ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

Assessing Methods for Predictive Cut-Point Estimation: A Simulation-Based Comparison

Di Gennaro Piergiacomo[1], Fordellone Mario[1], Nicolao Giovanni[1], Schiattarella Paola[1], Smimmo Annafrancesca[1], Speranza Teresa[1], Simeon Vittorio[1], Signoriello Simona[1], Chiodini Paolo[1]

(1) Unità di Statistica Medica, Università degli studi della Campania "Lugi Vanvitelli"

CORRESPONDING AUTHOR: *Di Gennaro Piergiacomo, piergiacomo.digennaro@unicampania.it*

INTRODUCTION

The identification of an optimal cut-point for continuous biomarkers plays a crucial role in defining patient subgroups likely to benefit from specific treatments. While the literature has extensively covered prognostic biomarkers, those that provide outcome prediction regardless of treatment, the methodological framework for identifying predictive effect, which inform treatment effect heterogeneity, is less developed. This is primarily due to the added complexity of modelling treatment-biomarker interactions, which poses challenges related to statistical power, overfitting, and bias.

OBJECTIVES

This study aimed to compare three statistical methods for the identification of predictive cut-points in time-to-event data. Our goal was to assess their performance in estimating the correct interaction effect and identifying a responder subgroup, under simulation settings that account for variability in treatment efficacy, biomarker predictive effect, and subgroup prevalence.

METHODS

We implemented three approaches: Procedure B of the Biomarker-Adaptive Threshold Design (M1), which combines test statistics across possible cut-points using a permutation test based on likelihood-ratio statistics; the Differential Hazard Ratio method (M2), which selects the cut-point with the largest difference in HRs across adjacent thresholds; and a Minimum P-value method (M3) adapted for interaction terms in the Cox model [1,2]. We conducted a simulation study with 1000 replications from an exponential distribution with an expected censoring rate of approximately 40%. Eight main scenarios were defined by all possible combinations of two sample sizes (n = 300 and n = 500), two treatment effect sizes (HR = 1 or 0.5), two interaction effect sizes (HR = 1 or 0.5), and a biomarker prognostic effect set to HR = 0.6. In addition, we included two extra scenarios calibrated to achieve 80% power: one based on the interaction effect test (β for treatment-biomarker interaction) and one on the subgroup effect test (β within responders). In each replication, the true cut-point was randomly drawn from the biomarker distribution between the 20th and 80th percentiles. For each method, we evaluated statistical power, cut-point estimation bias, subgroup and predictive coefficient estimation bias, and type I error. A significance level of 0.05 was used for all three methods. The procedures were also evaluated on a real case on a prostate cancer clinical trial conducted by the Second Veterans Administration Cooperative Urologic Research Group [3].

RESULTS

M1 consistently demonstrated robust performance, with type I error close to the nominal level (***S2***, 5.6%) and minimal bias in cut-point estimation (***S1***, $\hat{c}$: 0.005±0.06). It maintained

DOI: 10.54103/2282-0930/29367

good power even when the subgroup size was small. M2 showed unstable cut-point estimates (**S1**, $\hat{c}$: 0.055±0.42) and high variability in interaction estimates (**S1**, β_3: 0.463±1.46), yielding a very low power (**S1**, 16.2%). While the M3 achieved the highest power in some scenarios (**S1**, 82.1%), it exhibited significant type I error inflation (**S2**, 50.1%) and substantial bias due to multiple testing without correction (**S1**, β_3: 0.401±1.730). In small subgroups, all methods experienced reduced performance, but M1 remained the most stable. On the prostate cancer dataset, M1 identified a plausible treatment-responsive subgroup, while the other two methods produced conflicting or less reliable results.

CONCLUSIONS

Our results highlight the need for robust methods in predictive cut-point estimation. M1 showed the best balance between error control and accuracy. In contrast, M2 and M3 may lead to overfitting, unstable estimates, and inflated first error rates. Future research should extend these comparisons to more complex models including multivariate biomarkers.

