Binary regression: Total gain in positive and negative predictive values

Jens Klotsche*[,1,2], **Dietmar Ferger**[3], **David Leistner**[4], **Lars Pieper**[1,2], **Andreas M. Zeiher**[4], **Hans-Ulrich Wittchen**[1,2], and **Juergen Rehm**[1,2,5]

[1] Institute of Clinical Psychology and Psychotherapy, Technische Universitaet Dresden, Chemnitzer Straße 46, 01187 Dresden, Germany
[2] Center of Clinical Epidemiology and Longitudinal Studies (CELOS), Technische Universitaet Dresden, Chemnitzer Straße 46, 01187 Dresden, Germany
[3] Department of Mathematics, Technische Universitaet Dresden, Zellescher Weg 12–14, 01069 Dresden, Germany
[4] Department of Medicine III, Cardiology, Goethe-University Frankfurt, Theodor-Stern-Kai 7, 60596 Frankfurt, Germany
[5] Public Health and Regulatory Policy Section, Centre for Addiction and Mental Health, 33 Russell Street, Toronto, Ontario M5S 2S1, Canada

Received 5 May 2011; revised 20 August 2012; accepted 25 August 2012

Models that predict disease incidence or disease recurrence are attractive for clinicians as well as for patients. The usefulness of a risk prediction model is linked to the two questions whether the observed outcome is confirmed by the prediction and whether the risk prediction is accurate in predicting the future outcome, respectively. The first phrasing of the question is linked to considering sensitivity and specificity and the latter to the positive and negative predictive values. We present the measures of standardized total gain in positive and negative predictive values dealing with the performance or accuracy of the prediction model for a binary outcome. Both measures provide a useful tool for assessing the performance or accuracy of a set of predictor variables for the prediction of a binary outcome. This concept is a tool for evaluating the optimal prediction model in future research.

Keywords: Binary regression; Negative predictive value; Positive predictive value; Total gain.

1 Introduction

In the field of cardiology, intensive research is being conducted on new risk factors and biomarkers with the goal of improving prediction of a future cardiovascular event. Models that predict disease incidence or disease recurrence are attractive for clinicians as well as for patients. These models allow for the calculation of an individual risk value based on multiple risk factors. For example, a very popular risk calculator for the incidence of a cardiovascular event is the Framingham Risk Score (NCEP, 2001).

Various measures have been discussed for quantifying the predictive ability of a risk prediction model. A comprehensive overview about measures for summarizing and comparing the predictive capacity of predictor variables was recently provided by Gu and Pepe (2009). The usefulness of a risk prediction model is linked to the two questions whether the observed outcome is confirmed by the prediction and whether the risk prediction is accurate in predicting the future outcome, respectively.

*Corresponding author: e-mail: klotsche@psychologie.tu-dresden.de, Phone: +49-351-463-37462, Fax: +49-351-463-36984

The first form of the question considers the classification performance of a prediction model. A standard measure is the area under the receiver-operating characteristics curve (AUC) based on the conditional probabilities of sensitivity and specificity. Recently, Pencina et al. (2008) proposed the new complementary statistical measure of integrated discrimination improvement (IDI) for improved risk classification of two competing risk models. It summarizes a model's ability—in this case with an additionally added risk indicator—to correctly upgrade the predicted probability for an outcome in subjects who experienced a future event. Accordingly, it downgrades the model-based predicted probability for an outcome in subjects who did not.

The second question considers predictive accuracy more from a prevention point of view (Kraemer, 2008), for example, when one aims at the identification of high-risk persons in order to intervene properly before they are hit by the risk events. Standard measures for predictive accuracy are the positive and negative predictive values of a binary outcome and predictor variable. Both measures are clinically relevant and easily understood (Moskowitz and Pepe, 2004). Leisenring et al. (2000) discussed how to compare the two predictive values for two binary distributed predictor variables. Moskowitz and Pepe (2004) extended the approach. They considered and compared two continuously distributed variables in a marginal linear model.

The current paper explores the predictive accuracy of a prediction model. Section 2 provides a short description of the data used in this paper. The third section introduces the positive and negative predictive value curves. The measures of standardized total gain in positive and negative predictive values of a prediction model for a binary outcome are introduced in Section 4. Both estimators shared good final sample properties as highlighted by our simulation study in Section 5. The introduced estimators are applied to real data for investigating whether the incorporation of Nt-pro-BNP improved risk prediction on top of the standard risk factors in Section 6. Finally, a concluding discussion in Section 7 completes this paper.

2 Data for illustration

We illustrate the approach using data from the DETECT study (Diabetes Cardiovascular Risk-Evaluation: Targets and Essential Data for Commitment of Treatment; Wittchen et al., 2005). The methodology was applied to investigate the incorporation of Nt-pro-BNP for risk management on top of standard risk factors. We aim to illustrate our proposed methodology in a real data situation. This example does not claim to serve as a substantial contribution in the discussion about the adding of the biomarker Nt-pro-BNP to cardiovascular risk stratification. The real data example serves as a method illustration.

The DETECT study is a large nationally representative epidemiologic cross-sectional and prospective longitudinal study in German primary care settings (Wittchen et al., 2005). The health state of 55,518 patients was recorded on a target day in September 2003 using an extensive patient and physician questionnaire. A random subsample of 7519 individuals underwent an intensive standardized laboratory assessment and was followed up for a time period of five years. Our analyses were based on these longitudinally followed patients. The study was approved by the ethical commission of TU Dresden (AZ: EK149092003; date: September 16, 2003), and all patients gave their informed consent.

The main focus of the study was to examine the prevalence and comorbidity status of diabetes mellitus, hypertension, dyslipidemia, coronary heart disease, and their associated risk conditions. A total of 6826 individuals were successfully followed up for a time period of five years, constituting a follow-up response rate of 90.8%. Patients with any history of prior myocardial infarction, known cardiovascular disease, documented stroke, clinical signs of systolic or diastolic heart failure, and/or chronic kidney disease requiring hemodialysis ($N = 1181$) were excluded from analysis at baseline. Patients with no available laboratory values of Nt-pro-BNP ($N = 851$) were also excluded, resulting in a sample of 4794 individuals for our analyses. There were no statistically significant differences

Table 1 Baseline characteristics of standard risk factors and Nt-pro-BNP in subjects without history of cardiovascular events, heart failure and chronic kidney disease requiring hemodialysis at baseline followed up for five years ($N = 4794$).

