiratory rate)")
R> abline(h=0,lty=2,lwd=0.5)
R> average.risk.RR <- average.risk(Y=data.ecopd$PoorEvolution,
+ X=data.ecopd$RR, data=data.ecopd , point=20)
R> average.risk.RR

```



```
$average.risk.cat
```

```
[1] 20.06430 24.11343
```

```
$average.point
```

```
[1] 22.08885
```

7.2 The *CatPredi* package

CatPredi is a package of R functions that allows the user to categorise a continuous predictor variable either before or during the development of a prediction model. Different approaches have been implemented depending on the prediction model chosen, i.e., logistic regression (for binary response variables) or Cox PH model (for time to event outcomes).

The *CatPredi* package can be used to categorise a predictor variable in a univariate or a multivariate setting. It provides the optimal location of cut points for a chosen number of cut points, fits the prediction model with the categorised predictor variable and returns the estimated and bias-corrected discriminative ability index for this model. Additionally, it allows a comparison of two categorisation proposals for a different number of cut points and the selection of the optimal number of cut points.

The *CatPredi* package has been designed similarly to other packages in R. It has two main functions called `catpredi.binary()` and `catpredi.survival()`, which categorise a continuous predictor variable in a logistic regression model or a Cox PH model, respectively. Numerical and graphical summaries of the fitted objects can be obtained by using `print.catpredi.binary`, `summary.catpredi.binary` and `plot.catpredi.binary` for `catpredi.binary` type objects and analogously for `catpredi.survival` type objects. Table 7.1 contains a description of all the functions available in the package. Furthermore, two more main functions have been developed - `comp.cutpoints.binary` and `comp.cutpoints.survival` - to obtain the optimal number of cut points in a logistic regression or Cox PH model, respectively.

Below, we give a general overview of the package and its general use. As an illustrative example for the `catpredi.binary()` function, we will use the IRYSS-COPD study presented in Section 2.1. In addition, we will use the Stable-COPD study presented in Section 2.2 to exemplify the `catpredi.survival()` function.

Table 7.1: Summary of the functions in the `CatPredi` package.

Function	Description
<code>catpredi.binary</code>	Returns an object with the optimal cut points to categorise a continuous predictor variable in a logistic regression model.
<code>controlcatpredi.binary</code>	Function used to set several parameters to control the selection of the optimal cut points in a logistic regression model.
<code>print.catpredi.binary</code>	Print method for objects of type <code>catpredi.binary</code> .
<code>summary.catpredi.binary</code>	Produces a summary of the <code>catpredi.binary</code> object.
<code>plot.catpredi.binary</code>	Plots the relationship between the continuous predictor and the response variable obtained by fitting a GAM function, together with the location of the optimal cut points.
<code>comp.cutpoints.binary</code>	Compares two objects of type <code>catpredi.binary</code> .
<code>print.comp.cutpoints.binary</code>	Print method for objects of type <code>comp.cutpoints.binary</code> .
<code>catpredi.survival</code>	Returns an object with the optimal cut points to categorise a continuous predictor variable in a Cox PH regression model.
<code>controlcatpredi.survival</code>	Function used to set several parameters to control the selection of the optimal cut points in a Cox PH regression model.
<code>print.catpredi.survival</code>	Print method for objects of type <code>catpredi.survival</code> .

Continues on next page.

Function	Description
<code>summary.catpredi.survival</code>	Produces a summary of the <code>catpredi.survival</code> object.
<code>plot.catpredi.survival</code>	Plots the smooth relationship between the continuous predictor and the estimated smooth function in a Cox PH model, together with the location of the optimal cut points.
<code>comp.cutpoints.survival</code>	Compares two objects of type <code>catpredi.survival</code> .
<code>print.comp.cutpoints.survival</code>	Print method for objects of type <code>comp.cutpoints.survival</code> .

7.2.1 `catpredi.binary()` function

The `catpredi.binary()` function provides the optimal cut points to categorise a continuous predictor variable in a logistic regression model. This function creates an object of class `catpredi.binary`. The main arguments of this function are presented in Table 7.2. The call to the function is as follows:

```
catpredi.binary(formula, cat.var, cat.points, data,
method = c("addfor", "genetic"), range=NULL, correct.AUC=TRUE,
control = controlcatpredi.binary())
```

In the `formula` argument users must specify the prediction model setting in which they want to categorise the predictor variable X specified in the `cat.var="X"` argument. If the model is univariate, then the formula would be specified as $Y \sim 1$, with Y being the response variable available in the data set specified in the argument `data`. However, if the model is multivariate, and the aim is to categorise the predictor variable X together with another predictor Z , then the formula would be specified as $Y \sim Z$. Additionally, in the argument `cat.points` the user must specify the number of cut points to look for. The `range` argument allows for modifying the range of the predictor variable X in which to look for the cut points. By default it would be `NULL`, which represents the entire range of X . Finally, if `correct.AUC` is set to `TRUE`, the bias-corrected AUC would be estimated.

Table 7.2: Summary of the arguments in the `catpredi.binary()` function.

Argument	Description
<code>formula</code>	A formula giving the model to be fitted.
<code>cat.var</code>	Name of the continuous variable to categorise.
<code>cat.points</code>	Number of cut points to look for.
<code>data</code>	Data frame containing all needed variables.
<code>method</code>	The algorithm selected to search for the optimal cut points- "addfor" if the <i>AddFor</i> algorithm is chosen; otherwise, "genetic".
<code>range</code>	The range of the continuous variable in which to look for the cut points. By default NULL, i.e., the entire range.
<code>correct.AUC</code>	A logical value. If TRUE the bias-corrected AUC is estimated.
<code>control</code>	Output of the <code>controlcatpredi.binary()</code> function

The specific syntax for the eCOPD data using a univariate model is shown below. In this example, the *AddFor* algorithm with a grid of size $M = 100$ is used to look for two optimal cut points for the predictor variable PCO_2 .

```
R> library(CatPredi)
R> data.ecopd <- read.table("ecopd.txt")
R> cat.k2 <- catpredi.binary(VerySevereEvolution~1, cat.var="pco2",
+ cat.points=2, data=data.ecopd, method="addfor", correct.AUC=TRUE)
```

A numerical summary of the results of the categorisation method can be obtained by calling the functions `print.catpredi.binary()` or `summary.catpredi.binary()`. While the former gives the optimal cut points together with the corresponding estimated AUC and bias-corrected AUC, the latter gives, in addition to that, the fitted logistic regression model for the categorised predictor variable. When the method selected is the *AddFor*, the summary returns the estimated AUC for each of the selected cut points. For example, if $k = 2$ is chosen, it returns the estimated AUC for $k = 1$ and $k = 2$. Additionally, if `correct.AUC=TRUE` is chosen it returns the bias-corrected AUC for $k = 2$. If the method selected is the *Genetic*, estimated cut points, AUC and bias-corrected AUC will be given only for the selected number of

cut points.

