About the categorization of continuous variables in prediction models:

method development and implementation

Irantzu Barrio

Joint work with Inmaculada Arostegui and María Xosé Rodriguez-Álvarez

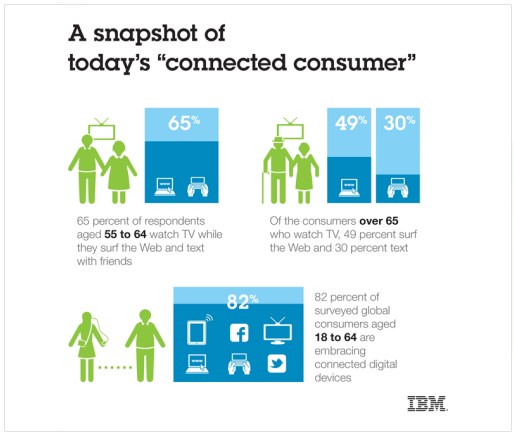
20 de diciembre de 2015

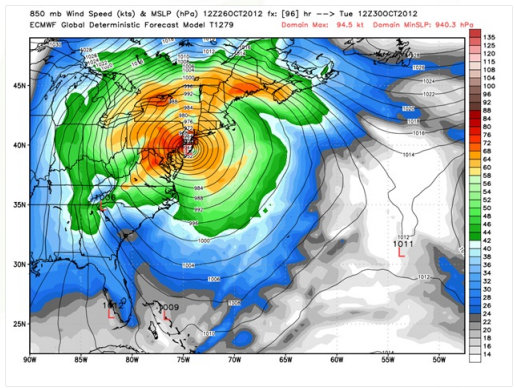
Outline

Motivation

BACKGROUND: Prediction Models

- Prediction models are important in various field: marketing, finance, medicine, meteorology...



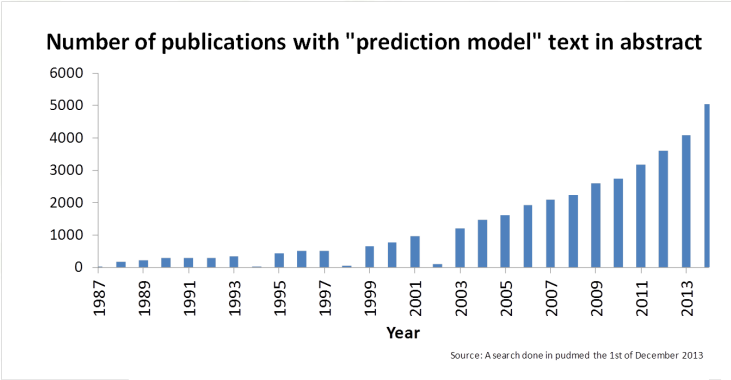


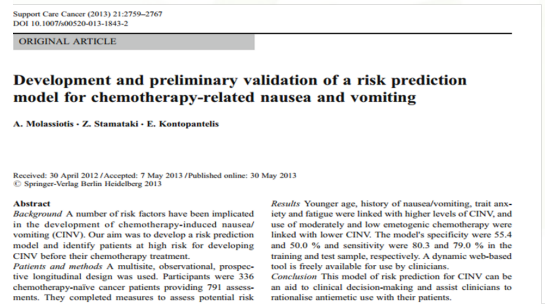
Source:
1) http://www.ibmbigdatahub.com/blog/movie-marketing-predictions-opening-weekend-box-office
2) http://swissnexsanfrancisco.org/Ourwork/events/emotionsandfinancialrisks
3) http://cliffmass.blogspot.com.es/2013/05/a-new-chapter-for-us-numerical-weather.html

Motivation

BACKGROUND: Prediction Models in Medicine

- Predictive models as support to decision making in health sciences
- It's use has increased exponentially in the last years
- The covariates (and their relationship with the response variable) are relevant in the development of prediction models




Support Care Cancer (2013) 21:2759–2767
DOI 10.1007/s00520-013-1843-2

ORIGINAL ARTICLE

Development and preliminary validation of a risk prediction model for chemotherapy-related nausea and vomiting

A. Molassiotis · Z. Stamataki · E. Kontopantelis

Received: 30 April 2012 / Accepted: 7 May 2013 / Published online: 30 May 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract
Background A number of risk factors have been implicated in the development of chemotherapy-induced nausea/vomiting (CINV). Our aim was to develop a risk prediction model and identify patients at high risk for developing CINV before their chemotherapy treatment.
Patients and methods A multisite, observational, prospective longitudinal design was used. Participants were 336 chemotherapy-naive cancer patients providing 791 assessments. They completed measures to assess potential risk
Results Younger age, history of nausea/vomiting, trait anxiety and fatigue were linked with higher levels of CINV, and use of moderately and low emetogenic chemotherapy were linked with lower CINV. The model's specificity were 55.4 and 50.0 % and sensitivity were 80.3 and 79.0 % in the training and test sample, respectively. A dynamic web-based tool is freely available for use by clinicians.
Conclusion This model of risk prediction for CINV can be an aid to clinical decision-making and assist clinicians to rationalise antiemetic use with their patients.