Characteristics	N (%)[a]	Mean (SD)
Age, years		55.8 (13.7)
Female	2970 (62.0)	
Obesity[b]	1046 (22.6)	
Current smoker	950 (21.5)	
Exsmoker	1090 (24.6)	
Systolic blood pressure, mmHg		131.7 (18.1)
Diastolic blood pressure, mmHg		80.1 (9.7)
Hyperlipidemia	1332 (27.8)	
Diabetes mellitus	594 (12.4)	
Nt-pro-BNP, pg/mL		121.9 (303.3)

[a]All percentages refer to number of subjects with existing data.
[b]Obesity defined by BMI above 30 kg/m^2.

in baseline characteristics between patients with and without measurements of Nt-pro-BNP, and the dropout was independent from the availability of Nt-pro-BNP values.

The details of the standardized methods used in the DETECT study and the baseline characteristics of the cross-sectional study sample have been described elsewhere (Wittchen et al., 2005). The standard risk factors age, gender, systolic blood pressure, diastolic blood pressure, smoking status, hyperlipidemia, diabetes mellitus, and obesity were incorporated into the analyses. The baseline rates of the standard risk factors in the sample are reported in Table 1. Death and the causes of death were documented by the physicians or were obtained by the death statistics in 2008 (Leistner et al., 2012).

The mean age was 55.8 (SD = 13.7) with 62% women among the subjects without any history of cardiovascular disease, stroke, heart failure, and/or chronic kidney disease. A total of 109 (2.3%) subjects died during the follow-up period. There was a prevalence of 27.8% for hyperlipidemia and 12.4% for diabetes mellitus in the study sample. A total of 22.6% of the subjects met the criteria for obesity and 21.5% were current smokers.

3 Positive and negative predictive value curves

We consider a data set (Z_i, Y_i), $1 \leq i \leq n$ of a patients sample of size n with the aim to predict whether a subject will experience a particular event ($Y = 1$) or not ($Y = 0$). The prediction is based on a vector of D covariates $Z = (Z_d)_{d=1,\ldots,D}$, such as gender, age, blood pressure, or smoking status.

The risk of outcome Y is approximated by a risk prediction model defined as $risk(z) := P[Y = 1|Z = z]$. We assume that larger values of $risk(Z)$ are positively associated with the probability $P[Y = 1]$, otherwise the assumption could be conformed by an appropriate transformation of the predictor variable $risk(Z)$. The continuous cumulative distribution function $risk(Z)$ is denoted by F, $F(x) = P[risk(Z) \leq x]$. The assumption of continuity on F can be relaxed (Ferger and Klotsche, 2009; Ferger et al., 2012). In practice $risk(z)$ can be estimated, for example, by a parametric regression model such as a logistic regression model (Hosmer and Lemeshow, 2000) or a Cox proportional hazard model (Cox and Oakes, 1984) in case of censored data, respectively.

Using the notion of order statistics, see for example Stirzaker (1994), we denote by $risk(Z)_{1:n} \leq \cdots \leq risk(Z)_{n:n}$ the rank ordered sample of $risk(Z)_1, \ldots, risk(Z)_n$ and by $Y_{[i:n]}$ the i-th concomitant satisfying $Y_{[i:n]} = Y_j$ if and only if $risk(Z)_{i:n} = risk(Z)_j$.

3.1 Event rates in low- and high-risk groups

We summarize the established method for dealing with the predictive values by comparing model performance (Pencina et al., 2008) in the first part of this section. The positive predictive value is $ppv(\nu) := P[Y = 1|F(risk(Z)) > \nu]$ for any selected quantile $\nu \in [0, 1]$ of $F(risk(Z))$. Accordingly, the negative predictive value is defined by $npv(\nu) := P[Y = 0|F(risk(Z)) \leq \nu]$. The conditioning of the probability $P[Y = 1]$ on a partition of the unit interval $F(risk(Z)) > \nu$ has a relevant interpretation in terms of the sample. The sample $(F(risk(Z)_i, Y_i), 1 \leq i \leq n$ is divided into two disjoint subsamples by the quantile ν. The quantile ν classifies a proportion ν of the sample as negative (low-risk group) and a proportion of $1 - \nu$ as positive (high-risk group) based on the covariates Z. The probabilities for the outcome Y in the high- and low-risk groups are given by the predictive values $ppv(\nu)$ and $(1 - npv(\nu))$ for a fixed $\nu \in [0, 1]$, respectively. A good set of predictor variables Z results in a high-event probability $ppv(\nu)$ in the high-risk group. In contrast, a low-event probability $(1 - npv(\nu))$ is preferable in the low-risk group (Gu and Pepe, 2009).

Reasonable sample estimates for the quantities $ppv(\nu)$ and $(1 - npv(\nu))$ are given by

$$\widehat{ppv}_n(\nu) := \frac{\sum_{i=1}^{n} 1_{\{F_n(risk(z)_i)>\nu\}} Y_i}{\sum_{i=1}^{n} 1_{\{F_n(risk(z)_i)>\nu\}}} \tag{1}$$

and

$$1 - \widehat{npv}_n(\nu) := \frac{\sum_{i=1}^{n} 1_{\{F_n(risk(z)_i)\leq\nu\}} Y_i}{\sum_{i=1}^{n} 1_{\{F_n(risk(z)_i)\leq\nu\}}} \tag{2}$$

for a fixed value $\nu \in (0, 1)$.

The asymptotic behavior of these estimates, following the strong law of large numbers, can be seen to approach $ppv(t)$ and $1 - npv(t)$ as $n \to \infty$

$$ppv(t) = \begin{cases} \dfrac{(1 - npv(1-\pi))(1-\pi-t)) + ppv(1-\pi)(\pi)}{1-t}, & 0 \leq t \leq 1-\pi \\ ppv(1-\pi), & 1-\pi < t \leq 1 \end{cases}, \tag{3}$$

and

$$1 - npv(t) = \begin{cases} 1 - npv(1-\pi), & 0 \leq t \leq 1-\pi \\ \dfrac{(1 - npv(1-\pi))(1-\pi) + (t+\pi-1)ppv(1-\pi)}{t}, & 1-\pi < t \leq 1, \end{cases} \tag{4}$$

where $\pi := P[Y = 1]$ is the probability of a positive outcome. The proof is available on request.