```
R> summary(cat.k2)
```

Call:

```
catpredi.binary(formula = VerySevereEvolution ~ 1, cat.var = "pco2",
cat.points = 2, data = data.ecopd, method = "addfor",
correct.AUC = TRUE)
```

```
*****
```

```
Addfor Search Algorithm
```

```
*****
```

Optimal cutpoints	Optimal AUC	Corrected AUC
50.87	0.6969	NA
62.67	0.7213	0.696

```
-----
Fitted model for the categorised predictor variable
-----
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.9479	0.1875	-21.060	< 2e-16 ***
pco2_cat(50.9,62.7]	1.0972	0.2967	3.698	0.000217 ***
pco2_cat(62.7,160]	2.2010	0.2570	8.565	< 2e-16 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.2.2 controlcatpredi.binary() function

The function `catpredi.binary()` has the argument `control`, which can be used to set several parameters for the categorisation process. This argument is the output of the `controlcatpredi.binary()` function. For instance, the grid size used

with the *AddFor* algorithm can be specified in the `addfor.g` argument (by default `addfor.g=100`). In addition, the number of bootstrap replicates used for the bias correction of the AUC can also be specified in argument `B`, which by default takes the value of 50. The argument `b.method` allows for specifying whether the bootstrap resampling should be done considering the outcome variable. The option “ncoutcome” indicates that the data is resampled without taking into account the response variable, while “coutcome” indicates that the data is resampled in regard to the response variable. Other arguments such as `min.p.cat` and `print.gen` are also available in the `controlcatpredi.binary()` function. The former allows for specifying the minimum number of individuals in each category. If the user wants to ensure a minimum number of individuals in each category, he or she should specify this number in the `min.p.cat` argument. Finally, the latter corresponds to the argument `print.level` of the `genoud()` function in the `rgenoud` package (Mebane and Sekhon 2011). This argument controls the level of printing that the `genoud` function does, which in our setting, corresponds to the printing of the optimisation process when the *Genetic* algorithm is used.

7.2.3 `plot.catpredi.binary()` function

The function `plot.catpredi.binary()` plots the relationship between the continuous predictor variable that is selected to be categorised and the response variable based on a GAM model. Additionally, the optimal cut points obtained with the `catpredi.binary()` function are drawn on the graph.

```
R> plot(cat.k2)
```

The result of the above code is shown in Figure 7.1.

7.2.4 `catpredi.survival()` function

The `catpredi.survival()` function provides the optimal cut points to categorise a continuous predictor variable in a Cox PH model. This function creates an object of class `catpredi.survival`. The main arguments of this function are presented in Table 7.3. The call to the function is as follows:

```
catpredi.survival(formula, cat.var, cat.points, data,
method = c("addfor","genetic"), conc.index = c("cindex","cpe"),
range = NULL, correct.index = TRUE ,
control = controlcatpredi.survival)
```

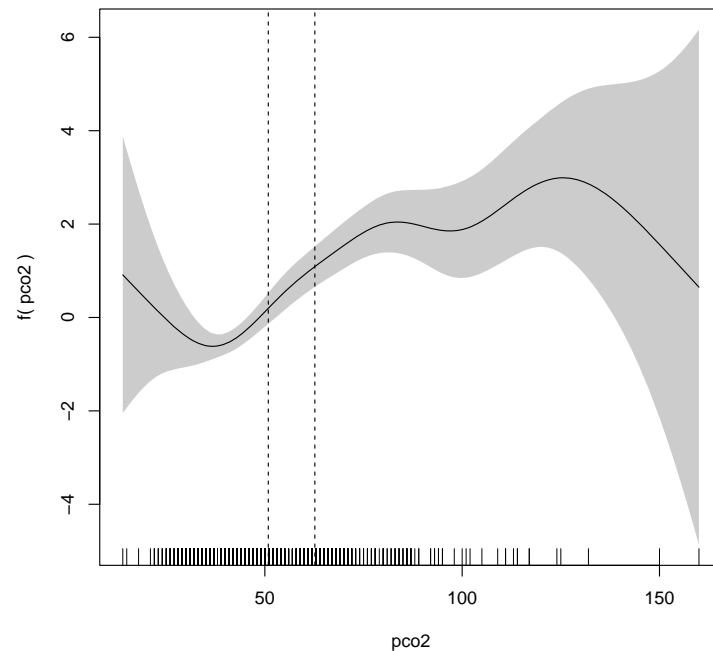


Figure 7.1: Relationship between the predictor variable PCO_2 and the response variable very severe evolution from the IRYSS-COPD study based on a logistic GAM, together with the optimal cut points obtained with the `catpredi.binary()` function.

In the `formula` argument, users must specify the prediction model setting in which they want to categorise the predictor variable X specified in the argument `cat.var="X"`. The response is written on the left of a \sim operator and the terms, separated by $+$ operators, on the right. The response must be a `Surv` object. If the model is univariate, then the formula would be specified as `Surv(SurvT, SurvS)~1`, being `SurvT` the observed survival time and `SurvS` the status indicator (0=censored and 1=event) in the data set specified in the argument `data`. If the model is multivariate, and the aim is to categorise the predictor variable X together with another predictor Z , then the formula would be specified as `Surv(SurvT, SurvS)~Z`. Additionally, in the argument `cat.points` we must specify the number of cut points to look for. The `range` argument allows for modifying the range of the predictor variable X in which to look for the cut points. By default it is `NULL` which represents the entire range of X . In the argument `conc.index` the user must specify the

discriminative ability index estimator chosen for maximisation purposes, i.e., “cpe” if the concordance probability estimator proposed by Gönen and Heller (2005) is chosen and “cindex” otherwise. If an estimation of the bias-corrected discriminative ability index is desired, then the argument `correct.AUC` should be set to `TRUE`.

Table 7.3: Summary of the arguments in the `catpredi.survival()` function.

Argument	Description
<code>formula</code>	A formula giving the model to be fitted. Left hand side of the formula must be an object of type. <code>Surv</code>
<code>cat.var</code>	Name of the continuous variable to categorise.
<code>cat.points</code>	Number of cut points to look for.
<code>data</code>	Data frame containing all needed variables.
<code>method</code>	The algorithm selected to search for the optimal cut points- “addfor” if the <i>AddFor</i> algorithm is chosen; otherwise, “genetic”.
<code>conc.index</code>	The discriminative ability index selected for maximisation purpose. “cindex” if the c-index proposed by Harrell et al. (1982) is selected and “cpe” if the concordance probability estimator proposed by Gönen and Heller (2005) is chosen.
<code>range</code>	The range of the continuous variable in which to look for the cut points. By default <code>NULL</code> , i.e., all the range.
<code>correct.index</code>	A logical value. If <code>TRUE</code> the bias-corrected discriminative ability index is estimated according to the estimator selected in the argument <code>conc.index</code> .
<code>control</code>	Output of the <code>controlcatpredi.survival()</code> function.