A New Risk Scheme to Predict Ischemic Stroke and Other Thromboembolism in Atrial Fibrillation: The ATRIA Study Stroke Risk Score

Daniel E. Singer, MD; Yuchiao Chang, PhD; Leila H. Borowsky, MPH; Margaret C. Fang, MD, MPH; Niela K. Pomernacki, RD; Natalia Udaltsova, PhD; Kristi Reynolds, PhD, MPH; Alan S. Go, MD

Background—More accurate and reliable stroke risk predic
patients with atrial fibrillation (AF). We developed a new AF
Factors in Atrial Fibrillation (ATRIA) AF cohort and external
inception AF cohort, ATRIA–Cardiovascular Research Netwo
Methods and Results—The derivation ATRIA cohort cons
person-years off warfarin and 685 thromboembolic events (
patients contributing 26 263 person-years off warfarin and
sex, diabetes mellitus, heart failure, hypertension, proteinuri
an age×prior stroke interaction term for the final model.
c-index in the ATRIA cohort was 0.73 (95% CI, 0.71 to 0.75
were considered. In the ATRIA-CVRN, c-indexes were 0.70 (9
severe events, respectively. The C-index was greater and net
the $CHADS_2$ or CHA_2DS_2-VASc scores.
Conclusions—The ATRIA stroke risk score performed bett
improvement in predicting severe events, which is of gre
decision for patients with AF and should provide a secure fo

The Journal of Urology
Available online 26 September 2013
In Press, Accepted Manuscript — Note to users

Measuring and Predicting Prostate Cancer Related Quality of Life Changes using the Expanded Prostate Cancer Index Composite for Clinical Practice (EPIC-CP)

Jonathan J. Chipman, MS[1], Martin G. Sanda, MD[2], Rodney L. Dunn, MS[3], John T. Wei, MD, MS[3], Mark S. Litwin, MD, MPH[4], Catrina M. Crociani, MPH[5], Meredith M. Regan, ScD[1, 6], Peter Chang, MD[5, 6],

RESEARCH ARTICLE Open Access

A decision rule to aid selection of patients with abdominal sepsis requiring a relaparotomy

Jordy JS Kiewiet[1*], Oddeke van Ruler[1], Marja A Boermeester[1] and Johannes B Reitsma[2]

Abstract
Background: Accurate and timely identification of patients in need of a relaparotomy is challenging since there are
a relaparotomy.
omized trial comparing surgical strategies for relaparotomy were used. Variables were
reports and common clinical sense and screened in a univariable regression analysis to
h the need for relaparotomy. Variables with the strongest association were considered
ich was constructed after backward elimination in a multivariable regression analysis.
of the model was expressed with the area under the curve (AUC). A cut-off analysis was
onsequences in clinical practice.
ighty-two patients were included; 46 were considered cases requiring a relaparotomy.
d containing 6 variables. This final model had an AUC of 0.80 indicating good
ever, acceptable sensitivity would require a low threshold for relaparotomy leading to
ative relaparotomies (63%). Therefore, the prediction model was incorporated in a
al until re-assessment and the use of Computed Tomography are related to the
prediction model that will provide a definite answer whether or not to perform a
s. However, our prediction model can be used to stratify patients on their underlying
monitoring of patients with abdominal sepsis in order to identify patients with

Motivation

CATEGORIZATION OF CONTINUOUS VARIABLES IN CLINICAL PREDICTION MODELS

- Despite the statistical recommendations not to categorize continuous variables (Royston et al., 2006), due to the loss of information and power in clinical practice physicians and healthcare managers claim to categorize continuous variables
- The aim of the categorization may be to separate patients on distinct risk groups so the prediction model mimics the decision-making process in daily clinical practice.
- One of the most common methodologies to develop prediction models is the logistic regression model
- There are approaches to categorize continuous variables based on clinical criteria and statistical methods; based on graphs or minimum p-value (Mazumdar and Glassman, 2000), but seek for a unique cut point.