3.2 Comparison of event rates in the real data example

Considering the DETECT data example, the risk for the event death by all causes for the two predictor variable sets (i) standard risk factors (Z^1) and (ii) Nt-pro-BNP on top of standard risk factors (Z^2) were estimated by logistic regression models. The quantile thresholds of $\nu_1 = 0.1$ and $\nu_2 = 0.2$ were considered, for example, 10% and 20% of the sample where estimated to be in the low-risk group based on Z^1 or Z^2. Applying the estimators (1) and (2), the event rates are $\widehat{ppv}(\nu_1) = 2.5\%$, $\widehat{ppv}(\nu_2) =$

2.8%, $1 - \widehat{npv}(\nu_1) = 0.2\%$, and $1 - \widehat{npv}(\nu_2) = 0.1\%$ for the predictor Z^2. In contrast, the predictor Z^1 resulted in $\widehat{ppv}(\nu_1) = 2.3\%$, $\widehat{ppv}(\nu_2) = 2.5\%$, $1 - \widehat{npv}(\nu_1) = 0.4\%$, and $1 - \widehat{npv}(\nu_2) = 0.1\%$. It appears that adding Nt-pro-BNP on top of standard risk factors improved prediction based on the predictive values (Gu and Pepe, 2009).

Pencina et al. (2008) suggested the comparison of predictive values for a set of meaningful thresholds as highlighted in our example above. However, this approach heavily depends on the selected quantile ν. Preferable in terms of model performance would be an index that overcomes this limitation being independent from a threshold.

3.3 The positive and negative predictive value curves

The positive and (1 − negative) predictive value curves were constructed by plotting $ppv(\nu)$ and $(1 - npv(\nu))$ as a function of quantiles ν, $\nu \in [0, 1]$. The solid lines display the predictive value curves in Fig. 1A and B for an artificial data example. Figure 1 assumes that (i) the probability $P[Y = 1]$ equals 0.4, (ii) $P[risk(Z)|Y = 0]$ follows a $N(0, 1)$-distribution, and (iii) $P[risk(Z)|Y = 1]$ follows a $N(2, 1)$ distribution.

The event rates for the high- and low-risk groups can directly be obtained from the predictive value curves for all possible quantile thresholds.

4 Total gain in positive and negative predictive values

We follow the idea of Bura and Gastwirth (2001) and define total gain in positive and negative predictive values as a threshold independent index for model performance.

The positive predictive value equals the probability $\pi := P[Y = 1]$ and the negative predictive value equals $1 - \pi = P[Y = 0]$ for a predictor variable Z providing no information in respect to the outcome Y. The larger the areas sandwiched between the horizontal dotted lines and the solid lines in Fig. 1A and B (light gray shadowed area), the more information is given about the model-based positive and negative predictive values provided by the predictor variable Z. Therefore, the area between the horizontal line π and the curve $ppv(t)$ is a reasonable estimator for the total gain in positive predictive value (see Fig. 1A). A similar measure can be defined for the negative predictive value, it is the area between the horizontal line $1 - \pi$ and $npv(t)$ (see Fig. 1B, graphs are mirrored because of presenting $1 - npv(t)$). The proposed measure of total gain are defined by $TG_{ppv} := \int_{[0,1]}(ppv(t) - \pi)\, \mathrm{d}t$ and $TG_{npv} := \int_{[0,1]}(npv(t) + \pi - 1)\, \mathrm{d}t$, respectively. The summary indices TG_{ppv} and TG_{npv} are appealing, they could directly be visualized from Fig. 1A and B (light gray shadowed area).

Estimators of TG_{ppv} and TG_{npv} are given by

$$\widehat{TG_{ppv}} = \sum_{k=0}^{n-1} \left(\frac{1}{n(n-k)} \sum_{i=k+1}^{n} Y_{[i:n]} \right) - \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{5}$$

and

$$\widehat{TG_{npv}} = \sum_{k=1}^{n} \left(\frac{1}{nk} \sum_{i=1}^{k} \left(1 - Y_{[i:n]}\right) \right) + \frac{1}{n} \sum_{i=1}^{n} Y_i - 1. \tag{6}$$

The derivation of $\widehat{TG_{ppv}}$ and $\widehat{TG_{npv}}$ is shown in the Appendix. It is notable that the area sandwiched by the negative predictive value curve and the horizontal line $1 - \pi$ equals the area between the line π and the (1 − negative predictive value) curve.

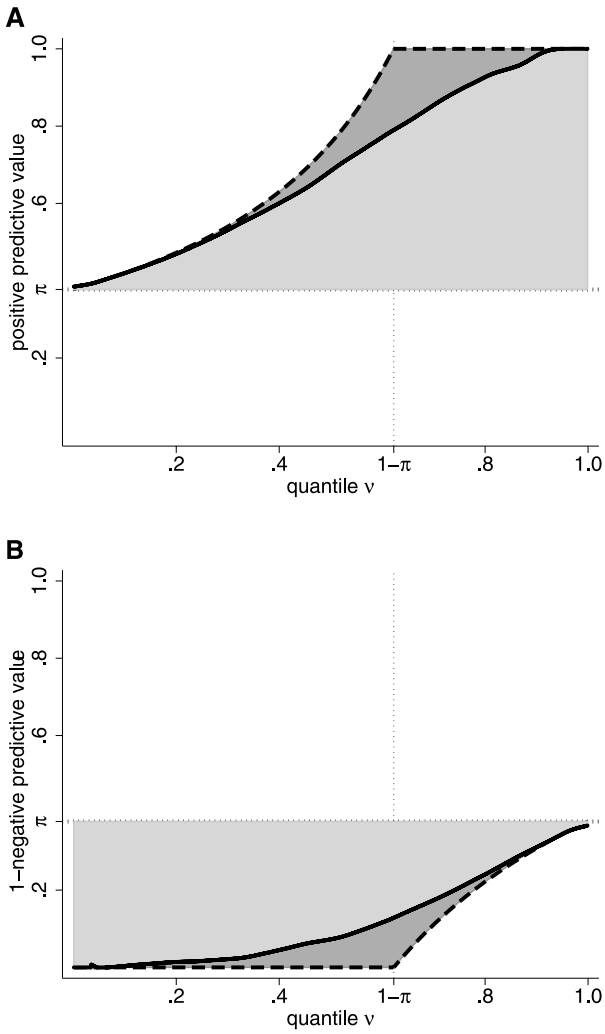

Figure 1 Positive predictive values (A) and (1 − negative) predictive values (B) as a function of quantiles $\nu \in [0, 1]$ in hypothetical data (solid line: positive predictive values (A) and (1 − negative) predictive values (B); dashed line: upper bound of positive predictive values (A) and lower bound of (1 − negative) predictive values (B) for a given $\pi = P[Y = 1]$; dotted line: $\pi = P[Y = 1]$; light gray shadowed area: (A) $\widehat{TG_{ppv}}$ and (B) $\widehat{TG_{npv}}$; gray shadowed area plus light gray shadowed area: maximum value of total gain.