The specific syntax for the sCOPD data using a multivariate model adjusted for age and dyspnoea, is shown below. In this example, the *Genetic* algorithm is used to look for two optimal cut points for the predictor variable $FEV_{1\%}$. The discriminative ability index chosen in this case was the c-index proposed by Harrell et al. (1982).

```
R> library(CatPredi)
R> data.scopd <- read.table("scopd.txt")
```



```
R> summary(data.scopd)
```

Surv.time	Event	Fev1	Age	Dyspnoea
Min. : 23	Min. :0.0000	Min. : 18	Min. :33.00	0-1:333
1st Qu.:1618	1st Qu.:0.0000	1st Qu.: 45	1st Qu.:63.00	2 :166
Median :1825	Median :0.0000	Median : 55	Median :70.00	3 : 23
Mean :1575	Mean :0.3076	Mean : 55	Mean :68.32	4 : 21
3rd Qu.:1825	3rd Qu.:1.0000	3rd Qu.: 65	3rd Qu.:75.00	
Max. :2045	Max. :1.0000	Max. :105	Max. :86.00	

```
R> cat.k2.surv <- catpredi.survival(Surv(Surv.time, Event) ~ Age
+ Dyspnoea, cat.var="Fev1", data=data.scopd, cat.points=2,
method="genetic", conc.index="cindex", correct.index = TRUE)
```

A numerical summary of the results of the categorisation method can be obtained by calling the functions `print.catpredi.survival()` or `summary.catpredi.survival()`. While the former gives the optimal cut points together with the corresponding estimated c-index and bias-corrected c-index, the latter gives in addition to that the fitted Cox PH multivariate model for the categorised predictor variable together with the covariates age and dyspnoea.

```
R> summary(cat.k2.surv)
```

```
Call:
```

```
catpredi.survival(formula = Surv(Surv.time, Event) ~ Age + Dyspnoea,
  cat.var = "Fev1", cat.points = 2, data = data.scopd,
  method = "genetic", conc.index = "cindex", correct.index =TRUE)
```

```
*****
```

```
Genetic Search AlgorithmConcordance C-index
```

```
*****
```

```
Optimal cutpoints
```

```
29.32
```

```
50.69
```

```
Optimal Cindex
```

```
0.7335
```

```
Corrected Cindex
0.7167
```

```
-----
Fitted cph model for the categorised predictor variable
-----
```

```
Cox Proportional Hazards Model
```

```
cph(formula = formula.n, data = data)
```

	Coef	S.E.	Wald Z	Pr(> Z)
Age	0.0909	0.0133	6.85	<0.0001
Dyspnoea=2	0.3320	0.1799	1.85	0.0649
Dyspnoea=3	0.8299	0.2895	2.87	0.0041
Dyspnoea=4	0.8251	0.3216	2.57	0.0103
Fev1_cat=(29.3,50.7]	-1.0540	0.3312	-3.18	0.0015
Fev1_cat=(50.7,105]	-1.5756	0.3454	-4.56	<0.0001

7.2.5 controlcatpredi.survival() function

The function `catpredi.survival()` has the argument `control` that can be used to set several parameters for the categorisation process. This argument is the output of the `controlcatpredi.survival()` function. For instance, the grid size used with the *AddFor* algorithm can be specified in the `addfor.g` argument (by default `addfor.g=100`). In addition, the number of bootstrap replicates used for the bias correction of the “cpe” or “cindex” can also be specified in the `B` argument, which by default takes the value of 50. The argument `b.method` allows for specifying whether the bootstrap resampling should be done considering the outcome variable. The option “`ncoutcome`” indicates that the data is resampled without taking into account the variable event indicator. Other arguments such as `min.p.cat` and `print.gen` are also available in the `controlcatpredi.survival()` function. The former allows

for specifying the minimum number of individuals in each category, which by default is five. Finally, the latter corresponds to the argument `print.level` of the `genoud()` function in the `rgenoud` package in the same way as we explained previously in the `controlcatpredi.binary()` function.

7.2.6 `plot.catpredi.survival()` function

The function `plot.catpredi.survival()` plots the functional form of the predictor variable we want to categorise. Additionally, the optimal cut points obtained with the `catpredi.survival()` function are drawn on the graph.

```
R> plot(cat.k2.surv)
```

The result of the above code is shown in Figure 7.2.

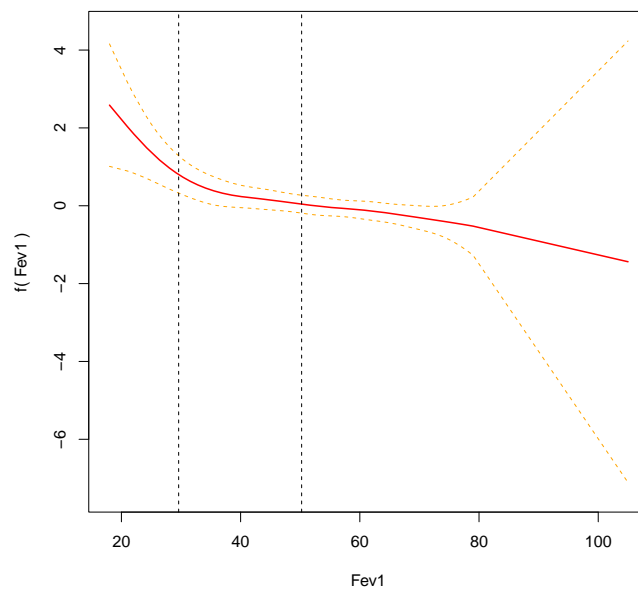


Figure 7.2: Smooth function for the predictor variable $FEV_{1\%}$ in the fitted multivariate Cox PH model, together with the optimal cut points obtained with the `catpredi.survival()` function.

7.2.7 `comp.cutpoints.binary()` function

The function `comp.cutpoints.binary()` allows for the comparison of two objects of type `catpredi.binary`. The aim of this function is to obtain the optimal number of cut points as explained in Section 5.1.2, above. The main arguments of this function are presented in Table 7.4. The call to the function is as follows:

```
comp.cutpoints.binary(obj1, obj2, V = 100)
```

In the arguments `obj1` and `obj2`, users must specify the two `catpredi.binary` objects they want to compare. It must be noted that the models fitted in both objects must be the same, where the only difference is the selected number of cut points. Finally, the argument `V` specifies the number of bootstrap resamples used to select the optimal number of cut points (by default `V=100`).

Table 7.4: Summary of the arguments in the `comp.cutpoints.binary()` function.