Motivation

OBJECTIVE

- Obtain the best categorization of a continuous predictor in a logistic regression model or Cox Proportional Hazards model, so that the prediction ability of the model does not decrease significantly
- Prediction ability in a logistic regression model is considered as the ability to discriminate diseased patients from healthy patients.

Proposed Methodology

Let Y be a dichotomous response variable and X a continuous variable which we want to categorize

Proposed Methodology

Let Y be a dichotomous response variable and X a continuous variable which we want to categorize

PROPOSAL

- To categorize X in such a way that we obtain the best logistic predictive model for Y (highest AUC)
- The AUC is the area under the receiver operative characteristic (ROC) curve, which measures the discrimination ability of a logistic regression model.

Proposed Methodology

Let Y be a dichotomous response variable and X a continuous variable which we want to categorize

PROPOSAL

- To categorize X in such a way that we obtain the best logistic predictive model for Y (highest AUC)
- The AUC is the area under the receiver operative characteristic (ROC) curve, which measures the discrimination ability of a logistic regression model.

AddFor algorithm

- Looks sequentially for the k optimal cut points in a grid of size m

X_{cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

AddFor algorithm

- Looks sequentially for the k optimal cut points in a grid of size m
- Look for x_1 (In a grid of size m) in such a way that the AUC of the model is maximized

$$P(Y|X_{Cat_1}) = \frac{exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}{1 + exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}$$

X_{cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

AddFor algorithm

- Looks sequentially for the k optimal cut points in a grid of size m
- Look for x_1 (In a grid of size m) in such a way that the AUC of the model is maximized

$$P(Y|X_{Cat_1}) = \frac{exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}{1 + exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}$$

- We fix x_1 and look for x_2 $(x_2 \neq x_1)$ so that the AUC of the model is maximized

$$P(Y|X_{Cat_2}) = \frac{exp(\beta_0 + \beta_1 1_{X_{Cat_2}=1}) + \beta_2 1_{X_{Cat_2}=2})}{1 + exp(\beta_0 + \beta_1 1_{X_{Cat_2}=1}) + \beta_2 1_{X_{Cat_2}=2})}$$

X_{cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

AddFor algorithm

- Looks sequentially for the k optimal cut points in a grid of size m
- Look for x_1 (In a grid of size m) in such a way that the AUC of the model is maximized

$$P(Y|X_{Cat_1}) = \frac{exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}{1 + exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}$$

- We fix x_1 and look for x_2 $(x_2 \neq x_1)$ so that the AUC of the model is maximized

$$P(Y|X_{Cat_2}) = \frac{exp(\beta_0 + \beta_1 1_{X_{Cat_2}=1}) + \beta_2 1_{X_{Cat_2}=2})}{1 + exp(\beta_0 + \beta_1 1_{X_{Cat_2}=1}) + \beta_2 1_{X_{Cat_2}=2})}$$

- ...

X_{cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

AddFor algorithm

- Looks sequentially for the k optimal cut points in a grid of size m
- Look for x_1 (In a grid of size m) in such a way that the AUC of the model is maximized

$$P(Y|X_{Cat_1}) = \frac{exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}{1 + exp(\beta_0 + \beta_1 1_{X_{Cat_1}=1})}$$

- We fix x_1 and look for x_2 ($x_2 \neq x_1$) so that the AUC of the model is maximized

$$P(Y|X_{Cat_2}) = \frac{exp(\beta_0 + \beta_1 1_{X_{Cat_2}=1}) + \beta_2 1_{X_{Cat_2}=2})}{1 + exp(\beta_0 + \beta_1 1_{X_{Cat_2}=1}) + \beta_2 1_{X_{Cat_2}=2})}$$

- ...
- We repeat the process until we complete the vector of k cut points $v = (x_1, \ldots, x_k)$

X_{cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

Genetic algorithm

- Looks for the vector of k optimal cut points $v = (x_1, \ldots, x_k)$ by using genetic algorithms

X_{Cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

Genetic algorithm

- Looks for the vector of k optimal cut points $v = (x_1, \ldots, x_k)$ by using genetic algorithms
- The aim is to maximize the AUC of the model

$$P(Y = 1|X_{cat_k}) = \frac{\exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}{1 + \exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}$$

X_{Cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

Genetic algorithm

- Looks for the vector of k optimal cut points $v = (x_1, \ldots, x_k)$ by using genetic algorithms
- The aim is to maximize the AUC of the model

$$P(Y = 1|X_{cat_k}) = \frac{\exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}{1 + \exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}$$

- The arguments used in developing the genetic algorithm:

X_{Cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

Genetic algorithm

- Looks for the vector of k optimal cut points $v = (x_1, \ldots, x_k)$ by using genetic algorithms
- The aim is to maximize the AUC of the model

$$P(Y = 1|X_{cat_k}) = \frac{\exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}{1 + \exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}$$

- The arguments used in developing the genetic algorithm:
  - AUC function to be maximized

X_{Cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

Genetic algorithm

- Looks for the vector of k optimal cut points $v = (x_1, \ldots, x_k)$ by using genetic algorithms
- The aim is to maximize the AUC of the model

$$P(Y = 1|X_{cat_k}) = \frac{\exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}{1 + \exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}$$

- The arguments used in developing the genetic algorithm:
  - AUC function to be maximized
  - k number of parameters to be estimated

X_{Cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

Genetic algorithm

- Looks for the vector of k optimal cut points $v = (x_1, \ldots, x_k)$ by using genetic algorithms
- The aim is to maximize the AUC of the model

$$P(Y = 1|X_{cat_k}) = \frac{\exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}{1 + \exp(\beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}})}$$

- The arguments used in developing the genetic algorithm:
  - AUC function to be maximized
  - k number of parameters to be estimated
  - Range of the covariate X in which we look for the cut points

X_{Cat_k} the categorized variable taking $k+1$ values $(l = 0, \ldots, k)$

AUC optimism correction

Step 1. Categorise the predictor variable on the basis of the original sample $\{(x_i, y_i)\}_{i=1}^{N}$ and compute the corresponding AUC, $\widehat{AUC}_{app}$.
Step 2. For $b = 1, \ldots, B$, generate the bootstrap resample $\{(x_{ib}^*, y_{ib}^*)\}_{i=1}^{N}$ 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 categorized version of the predictor and compute the corresponding AUC, $\widehat{AUC}_{boot}^{b}$ for $b = 1, \ldots, B$.
Step 4. Obtain the predicted probabilities for the original sample based on the fitted logistic regression model obtained in Step 3 and compute the AUC, $\widehat{AUC}_{o}^{b}$ for $b = 1, \ldots, B$.

The optimism O is calculated as $O = \frac{1}{B}\sum_{b=1}^{B} |\widehat{AUC}_{boot}^{b} - \widehat{AUC}_{o}^{b}|$ and the bias corrected AUC is then computed as $\widehat{AUC}_{app} - O$.

Optimal number of cut points

- To determine the optimal number of cut points we propose an approach based on the difference between the bias-corrected AUCs obtained for $k = l$ and $k = l + 1$ cut points.

Optimal number of cut points

- To determine the optimal number of cut points we propose an approach 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. An extra cut point is considered to be needed as long as the CI does not contain the zero

Optimal number of cut points

- To determine the optimal number of cut points we propose an approach 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. An extra cut point is considered to be needed as long as the CI does not contain the zero
- In this case, bootstrap-based methods are proposed for constructing the CIs

Optimal number of cut points

1. For $v = 1, \ldots, V$, generate the bootstrap resample $\{(x^*_{iv}, y^*_{iv})\}_{i=1}^{N}$.
2. Compute the bias corrected AUC for the categorised variable for $k = l$ and $k = l + 1$ ($\widehat{AUC}^*_{l,v}$ and $\widehat{AUC}^*_{l+1,v}$).
3. Compute the difference between the bias-corrected AUCs obtained for $k = l + 1$ and $k = l$

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

The $(1 - \alpha)$ % limits for the CI for the difference are given by

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

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

Validation of the proposed methodology

- Let X be a predictor variable where $X_H = N(0, 1)$ for healthy patients and $X_D = N(1.5, 1)$ for diseased patients, and Y a dichotomous response variable (0 healthy, 1 diseased).

Validation of the proposed methodology

- Let X be a predictor variable where $X_H = N(0, 1)$ for healthy patients and $X_D = N(1.5, 1)$ for diseased patients, and Y a dichotomous response variable (0 healthy, 1 diseased).
- Since both groups have the same variance, we know which are the optimal cut points for X (Tsuruta and Bax, 2006)