4.1 Standardized total gain in positive and negative predictive values

Predictive values depend on the probability of the outcome Y (Altman and Bland, 1994). This property carries over the areas under the predictive value curves. Therefore, their magnitudes cannot be generalized and compared beyond a particular study. A standardization of TG_{ppv} and TG_{npv} by their maximum values provides a measure that can be compared across different studies.

The maximum value of total gain is obtained in a model with perfect prediction of Y by the predictor Z, a complete separation of Y by the values of Z. It is $P[Y = 1|F(risk(Z)) \leq 1 - \pi] = 0$ and $P[Y = 1|F(risk(Z)) > 1 - \pi] = 1$. The quantile threshold $1 - \pi$ refers to the so-called split point or change point in Ferger and Klotsche (2009). That threshold divides the sample into a high- and a

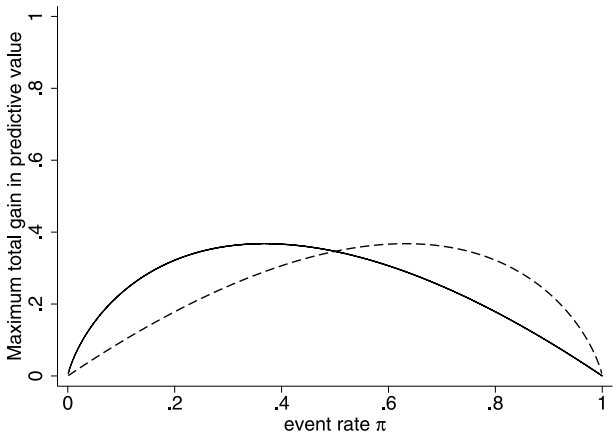

Figure 2 Maximum total gain in predictive values as a function of π in the artificial data example (solid line: maximum total gain in positive predictive value; dashed line: maximum total gain in negative predictive value).

low-risk group as described in Section 3.1. Both groups have a maximum difference in the event probability at $1-\pi$ (Ferger and Klotsche, 2009). In case of perfect prediction the quantity $(1-npv(1-\pi))$ equals zero and $ppv(1-\pi)$ equals one in (3) and (4). The predictive values are given for a perfect prediction model by

$$ppv_{max}(t) := \begin{cases} \dfrac{\pi}{1-t}, & 0 \le t \le 1-\pi \\ 1 \quad , & 1-\pi < t \le 1 \end{cases} \tag{7}$$

and

$$1-npv_{max}(t) := \begin{cases} 0 \quad , & 0 \le t \le 1-\pi \\ \dfrac{t+\pi-1}{t}, & 1-\pi < t \le 1 \end{cases} , \tag{8}$$

following from the limit variables presented in (3) and (4).

The maximum total gain in the positive and (1 − negative) predictive value curves are given by $\int_{[0,1]}(ppv_{max}(t)-\pi)\ \mathrm{d}t = -ln(\pi)\pi$ and $\int_{[0,1]}(\pi-(1-npv_{max}(t)))\mathrm{d}t = \ln(1-\pi)\pi - \ln(1-\pi)$. They are highlighted by the area sandwiched between the horizontal line π and the dashed lines in Fig. 1A and B (light gray shadowed area plus gray shadowed area). The functional dependency of the maximum total gain in predictive values and the event rate π is displayed in Fig. 2 for our artificial data example.

With the considerations above we can define the standardized total gain in positive (TG^{std}_{ppv}) and negative predictive values (TG^{std}_{npv}) as follows:

$$TG^{std}_{ppv} := \frac{TG_{ppv}}{-\ln(\pi)\pi} \quad \text{and} \quad TG^{std}_{npv} := \frac{TG_{npv}}{\ln(1-\pi)\pi - \ln(1-\pi)}.$$

Finally, estimates of the standardized measures of total gain can be obtained by

$$\widehat{TG^{std}_{ppv}} = \frac{\widehat{TG_{ppv}}}{-\ln(\hat{\pi})\hat{\pi}} \tag{9}$$

and

$$\widehat{TG^{std}_{npv}} = \frac{\widehat{TG_{npv}}}{\ln(1-\hat{\pi})\hat{\pi} - \ln(1-\hat{\pi})}, \tag{10}$$

where the probability for a positive outcome can be calculated by $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Both measures provide a tool for assessing the performance or accuracy of a set of predictor variables Z. The total gains TG^{std}_{ppv} and TG^{std}_{npv} equal zero for a set of predictor variables providing no information in respect to the outcome Y. On the other side, both measures equal one for a perfect prediction of Y.

4.2 Comparing two risk prediction models using total gain in predictive values

A hypothesis test is of interest in real data application comparing the standardized total gains in predictive values for two sets of predictor variables. We consider two risk models 1 and 2 estimated in the same population resulting in paired observations $(risk(Z^1)_i, risk(Z^2)_i, Y_i)$ for $i = 1, \ldots, n$. The standardized total gain in predictive values $\widehat{TG^{std,1}_{ppv}}$, $\widehat{TG^{std,1}_{npv}}$, and $\widehat{TG^{std,2}_{ppv}}$, $\widehat{TG^{std,2}_{npv}}$ were calculated for risk models 1 and 2 as defined in (9) and (10), respectively. To test whether $H_0 : TG^{std,1}_{ppv} = TG^{std,2}_{ppv}$ the test statistic

$$\chi_{ppv} := \frac{n\left(\widehat{TG^{std,1}_{ppv}} - \widehat{TG^{std,2}_{ppv}}\right)^2}{\widehat{V}_{ppv}} \tag{11}$$