Argument	Description
<code>obj1</code>	<code>catpredi.binary</code> type object for k number of cut points.
<code>obj2</code>	<code>catpredi.binary</code> type object for $m \neq k$ number of cut points.
<code>V</code>	Number of bootstraps resamples. By default $V = 100$.

Continuing with the example above for the eCOPD data, we will look for the optimal number of cut points for the predictor variable PCO_2 in a univariate setting.

```
R> cat.k2 <- catpredi.binary(VerySevereEvolution~1, cat.var="pco2",
+ cat.points=2, data=data.ecopd, method="addfor", correct.AUC=TRUE)
R>
R> cat.k3 <- catpredi.binary(VerySevereEvolution~1, cat.var="pco2",
+ cat.points=3, data=data.ecopd, method="addfor", correct.AUC=TRUE)
R>
R> comp.k2k3 <- comp.cutpoints.binary(cat.k2, cat.k3, V = 100)
```

A numerical summary of the result of the selection of optimal cut points can be obtained by calling the function `print.comp.cutpoints.binary()`.

```
*****
Compare optimal number of cut points
*****

Bias-corrected AUC difference:  0.0139
95% Bootstrap Confidence Interval: ( -3e-03 , 0.0415 )
```

7.2.8 `comp.cutpoints.survival()` function

The function `comp.cutpoints.survival()` allows for the comparison of two objects of type `catpredi.survival`. The aim of this function is to obtain the optimal number of cut points as explained in Section 6.1.2, above. The main arguments of this function are presented in Table 7.5. The call to the function is as follows:

```
comp.cutpoints.survival(obj1, obj2, V = 100)
```

In the arguments `obj1` and `obj2`, users must specify the two `catpredi.survival` objects they want to compare. It must be noted that the models fitted as well as the selected discrimination indexes must be the same in both objects, where the only difference is the selected number of cut points. Finally, the argument `V` specifies the number of bootstrap resamples used to select the optimal number of cut points (by default `V=100`).

Table 7.5: Summary of the arguments in the `comp.cutpoints.survival()` function.

Argument	Description
<code>obj1</code>	<code>catpredi.survival</code> type object for k number of cut points.
<code>obj2</code>	<code>catpredi.survival</code> type object for $m \neq k$ number of cut points.
<code>V</code>	Number of bootstraps resamples. By default $V = 100$.

Continuing with the example above for the `sCOPD` data, we will look for the optimal number of cut points for the predictor variable $FEV_{1\%}$ in a multivariate setting.

```
R> cat.k2.surv <- catpredi.survival(Surv(Surv.time, Event)~Age
+ Dyspnoea, cat.var="Fev1", data=data.scopd, cat.points=2,
method="genetic", conc.index="cindex",correct.index = TRUE)
```

```
R>
R> cat.k3.surv <- catpredi.survival(Surv(Surv.time, Event)~Age
+ Dyspnoea, cat.var="Fev1", data=data.scopd, cat.points=3,
method="genetic", conc.index="cindex",correct.index = TRUE)
R>
R> comp.k2k3.surv <- comp.cutpoints.survival(cat.k2.surv,
cat.k3.surv, V = 100)
```

A numerical summary of the result of the selection of optimal cut points can be obtained by calling the function `print.comp.cutpoints.survival()`.

Call:

```
comp.cutpoints.survival(obj1 = cat.k2.surv, obj2 = cat.k3.surv,
V = 100)
```

```
*****
```

```
Compare optimal number of cut points
```

```
*****
```

```
Bias-corrected concordance difference: 0.006
```

```
95% Bootstrap Confidence Interval: ( -0.0044 , 0.0134 )
```

Conclusions and Future Research

In clinical practice, decisions need to be made by reference to clinical parameters, which are usually continuous measurements. Accurate knowledge of the relationship between such parameters and the risk of developing a certain outcome helps identify individuals most at risk. Considering these parameters as continuous predictors is preferable from a statistical point of view, since categorisation may lead to loss of information and reduction in power (Royston et al. 2006). Nevertheless, in the development of clinical prediction models for application in clinical practice, it may be preferable for a certain amount of information to be sacrificed in the interests of enhanced utility and ease of use in daily clinical practice. Moreover, in a study such as the IRYSS-COPD (Quintana et al. 2011), ED clinical practice prevails over research requirements. Hence, the data available is the information routinely recorded at the ED for eCOPD patients. In our opinion, the sacrifice of some information in the subset of data recorded as a continuous variable to avoid information exclusion of the subset of data recorded as ordinal variable is a worthwhile trade-off.

In this dissertation, we proposed different approaches for the categorisation of continuous variables depending on the distribution of the response variable. As a first step, in Chapter 4 we proposed to categorise a continuous variable in a minimum of three categories based on the graphical relationship between the continuous predictor and the outcome given by a GAM with P-spline smoothers. This methodology was based on a previous proposal made by Hin et al. (1999) that considered the selection of a unique cut point. Our proposal provided the possibility of categorising the variable into more than two categories compared with what already existed in the literature. However, the selection for an extra cut point when more than three categories were needed was subjected to clinical significance and the graphical dis-

play. Hence, we considered developing a more accurate methodology to categorise continuous predictor variables in prediction models, which was presented in Chapter 5 and Chapter 6 of this dissertation.

As we mentioned before, previous work on categorisation has been done, but with the aim in almost all cases to dichotomise the continuous predictor variable, where the most common alternative is the minimum p-value approach (Miller and Siegmund 1982). Tsuruta and Bax (2006) proposed a parametric method for obtaining more than one cut point based on the overall discrimination of the prediction model. Tsuruta and Bax (2006) showed the optimal location of the cut points when the distribution of the predictor variable was known. In this dissertation, we proposed a methodology to categorise a continuous predictor variable considering any given number of cut points. Our proposal is based on the maximisation of the discriminative ability of the prediction model but without assuming any distribution for the predictor variable. Furthermore, two alternative algorithms were proposed, *AddFor* and *Genetic*, in order to look for the vector of cut points to maximise the discriminative ability. Additionally, our proposal allows for the categorisation of the predictor variable either in a univariate or multivariate model. To the best of our knowledge, none of the previous proposals allowed the categorisation in a multivariate setting, that is, during the development of the model. In fact, Mazumdar et al. (2003) stated that if the aim is to develop a multivariate model, the categorisation should be performed taking into account the effect that other covariates may have on the predictor variable we wished to categorise.

Furthermore, since the aim is to use the prediction models in practice, reporting the discriminative ability of the final model (with the categorised variable) is an important issue. However, since the same data has been used for the estimation of the optimal cut points and the model development, the discriminative ability may be overestimated. Hence, in this dissertation, we proposed a bootstrap based approach to correct the optimism of the discriminative ability of the model. This has been specifically developed for the categorisation proposal, although it could be extended to any other model development.