Validation of the proposed methodology

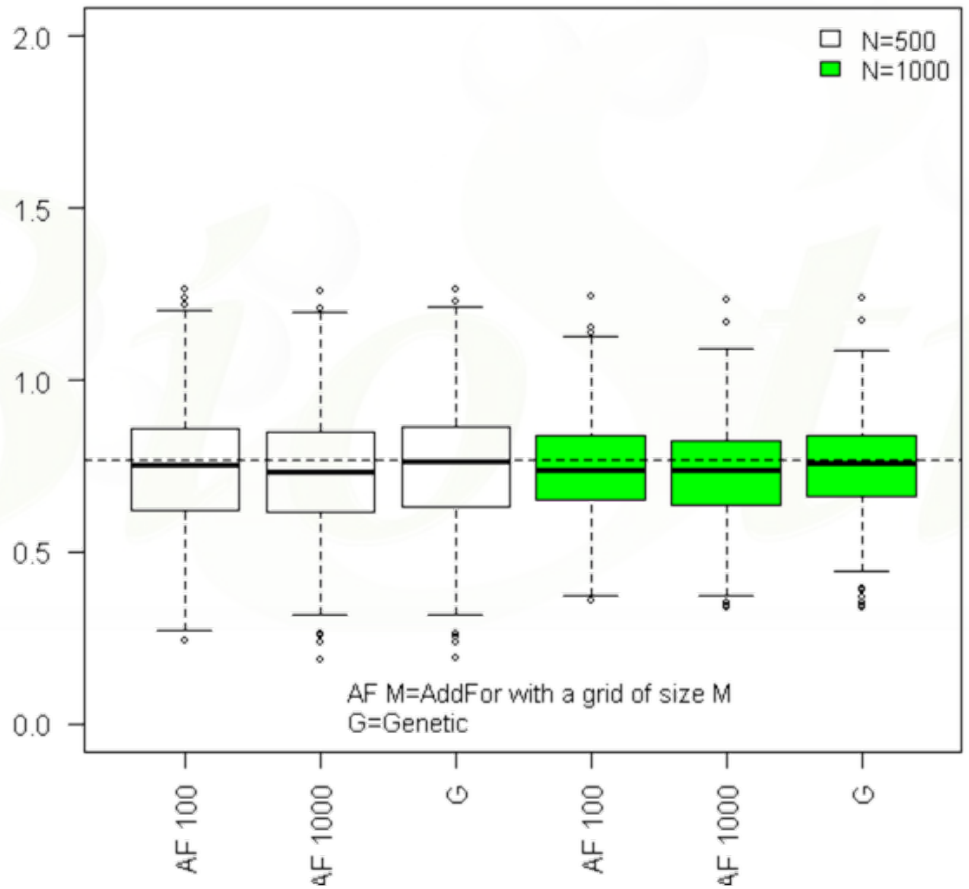
- Let X be a predictor variable where $X_H = N(0, 1)$ for healthy patients and $X_D = N(1.5, 1)$ for diseased patients, and Y a dichotomous response variable (0 healthy, 1 diseased).
- Since both groups have the same variance, we know which are the optimal cut points for X (Tsuruta and Bax, 2006)
- 500 data sets were generated, for 500 and 1000 sample sizes (n), with 1, 2 and 3 cut points (k)

Validation of the proposed methodology

- Let X be a predictor variable where $X_H = N(0, 1)$ for healthy patients and $X_D = N(1.5, 1)$ for diseased patients, and Y a dichotomous response variable (0 healthy, 1 diseased).
- Since both groups have the same variance, we know which are the optimal cut points for X (Tsuruta and Bax, 2006)
- 500 data sets were generated, for 500 and 1000 sample sizes (n), with 1, 2 and 3 cut points (k)
- For the Addfor 100 and 1000 grid sizes were used (m)