is used. $\widehat{V}_{ppv}$ is the variance estimate of the difference in total gains for the two models. The variance estimate $\widehat{V}_{ppv}$ can be calculated by a bootstrap. We sample with replacement n observations from $(risk(Z^1)_i, risk(Z^2)_i, Y_i)$. The empirical variance estimate (Efron and Tibshirani, 1993) is computed based on bootstrap samples $\sqrt{n}(\widehat{TG^{std,1}_{ppv,b}} - \widehat{TG^{std,2}_{ppv,b}})$, $b = 1, \ldots, B$. Similar arguments apply to $\chi_{npv} := \frac{n(\widehat{TG^{std,1}_{npv}} - \widehat{TG^{std,2}_{npv}})^2}{\widehat{V}_{npv}}$ to test whether $H_0 : TG^{std,1}_{npv} = TG^{std,2}_{npv}$. It can be shown that the estimators $\widehat{TG^{std}_{ppv}}$ and $\widehat{TG^{std}_{npv}}$ in (9) and (10) are asymptotically normally distributed. Then both test statistics χ_{ppv} and χ_{npv} follow a central χ^2_1 distribution under H_0 (Pfeiffer and Gail, 2011). Confidence intervals for the differences can be obtained by $(\widehat{TG^{std,1}_{ppv}} - \widehat{TG^{std,2}_{ppv}}) \pm 1.96(\widehat{V}_{ppv})^{1/2}$ and $(\widehat{TG^{std,1}_{npv}} - \widehat{TG^{std,2}_{npv}}) \pm 1.96(\widehat{V}_{npv})^{1/2}$. Confidence intervals for $\widehat{TG^{std}_{ppv}}$ and $\widehat{TG^{std}_{npv}}$ can be computed as $\widehat{TG^{std}_{ppv}} \pm 1.96(\hat{\sigma}^2_{ppv})^{1/2}$ and $\widehat{TG^{std}_{npv}} \pm 1.96(\hat{\sigma}^2_{npv})^{1/2}$, where $\hat{\sigma}^2_{ppv}$ and $\hat{\sigma}^2_{npv}$ are the empirical bootstrap variance estimates of (9) and (10).

5 Simulations

We conducted a simulation study to investigate the numerical performance of the nonparametric estimators $\widehat{TG_{ppv}}$ and $\widehat{TG^{std}_{ppv}}$, the coverage of the 95% confidence intervals, size and power function of the significance test introduced in Section 4.2. The log odds of the outcome variable Y was simulated by a logistic risk model in which the two covariates Z_1 and Z_2 were incorporated, $\text{logit}P[Y = 1|Z_1, Z_2] =$

Table 2 Results of the simulation study.

			$\widehat{TG_{ppv}}$			$\widehat{TG^{std}_{ppv}}$		
β_0; β_1	π; TG_{ppv}; TG^{std}_{ppv}	n	Bias	RMSE	Coverage	Bias	RMSE	Coverage
−6.8; 0.09	0.29; 0.044; 0.134	50	0.013	0.068	0.946	0.028	0.191	0.948
		100	0.007	0.047	0.957	0.015	0.133	0.955
		500	0.003	0.022	0.935	0.007	0.061	0.937
		1000	0.003	0.015	0.939	0.003	0.043	0.942
−8.73; 0.09	0.06; 0.015; 0.088	50	0.012	0.040	0.932	0.016	0.205	0.891
		100	0.005	0.027	0.947	0.009	0.157	0.938
		500	0.002	0.012	0.946	0.004	0.081	0.953
		1000	0.001	0.009	0.956	0.001	0.075	0.952
−7.2; 0.09	0.21; 0.044; 0.133	50	0.013	0.058	0.957	0.014	0.161	0.955
		100	0.007	0.044	0.941	0.011	0.146	0.932
		500	0.004	0.021	0.942	0.006	0.061	0.944
		1000	0.004	0.015	0.958	0.004	0.039	0.963

$\beta_0 + \beta_1 Z_1 + \beta_2 Z_2$. The distribution of the covariate Z_1 was simulated as $N(65, 10)$ and Z_2 as $N(0, 1)$. The simulation study is based on $N = 1000$ independent Monte Carlo replications. Sample sizes of $n \in \{50, 100, 500, 1000\}$ were considered and the bootstrap variance estimates were based on $B = 1000$ bootstrap samples.

5.1 Properties of the estimators $\widehat{TG_{ppv}}$ and $\widehat{TG^{std}_{ppv}}$

The regression coefficient β_2 was set to zero in the first part of the simulation study investigating the properties of the estimators and the coverage of the 95% confidence intervals. The detailed results are reported in Table 2. The prevalence π ranged between 0.06 and 0.29 in the three considered scenarios. The root mean square error (RMSE) was comparable across all simulation scenarios for a fixed sample size. The RMSE decreased by increasing the sample size n. There existed a remarkable bias in the estimate for a sample size of $n = 50$. The estimated coverage of the 95% confidence intervals was close to 0.95 for all parameter settings, even for a small sample size of $n = 50$. Finally, both estimators $\widehat{TG_{ppv}}$ and $\widehat{TG^{std}_{ppv}}$ were asymptotically normal distributed (data not shown). Similar results were obtained for the estimators $\widehat{TG_{npv}}$ and $\widehat{TG^{std}_{npv}}$.

5.2 Size and power of the hypothesis test

We compared the size and power of the hypothesis test, whether adding a predictor variable to an established risk model increases the standardized total gain in positive and negative predictive values. We considered risk model 1 including predictor variable Z_1, $\text{logit}P[Y = 1|Z_1] = -6.8 + 0.09Z_1$. Model 2 included both predictor variables Z_1 and Z_2, $\text{logit}P[Y = 1|Z_1, Z_2] = -6.8 + 0.09Z_1 + \beta_2 Z_2$. The regression coefficient β_2 ranges from 0 (noninformative) to 0.24. The values of 0.12 and 0.24 for β_2 correspond to an increase in standardized total gain in positive predictive value of 2.2% and 6.4%, respectively. If the statistical test performs well, the simulated significance levels should be close to the selected significance level of $\alpha = 0.05$ under H_0, for example, $\beta_2 = 0$. It appeared that the simulated significance levels were approximately 0.061 for $n = 100$, 0.053 for $n = 250$, and 0.048 for $n = 1000$. Figure 3 displays the power functions for sample sizes $n \in \{100, 250, 1000\}$. The statistical power achieved 95% for a 2.2% increase in $\widehat{TG^{std}_{ppv}}$ for $n = 1000$ and 69% for $n = 100$.

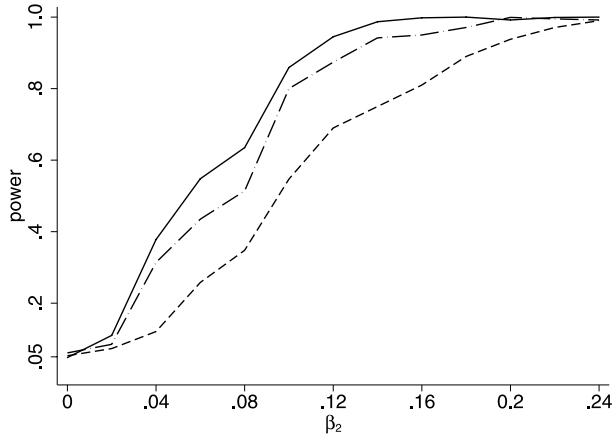

Figure 3 Power functions for the hypothesis test (solid line: $n = 1000$; long-dash dotted line: $n = 250$; dashed line: $n = 100$).