Table 1. Empirical power, type-I error and parameter estimation bias from setting scenarios in using the three compared methods

Scenarios (N,exp(β_1), exp(β_3))	Empirical power / Type-I error [a]			$\hat{c}^b$ (SD)			β_3^b (SD)			$\gamma_{\hat{c}}^b$ (SD)		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
S1 (300, 1.0, 0.5)	0.636	0.162	0.821	0.005 (0.06)	0.055 (0.42)	0.003 (0.24)		0.463 (1.46)	-0.401 (1.730)	-0.167 (0.26)	0.385 (1.30)	-0.356 (2.33)
S2 (300, 1.0, 1.0)	0.056	0.084	0.501	0.030 (0.15)	0.045 (0.37)	0.010 (0.37)		0.003 (1.19)	-0.040 (1.77)	-0.419 (2.98)	-0.009 (1.12)	-0.077 (2.55)
S3 (300, 0.5, 0.5)	0.998	0.144	0.737	-0.001 (0.09)	0.041 (0.41)	0.021 (0.28)		0.632 (2.47)	-0.955 (3.74)	-0.825 (1.02)	-0.161 (2.42)	-2.065 (5.02)
S4 (300, 0.5, 1.0)	0.918	0.062	0.476	-0.003 (0.19)	-0.005 (0.36)	-0.005 (0.36)		-0.033 (1.93)	-0.275 (2.97)	-0.818 (1.07)	-0.559 (1.68)	-1.165 (3.91)
S5 (500, 1.0, 0.5)	0.853	0.331	0.932	-0.000 (0.03)	0.100 (0.43)	-0.005 (0.19)		0.392 (0.60)	-0.226 (0.84)	-0.072 (0.19)	-0.259 (0.55)	-0.159 (1.02)
S6 (500, 1.0, 1.0)	0.057	0.158	0.504	0.006 (0.12)	0.075 (0.44)	0.009 (0.37)		0.023 (0.70)	0.020 (0.93)	0.077 (0.39)	-0.014 (0.38)	0.022 (0.85)
S7 (500, 0.5, 0.5)	1.000	0.178	0.881	0.000 (0.05)	0.062 (0.43)	0.019 (0.22)		0.486 (1.50)	-0.461 (2.07)	-0.707 (0.26)	-0.378 (1.42)	-1.272 (3.02)
S8 (500, 0.5, 1.0)	0.998	0.093	0.509	-0.001 (0.12)	0.020 (0.40)	0.002 (0.37)		0.063 (0.82)	0.015 (1.45)	-0.705 (0.26)	-0.533 (0.78)	-0.790 (1.84)
S9 (264, 1.0, 0.5)	0.616	0.189	0.812	0.006 (0.06)	0.074 (0.36)	0.017 (0.22)		0.630 (1.88)	-0.394 (2.21)	-0.228 (0.88)	0.556 (1.67)	-0.374 (2.85)
S10 (481, 1.0, 0.5)	0.838	0.305	0.920	0.003 (0.03)	0.082 (0.43)	-0.011 (0.18)		-0.292 (0.57)	-0.910 (0.83)	-0.774 (0.20)	-0.423 (0.49)	-0.853 (1.19)

0.8 Power target for test on $\gamma_{\hat{c}}$
0.8 Power target for test on β_3
[a]Depending on the scenario, the Empirical power/Type-I error column shows the power (when H0 is true) as non-underlined values and the type-I error (when H1 is true) as underlined values
[b]Represents the empirical bias, mean and (SD)

REFERENCES

1. Jiang W, Freidlin B, Simon R, Biomarker-Adaptive Threshold Design: A Procedure for Evaluating Treatment With Possible Biomarker-Defined Subset Effect. JNCI Journal of the National Cancer Institute, 2007;99(13):1036-1043
2. Rabbee N, Biomarker Analysis in Clinical Trials with R. Chapman and Hall/CRC, 2020
3. Byar DP, Corle DK, Selecting optimal treatment in clinical trials using covariate information. J Chronic Dis, 1977; 30: 445-59

DOI: 10.54103/2282-0930/29367