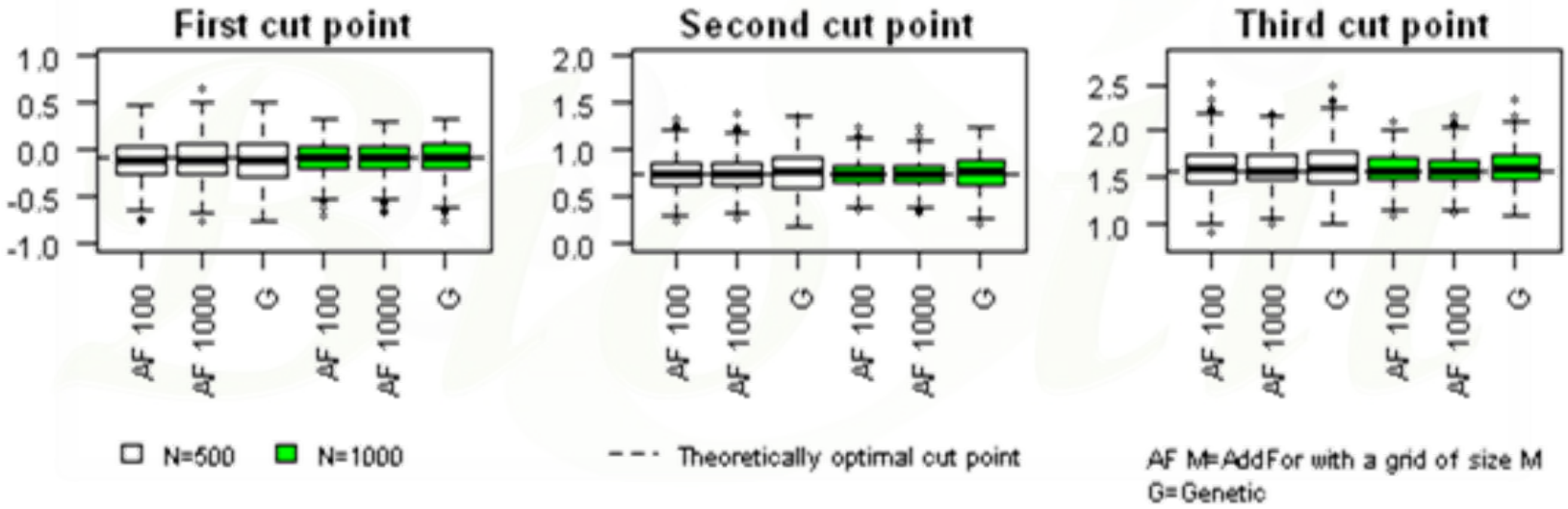
Theoretical Optimal cut-points: Results

Obtained cut points with the 500 simulations when $k = 1$ was chosen. Optimal cut point was $v = (0.77)$.



Theoretical Optimal cut-points: Results

Obtained cut points with the 500 simulations when $k = 3$ was chosen. Optimal cut points were $v = (-0.07, 0.75, 1.57)$.



CatPredi package

- We have implemented the proposed methods in the R package CatPredi

CatPredi package

- We have implemented the proposed methods in the R package CatPredi
- `catpredi.binary(formula, cat.var, cat.points = 1, data, method = c("addfor","genetic"), range = NULL, correct.AUC = TRUE , control = controlcatpredi.binary(), ...)`

CatPredi package

- We have implemented the proposed methods in the R package `CatPredi`
- `catpredi.binary(formula, cat.var, cat.points = 1, data, method = c("addfor","genetic"), range = NULL, correct.AUC = TRUE , control = controlcatpredi.binary(), ...)`
  - `formula`: $Y \sim 1$ or $Y \sim Z_1 + \ldots + Z_p$
  - `cat.var`: X continuous variable
  - `cat.points`: k number of cut points
  - `method`: *AddFor* or *Genetic* algorithm
  - `range`: range of X in which to search cut points
  - `control`: set control parameters

CatPredi package

- We have implemented the proposed methods in the R package `CatPredi`
- `catpredi.binary(formula, cat.var, cat.points = 1, data, method = c("addfor","genetic"), range = NULL, correct.AUC = TRUE , control = controlcatpredi.binary(), ...)`
  - `formula`: $Y \sim 1$ or $Y \sim Z_1 + \ldots + Z_p$
  - `cat.var`: X continuous variable
  - `cat.points`: k number of cut points
  - `method`: *AddFor* or *Genetic* algorithm
  - `range`: range of X in which to search cut points
  - `control`: set control parameters
    - `addfor.g`: m grid size for the *AddFor* algorithm. By default $m = 100$
    - `B`: number of bootstrap samples to correct the AUC bias. By default $B = 50$

Example

- Y response variable: 84 (4.1%) individuals died and 1966 did not (95.9%).

Example

- Y response variable: 84 (4.1%) individuals died and 1966 did not (95.9%).
- X a continuous variable we want to categorize.

Example

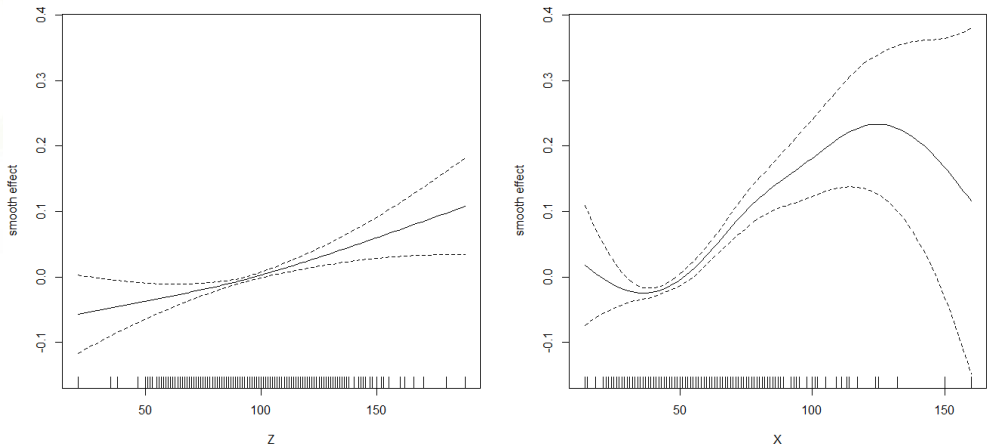
- Y response variable: 84 (4.1%) individuals died and 1966 did not (95.9%).
- X a continuous variable we want to categorize.
- Z a continuous predictor in the multivariate model.