5.3 Discrimination performance

This section is used to present briefly the statistical measures net reclassification improvement (NRI) and integrated discrimination performance (IDI). Pencina et al. (2008) suggested the two measures of NRI and IDI for investigating whether adding a predictor variable improves discrimination performance of an established risk prediction model. The definition of IDI is based on sensitivity and specificity. It complements our suggested two measures $\widehat{TG_{ppv}}$ and $\widehat{TG_{npv}}$.

We label the established predictor variables with Z^1 and the additional predictor variable on top of established predictors by Z^2. The resulting two conditional probabilities $risk(Z^1)$ and $risk(Z^2)$ are categorized into clinically meaningful ordinal categories of risk. The resulting classifications are cross-tabulated. Then the NRI is defined by

$$NRI := (P[up \mid D = 1] - P[down \mid D = 1]) - (P[up \mid D = 0] - P[down \mid D = 0]),$$

where up defines upward movement into a higher risk category based on Z^2 and $down$ the downward movement into a lower risk category. An improved prediction model based on Z^2 would result in an upward movement for cases ($Y = 1$) and a downward movement for controls ($Y = 0$). The second measure suggested by Pencina et al. (2008) is the IDI. It is built on the integral of sensitivity and specificity over all possible threshold values. The IDI can be written by $IDI = (IS_n - IS_o) - (IP_n - IP_o)$, where IS and IP are the integral of sensitivity and specificity over all possible threshold values. The subscript n refers to the model based on predictor variables Z^2 and subscript o to the model based on predictor variables Z^1.

6 Real data application

The real data application introduced in Section 2 is continued in this paragraph. All estimators were implemented in the statistical software package Stata 11.1 (StataCorp., 2009) and are available on request. Statistical analyses were also conducted in STATA 11.1.

The median baseline value of Nt-pro-BNP was elevated in patients who died (156.5 pg/ml, interquartile range 76.4, 421.1 pg/ml vs. 56.7 pg/ml, interquartile range 28.7, 113.9 pg/ml). We investigated the association of the predictor variables Nt-pro-BNP and standard risk factors with the endpoint by

Table 3 Parameter estimates for the risk model predicting death by all causes in subjects without history of cardiovascular events, heart failure, and chronic kidney disease requiring hemodialysis at baseline followed up for five years ($N = 4794$).

	Death by all causes ($N = 109$)		
	Estimate	95% CI[a]	p value
Odds ratio for 1-SD increase Nt-pro-BNP			
Crude[b)]	1.34	1.20–1.51	0.001
Adjusted for standard risk factors[b),c)]	1.15	1.04–1.27	0.005
Area under receiver operating characteristic curve[b)]			
Nt-pro-BNP	0.750	0.703–0.797	0.001
Standard risk factors[c),d)]	0.816	0.780–0.853	0.001
Standard risk factors[c)] plus Nt-pro-BNP	0.822	0.784–0.856	0.000
Estimated difference[d),e)]	0.011	0.002–0.020	0.012
Standardized total gain in positive predictive value[b)]			
$\widehat{TG}^{std}_{ppv}$ with standard risk factors[d)]	0.376	0.304–0.448	0.013
$\widehat{TG}^{std}_{ppv}$ with standard risk factors[d)] plus Nt-pro-BNP	0.424	0.349–0.497	0.001
Estimated difference[e)]	0.048	0.003–0.085	0.020
Standardized total gain in negative predictive value[b)]			
$\widehat{TG}^{std}_{npv}$ with standard risk factors[d)]	0.758	0.698–0.819	0.008
$\widehat{TG}^{std}_{npv}$ with standard risk factors[d)] plus Nt-pro-BNP	0.778	0.719–0.838	0.018
Estimated difference[d),e)]	0.028	−0.006–0.049	0.169
Improved risk classification[b)]			
Net reclassification improvement (*NRI*)	16.03	3.16–30.45	0.010
Integrated discrimination improvement (*IDI*)[f)]	1.64	0.83–4.80	0.002

[a)]CI: confidence interval.
[b)]Odds ratio for a 1-SD increase Nt-pro-BNP: test of odds ratio is equal to 1; area under receiver operating characteristic curve: test of AUC is equal to 0.5; standardized total gain in positive predictive value: test of statistic is equal to 0; standardized total gain in negative predictive value: test of statistic is equal to 0; improved risk classification: test of statistic is equal to 0.
[c)]Odds ratio for increase of 1 standard deviation.
[d)]Age, gender, genetic disposition, obesity, smoking, systolic blood pressure, diastolic blood pressure hyperlipidemia, diabetes mellitus.
[e)]Estimated difference with addition of Nt-pro-BNP to standard risk factors.
[f)]Difference of averaged increase in sensitivity and in 1 − specificity.

logistic regression analyses (Hosmer and Lemeshow, 2000). The AUC (with 95% confidence intervals) were estimated after fitting the logistic regression model for classification performance. The crude odds ratio for a one standard deviation increment of Nt-pro-BNP was 1.34 (95% CI: 1.20, 1.51) for death by all causes. This association was also statistically significant after adjusting for standard risk factors (OR = 1.15, 95% CI: 1.04, 1.27). The AUC significantly increased for prediction of death by all causes when adding Nt-pro-BNP into a model with the standard risk factors (estimated difference = 0.011, 95% CI: 0.002, 0.020).

The results of the comparison of the two risk models are reported in Table 3. Empirical variance estimates for the standardized total gain in predictive values and the difference between the two risk prediction models (see (11) for testing whether adding of Nt-pro-BNP on top of standard risk factors improves significantly the standardized total gain in predictive values) were computed by