Looking for the best number of categories in which to categorise a predictor variable is an area of interest in practice. We are aware that in theory the optimal number of cut points for the categorisation of a continuous variable does not exist, since above all the possible number of cut points, the best option would be the continuous variable. However, in clinical practice categorical versions of the continuous variables are usually preferred without it always being clear how many

categories is best. It is necessary to find a balance between the clinical sense of the categories and the minimal loss of information. Therefore, in this dissertation, we proposed a bootstrap based approach to select the optimal number of cut points based on the differences of the bias-corrected AUCs for $k = l + 1$ and $k = l$ number of cut points. To the best of our knowledge, none of the previous works about categorisation considered selecting the best number of cut points.

Jointly with the logistic regression, the survival model is the most widely used prediction model in clinical practice. Among the many outcomes of interest are either the event or the time until the event occurs. To model the latter, the Cox PH model is the most broadly used method. Therefore, all the proposed approaches mentioned above have been developed for these different settings, i.e., the logistic regression model and the Cox PH model. Additionally, the proposed methods have been validated considering different contexts and scenarios.

The statistical software R is a free software environment for statistical computing and graphics (R Core Team 2014). The methodology developed and proposed in this dissertation has been implemented in an easy to use R package called *CatPredi*. This package helps to obtain optimal cut points to categorise continuous predictor variables either in a logistic regression or a Cox PH model. The aim of this package is to provide an easy-to-use tool for clinical researchers to obtain optimal cut points to categorise continuous predictor variables whenever it is deemed necessary. As far as we know, there is no other package in R that provides the optimal categorisation of a continuous variable in more than two categories. Therefore, the availability and easy use of this package allows researchers to obtain optimal cut points whenever they consider it to be necessary.

Nevertheless, this proposal also has some limitations that should be taken into account. First of all, to search for the vector of optimal cut points, we proposed two alternative algorithms, namely, *AddFor* and *Genetic*. The *AddFor* algorithm looks for one cut point at a time, i.e., once the first one is selected it is fixed and the second cut point is sought, which means that in some circumstances the vector of cut points obtained with the *AddFor* might not be “optimal”. This happened especially when we looked for two cut points where simulation results showed that the *Genetic* algorithm performed more successfully. However, the *Genetic* algorithm is computationally more expensive, which for very large sample sizes might be not feasible. Secondly, two concordance probability estimators were considered for maximisation purposes when the methodology was developed for Cox PH models: the c-index proposed by Harrell et al. (1982) and the CPE proposed by Gönen and Heller (2005),

respectively. Other estimators for the concordance probability have been developed in the literature, which we did not take into account in this dissertation. In addition, time dependent discriminatory measures were out of the scope of this dissertation and were not studied here. Lastly, we must note that in this dissertation we do not recommend the categorisation as a modelling solution, but our goal is to propose a valid way to do so whenever it is considered necessary by clinical researchers.

In general, we would like to provide some recommendations on the use of the methodology developed in this thesis. First of all, as long as it can be computationally attained, we recommend the use of the *Genetic* algorithm rather than the *AddFor* algorithm. Secondly, when the aim is to categorise a continuous predictor variable in a Cox PH model, we recommend not to use this methodology to select a unique cut point for dichotomisation, unless the average risk value is centred in the continuous predictors distribution. For more than one cut point, as a general rule, we recommend the use of the CPE for low censoring rates and the c-index for high censoring rates.

In conclusion, we have proposed and validated a methodology to categorise continuous predictor variables in prediction models. This methodology would be very valuable in the development of prediction models and in the application of these models in practice. The *CatPredi* package we developed will allow clinical researchers to apply this methodology in an optimal and easy way.

Further research

Some aspects related to this dissertation are subject to further research. Firstly, we are interested in the categorisation of more than a unique continuous variable. In this dissertation, we focused on the categorisation of a unique predictor variable either in a univariate or a multivariate setting. From a clinical point of view it might be interesting to categorise two predictor variables at a time; thus, the categorisation of one variable would depend on the categorisation of the other. Initially, we thought of two possible approaches to do this. On the one hand, it would be the extension of the *AddFor* and *Genetic* algorithms to search for cut points in a two-dimensional setting rather than in a one-dimensional setting as we proposed in this dissertation. On the other hand, we think we could apply P-spline ANOVA-type methodology (Lee and Durbán 2011) to this specific setting, in order to categorise two continuous predictors based on different risk areas obtained from the estimation of interaction terms which would be decomposed as a sum of smooth functions.

Secondly, it would be also interesting to evaluate time dependent discriminative ability measures as the maximisation objective in the categorisation of continuous variables in a Cox PH model. In this dissertation we dealt with the concordance probability, which is a global discriminative ability measure. This implies that the optimal cut points are considered to be the same whatever the time of interest is. In our opinion, it might be of interest to compare the estimated cut points when time-dependent discriminative ability measures are used. Would the estimated cut points be different for different event times? Apparently, this question makes sense clinically. Hence, we will study the performance of different time dependent indices in the categorisation of predictor variables in a Cox PH model. Several time dependent discriminative ability estimators have been proposed in the literature. For example, Zheng and Heagerty (2004) proposed a semi-parametric estimator for the time-dependent ROC curve, and Antolini et al. (2005) proposed a time-dependent discrimination index specifically developed for survival data. Other authors have proposed alternative estimators for time-dependent discriminative ability measures (Chambless and Diao 2006, Uno et al. 2007). We think it would be interesting to study all these estimators and their performance in the categorisation of a continuous predictor variable in a Cox PH setting.

The categorisation of continuous predictor variables considering other regression modelling approaches is also of great interest. Studying the categorisation of a predictor variable for outcomes with the Poisson distribution, the Beta-binomial distribution or in a mixed effects model may be of interest in practice.



In this dissertation we have focused on the categorisation of continuous variables considering a minimal loss of information in regards to the discriminative ability of the model. In the development of prediction models it is also of great importance to assess the goodness of fit of the model. It would be of interest to study the influence categorisation and more specifically, the number of categories have on the calibration of the model.

Finally, although the software presented in this dissertation covers the proposed methods, it would be worthwhile to implement some extensions of interest. Thus, for instance, it could be useful to extend the `CatPredi` package to allow for categorising a continuous predictor variable in the presence of smooth covariates.

Contributions





In this section we present the contributions that have come directly from the work of this dissertation. Additionally, we present the grants and awards that the author has obtained in recognition of the work carried out and directly related to this dissertation.