Example

- Y response variable: 84 (4.1%) individuals died and 1966 did not (95.9%).
- X a continuous variable we want to categorize.
- Z a continuous predictor in the multivariate model.
- $logit(P(Y = 1|X_{cat_k})) = \beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}} + \beta_{k+1} Z$

Example

- Y response variable: 84 (4.1%) individuals died and 1966 did not (95.9%).
- X a continuous variable we want to categorize.
- Z a continuous predictor in the multivariate model.
- $logit(P(Y = 1|X_{cat_k})) = \beta_0 + \sum_{l=1}^{k} \beta_l 1_{\{X_{cat_k}=l\}} + \beta_{k+1} Z$



Example: `catpredi.binary`

- Search **2 optimal cut points** with the *AddFor* algorithm and 100 grid size.

Example: `catpredi.binary`

- Search **2 optimal cut points** with the *AddFor* algorithm and 100 grid size.
- `catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 2, data = datos, method = "addfor")`

Example: `catpredi.binary`

- Search **2 optimal cut points** with the *AddFor* algorithm and 100 grid size.
- `catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 2, data = datos, method = "addfor")`
- Search **3 optimal cut points** with the *AddFor* algorithm and 100 grid size.

Example: `catpredi.binary`

- Search **2 optimal cut points** with the *AddFor* algorithm and 100 grid size.
- `catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 2, data = datos, method = "addfor")`
- Search **3 optimal cut points** with the *AddFor* algorithm and 100 grid size.
- `catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 3, data = datos, method = "addfor")`

Example: `catpredi.binary`

- Search **2 optimal cut points** with the *AddFor* algorithm and 100 grid size.
- `catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 2, data = datos, method = "addfor")`
- Search **3 optimal cut points** with the *AddFor* algorithm and 100 grid size.
- `catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 3, data = datos, method = "addfor")`
- Compare 2 and 3 cut points

Example: `summary.catpredi.binary` for $k = 2$

```
> catX <-catpredi.binary(Y ~ Z, cat.var = "X", cat.points = 2, data = datos, method =
> plot(catX)
> summary(catX)

Call:
catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 2,
    data = datos, method = "addfor", range = NULL, correct.AUC = TRUE)


************************************************
Addfor Search Algorithm
************************************************

 Optimal cutpoints Optimal AUC Corrected AUC
 47.92             0.7539      NA
 62.67             0.7707      0.748


------------------------------------------------
Fitted model for the categorized predictor variable
------------------------------------------------


Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.543282   0.626245 -10.448  < 2e-16 ***
Z                 0.025077   0.005667   4.425 9.63e-06 ***
X_cat(47.9,62.7]  1.043084   0.299178   3.486 0.000489 ***
X_cat(62.7,160]   2.254750   0.281895   7.999 1.26e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Example: `summary.catpredi.binary` for $k = 3$

```
> catX.k3 <-catpredi.binary(Y ~ Z, cat.var = "X", cat.points = 3, data = datos, method = "addfor", range$
> summary(catX.k3)

Call:
catpredi.binary(formula = Y ~ Z, cat.var = "X", cat.points = 3,
    data = datos, method = "addfor", range = NULL, correct.AUC = TRUE)


************************************************
Addfor Search Algorithm
************************************************

 Optimal cutpoints Optimal AUC Corrected AUC
 47.92             0.7539      NA
 62.67             0.7707      NA
 19.90             0.7787      0.755


--------------------------------------------------
Fitted model for the categorized predictor variable
--------------------------------------------------


Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.040649   1.357169  -2.240  0.02506 *
Z                 0.025320   0.005688   4.452 8.52e-06 ***
X_cat(19.9,47.9] -3.573308   1.268797  -2.816  0.00486 **
X_cat(47.9,62.7] -2.485434   1.267248  -1.961  0.04985 *
X_cat(62.7,160]  -1.273543   1.263033  -1.008  0.31330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