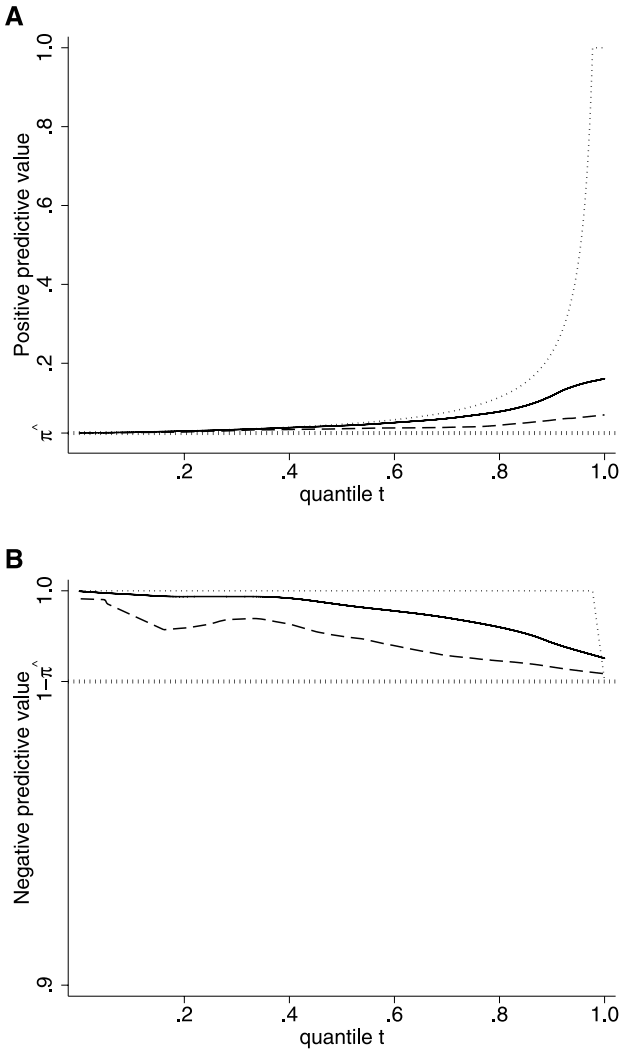

Figure 4 The positive predictive value (A) and negative predictive value (B) as a function of percentile thresholds $t \in [0, 1]$ for the prediction of death by all causes in the DETECT study (solid line: standard risk factors plus Nt-pro-BNP as predictor variables; dashed line: standard risk factors as predictor variables; dotted line: upper bound of positive predictive value for predicting death by all causes ($\hat{\pi} = 2.3\%$)).

bootstrapping (Efron and Tibshirani, 1993) as described in Section 4.2. The number of bootstrap replications was 1000 (Carpenter and Bithell, 2000). The maximum total gain in positive predictive value for a perfect prediction is 0.086 for the event death by all causes with a prevalence of 2.3%. The risk model establishing the standard risk factors resulted in 37.6% and the model adding Nt-pro-BNP on top of standard risk factors resulted in 42.4% of the maximum total gain in positive predictive value (Fig. 4A). The variance estimate $\widehat{V}_{ppv}$ of the difference of total gain in positive predictive values was $\widehat{V}_{ppv} = 2.01$. The difference is statistically significant (difference $= 4.8\%$, $\chi_{ppv} = 5.38$, $p = 0.020$). The standardized total gain in (1 − negative) predictive value was not significantly improved at the 5% significance level (75.8% vs. 77.8%, $\chi_{npv} = 1.80$, $p = 0.169$, Fig. 4B).

Finally, the *IDI* and *NRI* were calculated. The *IDI* was 1.64 ($p < 0.01$) for death by all causes when adding Nt-pro-BNP into the model with standard risk factors. The improvement in averaged sensitivity was 1.60 ($p < 0.01$) and improvement in averaged specificity 0.04 ($p = 0.07$).

7 Discussion

We introduced estimators for the total gain in positive and negative predictive values for a binary outcome provided by a set of predictor variables. Both estimators are based on the area under the positive and negative predictive value curves. Total gain in the predictive values is linked to the question whether a set of predictor variables increases the accuracy in prediction of a binary outcome. The approach of Pencina et al. (2008) does not address this important question. This concept provides an additional useful tool for evaluating the optimal prediction model in future research.

The nonparametric estimators for $\widehat{TG^{std}_{ppv}}$ and $\widehat{TG^{std}_{npv}}$ can easily be calculated. The results of our small simulation study suggest that the estimators well performed for sample sizes greater than 100. A remarkable bias existed for smaller sample sizes ($n < 100$). The only assumption on the statistical model is that $risk(Z)$ has a continuous distribution function. Moreover, it is notable that we do not assume the strict monotonicity in the association of predictor variables Z and the outcome Y such as Moskowitz and Pepe (2004). We only assume that higher values of Z are associated with a higher probability at all for the outcome of interest.

The measures $\widehat{TG^{std}_{ppv}}$ and $\widehat{TG^{std}_{npv}}$ provide a tool for assessing the performance or accuracy of a set of predictor variables Z for the prediction of Y, whereas sensitivity and specificity by themselves could not assess the performance or accuracy. Sensitivity and specificity are measures of classification performance.

We applied our estimators to data from the DETECT study to investigate whether the incorporation of Nt-pro-BNP on top of the standard risk factors for cardiovascular risk management improved the prediction of outcome. The basic condition is statistical significance for a positive response of including Nt-pro-BNP in cardiovascular risk management. However, statistical significance only stands for the fact that the sample is large enough to obtain a statistically significant result. Statistical significance does not imply clinical significance or clinical importance of a predictor variable (Pencina et al., 2008). The total gain in predictive values are statistical measures for improvement in model performance. Unfortunately, these measures lack a clinical interpretation such as the AUC.

Moskowitz and Pepe (2004) first presented the positive and negative predictive value curves (see Fig. 1) for a single continuous risk indicator. The statistical modeling framework for comparing the predictive values of two sets of predictor variables does not provide a simple estimate and comparison for the positive and negative predictive value curves as already stated by Huang et al. (2007). Moskowitz and Pepe (2004) compare single predictor variables. The statistical model in Moskowitz and Pepe (2004) could also be extended to explore a combination of predictor variables as suggested in our paper.

Acknowledgments The authors would like to thank the associate editor and the anonymous reviewer for their constructive comments.

Diabetes cardiovascular risk-evaluation: targets and essential data for commitment of treatment (DETECT) is a cross-sectional and prospective-longitudinal, nationwide clinical epidemiological study. DETECT is supported by an unrestricted educational grant of Pfizer GmbH, Karlsruhe, Germany. Principal investigator: Prof.

Dr. H.-U. Wittchen; staff members: Dr. H. Glaesmer, E. Katze, Dr. J. Klotsche, Dipl.-Psych. L. Pieper, Dipl.-Psych. A. Bayer, Dipl.-Psych. A. Neumann. Steering Committee: Prof. Dr. H. Lehnert (Magdeburg, Coventry), Prof. Dr. G. K. Stalla (Muenchen), Prof. Dr. M. A. Zeiher (Frankfurt); Advisory Board: Prof. Dr. W. Maerz (Graz), Prof. Dr. S. Silber (Muenchen), Prof. Dr. Dr. U. Koch (Hamburg), PD Dr. D. Pittrow (Muenchen, Dresden).

Conflict of interest
All authors of the paper declare that they have no conflict of interest.