Research articles

-  Barrio, I., Arostegui, I., Quintana, J.M., and IRYSS-COPD Group. Use of generalised additive models to categorise continuous variables in clinical prediction. *BMC medical research methodology* 2013; **13**:83.
-  Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. A new approach to categorising continuous variables in prediction models: proposal and validation. *Statistical Methods in Medical Research* (under review)







Conferences

Invited

-  5th International Conference of the ERCIM WG on Computing & Statistics. *Development and implementation of a methodology to select optimal cut-points to categorize continuous covariates in prediction models*. Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Oviedo, December 2012.
-  One Day Meeting on Statistics and Applications. *Categorization of continuous variables in clinical prediction models. Development, validation and application of a new approach*. Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Guimaraes, December 2013.
-  IV Jornadas de Investigación de la Facultad de Ciencia y Tecnología. *Development of statistical methodology for research: categorization of continuous variables in prediction models*. Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Leioa, February 2014.
-  Joint Meeting of the International Biometric Society (IBS) Austro-Swiss and Italian regions. *Categorization of continuous predictors in the development*

of prediction models by maximization of the AUC. Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Milan, June 2015.

Contributions

-  HTA in Integrated Care for a Patient Centered System. *Continuous variables categorization to apply into the development of predictive models for patients with COPD exacerbation.* Barrio, I., Arostegui, I., Quintana, J.M., Esteban, C., and IRYSS-COPD Group. Bilbao, June 2012.
-  International Workshop on Statistical Modelling. *Location of optimal cut-points to categorize continuous variables in clinical studies.* Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Prague, July 2012.
-  II Congreso de Jóvenes Investigadores en Diseño de Experimentos y Bioestadística. *Validación de métodos de obtención de puntos de corte óptimos para categorizar variables continuas.* Arostegui, I., Barrio, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Tenerife, July 2012.
-  Reunión Científica de la Sociedad Española de Epidemiología. *Metodología de categorización de variables continuas a aplicar en el desarrollo de modelos predictivos para los pacientes con una exacerbación de EPOC.* Barrio, I., Arostegui, I., García-Gutierrez, S., Quintana, J.M., and the IRYSS-COPD Group. Santander, October 2012.
-  XIV Conferencia Española de Biometría. *Methodology to categorize continuous variables in prediction models: proposal and validation.* Barrio, I., Arostegui, I., Rodríguez-Álvarez, M.X., and Quintana, J.M. Ciudad Real, May 2013.
-  XXVII International Biometric Conference. *Optimal cut points to categorize continuous predictor variables in a Cox Proportional Hazards Model.* Barrio, I., Rodríguez-Álvarez, M.X., Meira-Machado, L., Quintana, J.M., and Arostegui, I. Florence, July 2014.

Grants and Awards

- Award for the **best student** of the Master's degree *Master en Modelización Matemática Estadística y Computación* of 2009/2010 class.

-
- The prize for **best oral presentation** for the talk entitled, *Development and implementation of a methodology to select optimal cut-points to categorize continuous covariates in prediction models*.
 - Third prize for **young researcher** for the talk entitled *Methodology to categorize continuous variables in prediction models: proposal and validation*.
 - Selected to perform a one week research-stay in Guimaraes funded by BIOSTAT-NET.
 - Grant from the Spanish Biometric Society to participate in the XIVth Spanish Biometric Conference.
 - Grant from the Spanish Biometric Society to participate in the Joint Meeting of the International Biometric Society (IBS), Austro-Swiss and Italian regions.

References

- Airola, A., Pahikkala, T., Waegeman, W., De Baets, W., and Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the roc curve. *Computational Statistics & Data Analysis*, 55:1828–1844.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.
- Almagro, P., Martinez-Camblor, P., Soriano, J., Marin, J., Alfageme, I., Casanova, C., Esteban, C., Soler-Cataluña, J., De-Torres, J., and Celli, B. (2014). Finding the best thresholds of FEV1 and dyspnea to predict 5-year survival in COPD patients: the COCOMICS study. *PLoS One*, 9:e89866.
- Altman, D. G., Lausen, B., Sauerbrei, W., and Schumacher, M. (1994). Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86:829–835.
- Altman, D. G. and Lyman, G. (1998). Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment*, 52:289–303.
- Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24:3927–3944.
- Azarisman, M., Fauzi, M., Faizal, M., Azami, Z., Roslina, A., and Roslan, H. (2007). The SAFE (SGRQ score, air-flow limitation and exercise tolerance) index: a new composite score for the stratification of severity in chronic obstructive pulmonary disease. *Postgraduate Medical Journal*, 83:492–497.

- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415.
- Bandos, A. I., Rockette, H. E., and Gur, D. (2005). A permutation test sensitive to differences in areas for comparing roc curves from a paired design. *Statistics in Medicine*, 24:2873–2893.
- Bandos, A. I., Rockette, H. E., and Gur, D. (2006). A permutation test for comparing ROC curves in multireader studies: a multi-reader ROC, permutation test. *Academic Radiology*, 13:414–420.
- Begg, C. B., Cramer, L. D., Venkatraman, E., and Rosai, J. (2000). Comparing tumour staging and grading systems: a case study and a review of the issues, using thymoma as a model. *Statistics in Medicine*, 19:1997–2014.
- Bennette, C. and Vickers, A. (2012). Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, 12:21.
- Bestall, J. C., Paul, E. A., Garrod, R., Garnham, R., Jones, P. W., and Wedzicha, J. A. (1999). Usefulness of the medical research council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease. *Thorax*, 54:581–586.
- Braun, T. M. and Alonzo, T. A. (2008). A modified sign test for comparing paired ROC curves. *Biostatistics*, 9:364–372.
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.
- Buist, A. S., Vollmer, W. M., and McBurnie, M. A. (2008). Worldwide burden of COPD in high-and low-income countries. Part I. The Burden of Obstructive Lung Disease (BOLD) Initiative. *The International Journal of Tuberculosis and Lung Disease*, 12:703–708.
- Celli, B. R., Cote, C. G., Marin, J. M., Casanova, C., Montes de Oca, M., Mendez, R. A., Pinto Plata, V., and Cabral, H. J. (2004). The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *New England Journal of Medicine*, 350:1005–1012.

- Chambless, L. E. and Diao, G. (2006). Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in Medicine*, 25:3474–3486.
- Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, 40:373–383.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7:249–253.
- Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115:928–935.
- Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the roc curve. *Clinical Chemistry*, 54:17–23.
- Copas, J. B. and Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*, 89:315–331.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.
- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. CRC Press.
- Currie, I. D., Durban, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, 68:259–280.
- De Boor, C. (2001). *A Practical Guide to Splines. Revised Edition*. New York: Springer-Verlag.
- DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845.
- Demler, O. V., Pencina, M. J., and D’Agostino, R. B. (2012). Misuse of DeLong test to compare AUCs for nested models. *Statistics in Medicine*, 31:2577–2587.
- Efron, B. (1977). The efficiency of Cox’s likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565.

- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Eiben, A. E. and Smith, J. E. (2003). *Introduction to Evolutionary Computing*. Berlin: Springer.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121.
- Esteban, C., Arostegui, I., Aburto, M., Moraza, J., Quintana, J. M., Aizpiri, S., Basualdo, L. V., and Capelastegui, A. (2014). Influence of changes in physical activity on frequency of hospitalization in chronic obstructive pulmonary disease. *Respirology*, 19:330–338.
- Esteban, C., Quintana, J. M., Aburto, M., Moraza, J., and Capelastegui, A. (2006). A simple score for assessing stable chronic obstructive pulmonary disease. *QJM - An International Journal of Medicine*, 99:751–759.
- Esteban, C., Quintana, J. M., Moraza, J., Aburto, M., Egurrola, M., España, P. P., Pérez-Izquierdo, J., Aguirre, U., Aizpiri, S., and Capelastegui, A. (2009). Impact of hospitalisations for exacerbations of COPD on health-related quality of life. *Respiratory Medicine*, 103:1201–1208.
- Faraggi, D. and Simon, R. (1996). A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Statistics in Medicine*, 15:2203–2213.
- Fletcher, C. M., Elmes, P. C., Fairbairn, A. S., and Wood, C. H. (1959). The significance of respiratory symptoms and the diagnosis of chronic bronchitis in a working population. *British Medical Journal*, 2:257.
- Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the youden index and its associated cutoff point. *Biometrical Journal*, 47:458–472.
- Gelman, A. and Park, D. K. (2009). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician*, 63:1–8.
- Global Initiative for Chronic Obstructive Lung Disease (updated 2013). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. <http://www.goldcopd.com/>.

- Gönen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92:965–970.
- Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.
- Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.
- Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.
- Haroon, S., Adab, P., Riley, R. D., Marshall, T., Lancashire, R., and Jordan, R. E. (2015). Predicting risk of COPD in primary care: development and validation of a clinical risk score. *BMJ Open Respiratory Research*, 2:e000060.
- Harrell, F. E. (2001). *Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer.
- Harrell, F. E. (2015). *rms: Regression Modeling Strategies*. R package version 4.3-0.
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association*, 247:2543–2546.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.
- Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105.
- Hin, L. Y., Lau, T. K., Rogers, M. S., and Chang, M. Z. (1999). Dichotomization of continuous measurements using generalized additive modelling - application in predicting intrapartum caesarean delivery. *Statistics in Medicine*, 18:1101–1110.
- Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. New Jersey: Wiley.

- Jones, R. C., Donaldson, G. C., Chavannes, N. H., Kida, K., Dickson-Spillmann, M., Harding, S., Wedzicha, J. A., Price, D., and Hyland, M. E. (2009). Derivation and validation of a composite index of severity in chronic obstructive pulmonary disease: the DOSE index. *American Journal of Respiratory and Critical Care Medicine*, 180:1189–1195.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.
- Lee, D.-J. and Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, 11:49–69.
- Lim, B. L. and Kelly, A. M. (2010). A meta-analysis on the utility of peripheral venous blood gas analyses in exacerbations of chronic obstructive pulmonary disease in the emergency department. *European Journal of Emergency Medicine*, 17:246–248.
- Lindström, J. and Tuomilehto, J. (2003). The diabetes risk score a practical tool to predict type 2 diabetes risk. *Diabetes Care*, 26:725–731.
- López-Ratón, M., Rodríguez-Alvarez, M. X., Cadarso-Suárez, C., and Gude-Sampedro, F. (2014). OptimalCutpoints: An R package for selecting optimal cut-points in diagnostic tests. *Journal of Statistical Software*, 61.
- MacCallum, R. C., Zhang, S., Preacher, K. J., and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7:19.
- Make, B. J., Eriksson, G., Calverley, P. M., Jenkins, C. R., Postma, D. S., Peterson, S., Östlund, O., and Anzueto, A. (2015). A score to predict short-term risk of COPD exacerbations (SCOPEX). *International Journal of Chronic Obstructive Pulmonary Disease*, 10:201.
- Marin, J. M., Alfageme, I., Almagro, P., Casanova, C., Esteban, C., Soler-Cataluña, J. J., de Torres, J. P., Martínez-Cambor, P., Miravittles, M., Celli, B. R., and Soriano, J. B. (2013). Multicomponent indices to predict survival in COPD: the COCOMICS study. *European Respiratory Journal*, 42:323332.
- Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28:193–209.

- Mazumdar, M. and Glassman, J. R. (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine*, 19:113–132.
- Mazumdar, M., Smith, A., and Bacik, J. (2003). Methods for categorizing a prognostic variable in a multivariable setting. *Statistics in Medicine*, 22:559–571.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd ed.* London: Chapman & Hall.
- Mebane, W. R. and Sekhon, J. S. (2011). Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software*, 42:1–26.
- Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283–298.
- Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics*, 38:1011–1016.
- Mo, Q., Gonen, M., and Heller, G. (2012). *CPE: Concordance Probability Estimates in Survival Analysis*. R package version 1.4.4.
- Moise, A., Clement, B., and Raissis, M. (1988). A test for crossing receiver operating characteristic (ROC) curves. *Communications in Statistics-Theory and Methods*, 17:1985–2003.
- Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J. A., and Abdalla, S. (2013). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380:2197–2223.
- National Cholesterol Education Program (2002). Third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii) final report. *Circulation*, 106:3143–3421.
- O’Brien, S. M. (2004). Cutpoint selection for categorizing a continuous predictor. *Biometrics*, 60:504–509.
- Pencina, M. J., D’Agostino, R. B., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27:157–172.

- Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.
- Puhan, M. A., Garcia-Aymerich, J., Frey, M., ter Riet, G., Antó, J. M., Agustí, A. G., Gómez, F. P., Rodríguez-Roisín, R., Moons, K. G., Kessels, A. G., and Held, U. (2009). Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated bode index and the ado index. *The Lancet*, 374:704–711.
- Quintana, J. M., Esteban, C., Barrio, I., Garcia, S., Gonzalez, N., Arostegui, I., Lafuente, I., Bare, M., Blasco, J. A., Vidal, S., and I, G. T. (2011). The IRYSS-COPD appropriateness study: objectives, methodology, and description of the prospective cohort. *BMC Health Services Research*, 11:322.
- Quintana, J. M., Esteban, C., Unzurrunzaga, A., Garcia-Gutierrez, S., Gonzalez, N., Barrio, I., Arostegui, I., Lafuente, I., Bare, M., Fernandez-de Larrea, N., Vidal, S., and IRYSS-COPD, G. (2014a). Predictive score for mortality in patients with COPD exacerbations attending hospital emergency departments. *BMC Medicine*, 12:66.
- Quintana, J. M., Esteban, C., Unzurrunzaga, A., Garcia-Gutierrez, S., Gonzalez, N., Lafuente, I., Bare, M., de Larrea, N. F., Vidal, S., and IRYSS-COPD, G. (2014b). Prognostic severity scores for patients with COPD exacerbations attending emergency departments. *The International Journal of Tuberculosis and Lung Disease*, 18:1415–1420.
- Quintana, J. M., Garcia-Gutierrez, S., Aguirre, U., and Gonzalez-Hernandez, N. (2008). *Estándares de uso adecuado de tecnologías sanitarias. Creación de criterios explícitos de indicación de ingreso hospitalario en la exacerbación de EPOC*. Madrid: Agencia Laín Entralgo.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
- Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., van Weel, C., and Zielinski, J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American Journal of Respiratory and Critical Care Medicine*, 176:532–555.

- Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sample noisy curves. *Biometrics*, 57:253–259.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., and Mller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.
- Rota, M. and Antolini, L. (2014). Finding the optimal cut-point for gaussian and gamma distributed biomarkers. *Computational Statistics & Data Analysis*, 69:1–14.
- Rota, M., Antolini, L., and Valsecchi, M. G. (2015). Optimal cut-point definition in biomarkers: the case of censored failure time outcome. *BMC Medical Research Methodology*, 15:24.
- Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25:127–141.
- Schmid, M. and Potapov, S. (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, 31:2588–2609.
- Seshan, V. E., Gönen, M., and Begg, C. B. (2013). Comparing ROC curves derived from regression models. *Statistics in Medicine*, 32:1483–1493.
- Sima, C. S. and Gönen, M. (2013). Optimal cutpoint estimation with censored data. *Journal of Statistical Theory and Practice*, 7:345–359.
- Song, X. and Zhou, X.-H. (2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica*, 18:947–965.
- Steyerberg, E. W. (2009). *Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating*. New York: Springer.
- Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., and Group, P. (2013). Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine*, 10:e1001381.
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21:128.

- Taylor, J. M. G. and Yu, M. (2002). Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis*, 83:248–263.
- Teasdale, G. and Jennett, B. (1974). Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, 304:81–84.
- Therneau, T. (2014). *A Package for Survival Analysis in S*. R package version 2.37-7.
- Tsuruta, H. and Bax, L. (2006). Polychotomization of continuous variables in regression models based on the overall C index. *BMC Medical Informatics and Decision Making*, 6:41.
- Turner, E., Dobson, J., and Pocock, J. (2010). Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiologic Perspectives & Innovations*, 7:9.
- Uno, H., Cai, T., Pencina, M. J., D’Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30:1105–1117.
- Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102:527–537.
- Varadhan, R. and Gilbert, P. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, 32:1–26.
- Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics*, 56:1134–1138.
- Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83:835–848.
- Vermont, J., Bosson, J. L., François, P., Robert, C., Rueff, A., and Demongeot, J. (1991). Strategies for graphical threshold determination. *Computer Methods and Programs in Biomedicine*, 35:141–150.

- Vestbo, J., Hurd, S. S., Agustí, A. G., Jones, P. W., Vogelmeier, C., Anzueto, A., Barnes, P. J., Fabbri, L. M., Martinez, F. J., Nishimura, M., Stockley, R. A., Sin, D. D., and Rodriguez-Roisin, R. (2013). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American Journal of Respiratory and Critical Care Medicine*, 187:347–365.
- Vickers, A. J., Cronin, A. M., and Begg, C. B. (2011). One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology*, 11:13.
- Vickers, A. J., Cronin, A. M., Elkin, E. B., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*, 8:53.
- Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26:565–574.
- Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. Florida: CRC Press.
- Wilson, P. W. F., DAgostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97:1837–1847.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B*, 65:95–114.
- Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3:32–35.
- Zheng, Y. and Heagerty, P. J. (2004). Semiparametric estimation of time-dependent roc curves for longitudinal marker data. *Biostatistics*, 5:615–632.

Appendix A

Appendix A

In this appendix we demonstrate the result we mentioned in Chapter 5, above.

Result:

Let X be a predictor variable with a normal distribution separately in each of the populations defined by the outcome ($Y = 0$ and $Y = 1$), i.e., $X|Y = 0 \simeq N(\mu_0, \sigma_0)$ and $X|Y = 1 \simeq N(\mu_1, \sigma_1)$. When σ_0 and σ_1 are equal, X is linearly related to the log odds of the response.

Proof:

Due to Bayes and Total Probability theorems we have that:

$$\begin{aligned}
 P(Y = 1|X) &= \frac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 1)P(Y = 1) + P(X|Y = 0)P(Y = 0)} \\
 &= \frac{\frac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 0)P(Y = 0)}}{\frac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 0)P(Y = 0)} + 1}.
 \end{aligned} \tag{A.1}$$

If we denominate the constant fraction $\frac{P(Y=1)}{P(Y=0)}$ as ζ we can rewrite the expression in A.1 as follows:

$$\frac{\frac{P(X|Y = 1)}{P(X|Y = 0)}\zeta}{\frac{P(X|Y = 1)}{P(X|Y = 0)}\zeta + 1}. \tag{A.2}$$

On the other hand, we want to verify that the linear relationship between X and the *logit* function holds, that is, that the *logit* function can be written as a linear

function of X :

$$\text{logit}(p) = \ln \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = \beta_0 + \beta_1 X, \quad (\text{A.3})$$

where β_0 and β_1 are constants.

Using expression A.2 we can write the *logit* function as:

$$\begin{aligned} \ln \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} &= \frac{\frac{P(X|Y = 1)}{P(X|Y = 0)}^\zeta}{1 - \frac{P(X|Y = 1)}{P(X|Y = 0)}^\zeta} \\ &= \ln \frac{P(X|Y = 1)}{P(X|Y = 0)}^\zeta = \ln \zeta + \ln \frac{P(X|Y = 1)}{P(X|Y = 0)}. \end{aligned} \quad (\text{A.4})$$

Since $X|(Y = 0) \simeq N(\mu_0, \sigma_0)$ and $X|(Y = 1) \simeq N(\mu_1, \sigma_1)$ expression A.4 can be rewritten as follows:

$$\ln \zeta + \ln \left(\frac{\frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x - \mu_1)^2}{\sigma_1^2}}}{\frac{1}{\sigma_0 \sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x - \mu_0)^2}{\sigma_0^2}}} \right) = \ln \zeta + \ln \left(\frac{\sigma_0}{\sigma_1} e^{-\frac{(x - \mu_1)^2 \sigma_0^2 - (x - \mu_0)^2 \sigma_1^2}{2\sigma_1^2 \sigma_0^2}} \right). \quad (\text{A.5})$$

In the particular case in which $\sigma_1 = \sigma_0 = \sigma$ the squared term of x disappears and expression A.5 can be rewritten as shown in expression A.6:

$$\ln \zeta - \frac{1}{2\sigma^2}(\mu_1^2 - \mu_0^2) + \frac{1}{\sigma^2}(\mu_1 - \mu_0)x = \beta_0 + \beta_1 x. \quad (\text{A.6})$$