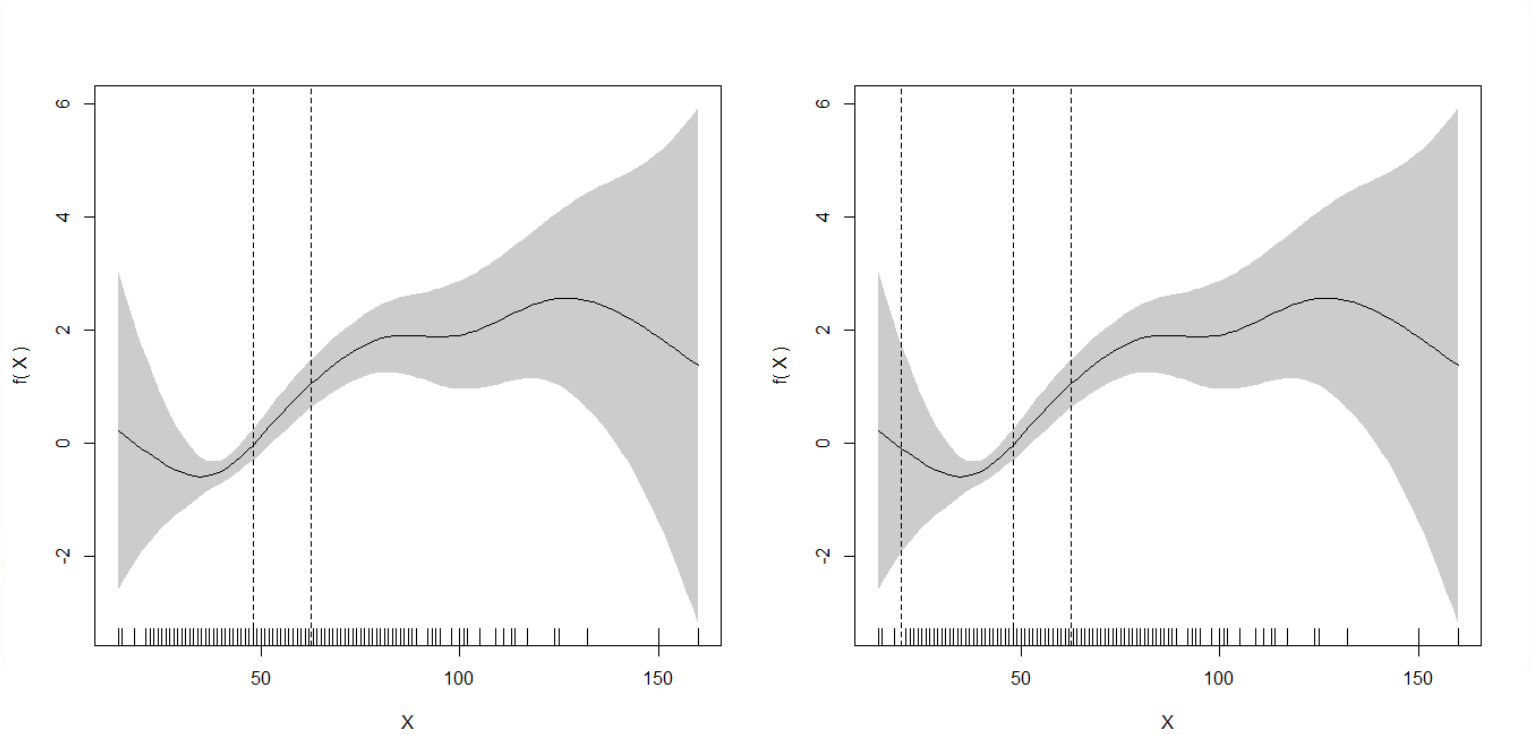
Example: comp.cutpoints.binary

```
> compare <-comp.cutpoints.binary(catX.k2, catX.k3, V = 100)
> compare


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

Bias corrected AUC difference:  0.0076
95% Bootstrap Confidence Interval: ( -0.006 , 0.0317 )
>
```

Example: `plot.catpredi.binary`



Conclusions

- We have developed a valid method to obtain optimal cut points in logistic prediction models
- We have implemented these methods in a R package witch is an easy to use tool
- This methodology has been extended to time to event outcomes and it is also implemented in the R package

References I

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(1)**,113 – 132.

Royston P, Altman D.G., Sauerbrei W. (2006)
Dichotomizing continuous predictors in multiple regression: a bad idea.
Statistics in Medicine, **25(1)**,127 – 41.

Tsuruta H., Bax L.(2006)
Polychotomization of continuous variables in regression models based on the overall C index.
BMC Medical Informatics and Decision Making, **6**,41.

Muchas Gracias