Appendix

Estimators for total gain in positive and negative predictive values

In the statistical model, (X_i, Y_i), $1 \le i \le n$ are independent and identical distributed realizations of a random vector $(X, Y) \in R \times \{0, 1\}$. The variable X has an unknown marginal distribution function F. Estimators for $TG_{ppv} := \int_{[0,1]}(ppv(t) - \pi)\,\mathrm{d}t$ and $TG_{npv} := \int_{[0,1]}(npv(t) + \pi - 1)\mathrm{d}t$ in data $(F(X), Y) \in (0, 1) \times \{0, 1\}$ are given by

$$\widehat{TG_{ppv}} = \sum_{k=0}^{n-1}\left(\frac{1}{n(n-k)}\sum_{i=k+1}^{n} Y_{[i:n]}\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$\widehat{TG_{npv}} = \sum_{k=1}^{n}\left(\frac{1}{nk}\sum_{i=1}^{k}\left(1 - Y_{[i:n]}\right)\right) + \frac{1}{n}\sum_{i=1}^{n} Y_i - 1.$$

Proof. We consider the order statistics of the sample $((X_1, Y_1), \ldots, (X_n, Y_n))$. The i-th order statistics is denoted by $X_{i:n}$ and by $Y_{[i:n]}$ the i-th concomitant satisfying $Y_{[i:n]} = Y_j \iff X_{i:n} = X_j$. Recall the empirical distribution function of $(X_1, \ldots, X_n)$, $F_n(t) := \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \le t\}}$, $t \in [0, 1]$.

A sample estimate of $P[Y = 1 | F(X) > t]$ is given by

$$\widehat{ppv}_n(t) := \frac{\sum_{i=1}^{n} 1_{\{F_n(X_{i:n}) > t\}} Y_{[i:n]}}{\sum_{i=1}^{n} 1_{\{F_n(X_{i:n}) > t\}}},$$

for $t \in \left\{\frac{1}{n}, \ldots, \frac{n-1}{n}, 1\right\}$.

We set $\widehat{ppv}_n := 0$ for $t \in [0, \frac{1}{n})$, $F_n(X_{0:n}) := 0$ and $F_n(X_{n+1:n}) := 1$. It follows that

$\sum_{i=1}^{n} 1_{\{F_n(X_{i:n})>t\}} = n - \sum_{i=1}^{n} 1_{\{F_n(X_{i:n})\le t\}} = n - \sum_{i=1}^{n} 1_{\{\frac{i}{n}\le t\}} = n-k$ for $t \in \left[\frac{k}{n}, \frac{k+1}{n}\right)$. Then

$$\begin{aligned}
\widehat{PPV}_n(t) &:= \int_{\left[\frac{1}{n},\ldots,1\right)} \widehat{ppv}_n(t)dt \\
&= \int_{\left[\frac{1}{n},\ldots,1\right)} \frac{\sum_{i=1}^{n} 1_{\{F_n(X_{i:n})>t\}} Y_{[i:n]}}{\sum_{i=1}^{n} 1_{\{F_n(X_{i:n})>t\}}} dt \\
&= \sum_{k=0}^{n-1} \int_{\left[\frac{k}{n},\frac{k+1}{n}\right)} \frac{\sum_{i=1}^{n} 1_{\{F_n(X_{i:n})>t\}} Y_{[i:n]}}{\sum_{i=1}^{n} 1_{\{F_n(X_{i:n})>t\}}} dt \\
&= \sum_{k=0}^{n-1} \int_{\left[\frac{k}{n},\frac{k+1}{n}\right)} \frac{1}{n-k} \sum_{i=k+1}^{n} Y_{[i:n]} dt \\
&= \sum_{k=0}^{n-1} \left(\frac{1}{n(n-k)} \sum_{i=k+1}^{n} Y_{[i:n]} \right).
\end{aligned}$$

The quantity $\pi = P[Y=1]$ can be estimated by $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. It yields the estimator $\widehat{TG_{ppv}}$ for TG_{ppv}. The total gain in negative predictive value could be calculated by similar arguments.

References

Altman, D. G. and Bland, J. M. (1994). Diagnostic tests 2: predictive values. *British Medical Journal* **309**, 102.

Bura, E. and Gastwirth, J. L. (2001). The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biometrical Journal* **43**, 5–21.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19**, 1141–1164.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, London.

Executive Summary of the 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) (2001). *Journal of the American Medical Association* **285**, 2486–2497.

Ferger, D. and Klotsche, J. (2009). Estimation of split-points in binary regression. *Statistics and Decision* **27**, 1001–1044.

Ferger, D., Klotsche, J. and Lueken, U. (2012). Estimation and testing of crossing-points in fixed design regression. *Statistica Neerlandica*, in press.

Gu, W. and Pepe, M. (2009). Measures to summarize and compare the predictive capacity of markers. *The International Journal of Biostatistics* **5**, Article 27.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley, New York.

Huang, Y., Pepe, M. S. and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63**, 1181–1188.

Kraemer, H. C. (2008). Comments on "Evaluating the added predictive ability of a new marker". *Statistics in Medicine* **27**, 196–198.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired design. *Biometrics* **56**, 345–351.

Leistner, D. M., Klotsche, J., Pieper, L., Stalla, G. K., Lehnert, H., Silber, S., Marz, W., Wittchen, H. U. and Zeiher, A. M. (2012). Circulating troponin as measured by a sensitive assay for cardiovascular risk assessment in primary prevention. *Clinical Chemistry* **58**, 200–208.

Moskowitz, C. S. and Pepe, M. S. (2004). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics* **5**, 113–127.

Pencina, M. J., D'Agostino, R. B., 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**, 157–72.

Pfeiffer, R. M. and Gail, M. H. (2011). Two criteria for evaluating risk prediction models. *Biometrics* **67**, 1057–1065.

StataCorp. 2009. *Stata Statistical Software: Release 11*. StataCorp LP, College Station, TX.

Stirzaker, D. (1994). *Elementary Probability.* Cambridge University Press, Cambridge.

Wittchen, H. U., Glaesmer, H., Marz, W., Stalla, G., Lehnert, H., Zeiher, A. M., et al. (2005). Cardiovascular risk factors in primary care: methods and baseline prevalence rates - the DETECT program. *Current Medical Research And Opinion* **21**(4), 619–629.

Wittchen, H. U., Glaesmer, H., Marz, W., Stalla, G., Lehnert, H., Zeiher, A. M., Silber, S., Koch, U., Bohler, S., Pittrow, D., Ruf, G. and Grp, D. E.-S. (2005). Cardiovascular risk factors in primary care: methods and baseline prevalence rates—the DETECT program. *Current Medical Research and Opinion* **21**, 619–629.

