Published in final edited form as:

Am Stat. 2012 ; 66(3): 173–179. doi:10.1080/00031305.2012.703873.

A Simulation Based Evaluation of the Asymptotic Power Formulae for Cox Models in Small Sample Cases

Mehmet Kocak* and

University of Tennessee Health Science Center, Department of Preventive Medicine 66 N. Pauline Street, Memphis, TN 38105

Arzu Onar-Thomas

St. Jude Children's Research Hospital, Department of Biostatistics 262 Danny Thomas Place, Memphis, TN 38105 arzu.onar@stjude.org

Abstract

Cox proportional hazards (PH) models are commonly used in medical research to investigate the associations between covariates and time to event outcomes. It is frequently noted that with less than ten events per covariate, these models produce spurious results, and therefore, should not be used. Statistical literature contains asymptotic power formulae for the Cox model which can be used to determine the number of events needed to detect an association. Here we investigate via simulations the performance of these formulae in small sample settings for Cox models with 1- or 2-covariates. Our simulations indicate that, when the number of events is small, the power estimate based on the asymptotic formulae is often inflated. The discrepancy between the asymptotic and empirical power is larger for the dichotomous covariate especially in cases where allocation of sample size to its levels is unequal. When more than one covariate is included in the same model, the discrepancy between the asymptotic power and the empirical power is even larger, especially when a high positive correlation exists between the two covariates.

Keywords

Time to event data; Cox Proportional Hazards Model; number of events per covariate; event size formula

1. INTRODUCTION AND BACKGROUND

Finding prognostic factors for a disease or identifying characteristics that are associated with an outcome of interest is a common aim in medical research. Time-to events such as time to recovery following a medical procedure, time to engraftment after a transplant, or time to tumor recurrence/progression or death, are often studied with the intent to investigate their associations with other covariates. Since time to event data very often include censored observations, the proportional hazards model proposed by D.R. Cox (1972) (Cox PH from this point onwards) is regularly used for analysis. When censoring is present, statistical power to detect an association is primarily determined by the number of events observed rather than the number of subjects in the sample. Further, number of events needed to achieve a desired level of power depends on the number of covariates included in the model. Hence most empirical studies of power published in the literature report number of events

*Corresponding Author Majority of this research was done while the corresponding author was employed by St. Jude Children's Research Hospital, Memphis, TN mkocak1@uthsc.edu.

per variable (EPV) which is simply the total number of events in the sample divided by the total number of variables in the model.

Concato et al. (1995) and Peduzzi et al. (1995) investigated the effect of EPV on various aspects of the estimates obtained from a Cox PH model. The authors used data from a randomized trial of 673 patients with seven covariates, where they studied EPV values of 2, 5, 10, 15, 20, and 25. They observed that, for EPV = 10, the regression coefficients were increasingly biased, confidence intervals did not have proper coverage, test statistics were not valid, and serious loss of power led to underfitting. Vittinghoff and McCulloch (2007) also investigated the impact of EPV on confidence interval coverage, Type-I error, relative bias, and other performance measures. The sample sizes they simulated ranged from 128 to 1024, where EPV varied from 2 to 16, and the number of covariates ranged from 2 to 16. They suggested that the EPV >10 recommendation by Concato et al. (1995) and Peduzzi et al. (1995) may be too conservative, as their simulations showed that the relative bias was within acceptable limits, Type-1 error was fairly well maintained even with 5-9 EPV, and the confidence interval coverage was similar for 5-9 and 10-16 EPV. Hence they concluded that statistically significant findings based on 5-9 EPV may still have merit and should not be discounted. They noted however that their work didn't directly investigate the influence of EPV on power and that they expected Type-2 errors to be common in cases where EPV is low.

In this manuscript, our primary interest is determining the power available to identify variables associated with a time to event outcome in small sample settings. Asymptotic formula currently available in statistical literature can serve as a starting point. Notably, for a single dichotomous covariate, Schoenfeld (1983) proposed the following asymptotic formula to estimate the minimum number of events, D :

$$D = \frac{(Z_{1-\alpha} - Z_{1-\beta})^2}{P(1-P)(\log \Delta_0)^2},$$

where Δ_0 is the hazard ratio to be detected, P is the proportion of subjects assigned to one level of the dichotomous covariate, and α and β are the type-I and type-II error rates, respectively for the underlying hypotheses $H_0: \Delta = 1$ versus $H_a: \Delta = \Delta_0$. Hsieh and Lavori (2000) provided a modified version of Schoenfeld's formula for a continuous covariate:

$$D = \frac{(Z_{1-\alpha} - Z_{1-\beta})^2}{\sigma^2(\log \Delta_0)^2},$$

where σ^2 is the variance of the covariate. Note that neither of these formulae incorporates any information regarding the censoring mechanism or the censoring fraction. Instead they provide an estimate of the number of events needed to detect a given effect size with a desired power and it is up to the investigator to choose the sample size or the follow-up time or both to ensure that the desired number of events is attained. Hsieh and Lavori (2000) and Latouche et al. (2000) also provided the following extension for multivariable Cox PH models:

$$D = \frac{(Z_{1-\alpha} - Z_{1-\beta})^2}{\sigma^2(\log \Delta_0)^2(1 - \rho^2)},$$

where ρ denotes the correlation between two covariates one of which is considered primary. The correlation is quantified via Pearson's correlation for the two continuous covariate case and via point biserial correlation, which is mathematically equivalent to Pearson's correlation, in the one continuous and one dichotomous covariate case. Thus, compared to the single-covariate case, when two covariates are present in the same model the number of events needed to achieve the specified power is inflated by a factor of $1/(1-\rho^2)$. Generalizing this further, in multivariable cases with a single primary covariate, the formula remains the same except ρ now denotes the multiple correlation between the primary covariate and the other covariates.

Given the asymptotic nature of these formulae, we conducted an extensive simulation study to investigate their performance in terms of power in small sample settings. We mainly focused on single-variable Cox models but also considered the case where a second covariate was included in the same model. In addition, for the dichotomous covariates we explored the effect of imbalance in sample size assignment between the levels of the covariate on power. For all cases we considered the use of the Score and Wald tests. In section 2, we provide a description of the simulations we conducted which are followed by a summary of results in section 3. We conclude with a brief discussion in section 4.

2. DETAILS OF SIMULATION STUDIES

In this simulation study, we generated the time to event endpoints from an exponential distribution. In all simulations, we used an event censoring approach and varied the censoring fraction from 90% to 0%. For example for sample size=12, the following numbers of events were studied: 5, 6, 7, 8, 9, 10, 11 and 12, corresponding to censoring fractions 58% (7/12) - 0% (0/12). Annual patient accrual was assumed to be uniform over three years and baseline hazard rates were chosen so that approximately 30% of the patients experienced an event at the completion of accrual. Three hazard ratios were targeted: 0.4, 1, and 2, for the primary covariate. We generated complete lifetimes for all patients and for each of the targeted number of events, say d_i , we ordered the generated survival times chronologically (taking into account the time of accrual) and "terminated the trial" when the first d_i events were observed. The rest of the patients were censored at the time of the d_i event. One sided tests were used for detecting associations. Following each simulation run, parameter estimates, standard errors, and corresponding p-values were recorded. The results reported in the next section are based on 10,000 simulated trials for each scenario which were performed on SAS® Version 9.3 for Windows®. Table 1 summarizes the set of scenarios investigated.

3. RESULTS

Our results indicated that the Likelihood Ratio and Score tests had practically identical empirical power. For the single dichotomous covariate case, p-values from the Log-rank test and Score tests were also practically identical. Hence the rest of this manuscript focuses on results from the Score and Wald tests only. Using Schoenfeld's (1983) and Hsieh and Lavori's (2000) formulae, the numbers of events needed to detect Hazard Ratios of 0.4 and 2.0 with 80% power and 5% type-1 error rate are presented in Table 2. Table-2 also shows the number of events needed based on empirical results for the same scenario.

3.1 Single Covariate Case

As summarized in Table 2 and presented in Figure 1, based on Hsieh and Lavori's (2000) formula for a continuous covariate, 18 events are needed to detect a HR of 2 with at least 80% power whereas our simulations suggest that at least 24 events are needed. Among

several sample sizes we simulated with EPV=18, the maximum empirical power achieved was 69%.

The difference between empirical power and power based on Hsieh and Lavori's (2000) formula diminished with increasing number of events, as expected. The Score test uniformly provided better power, which was more pronounced for smaller EPV. The empirical power using normally distributed covariates was practically identical to that observed for covariates generated from a uniform distribution (results not shown). Similarly for the dichotomous covariate, the empirical power estimates associated with both the Wald and Score tests were less than the power provided by Schoenfeld's (1983) formula. For example, when hazard ratio of 2.0 was targeted, Schoenfeld's (1983) formula suggested 52 events, while 80% empirical power was achieved with a minimum of 67 events. As expected, the difference between the empirical and asymptotic power estimates decreased as the number of events increased.

The results in Table 2 are based on 50% allocation for each level of the dichotomous covariate, which maximizes power. When imbalance was introduced in the allocated sample size to the levels of the dichotomous covariate, our simulations confirm the expected loss of power (Figure 2). When targeting $HR > 1$, the discrepancy between the empirical power and the asymptotic power increased as a larger proportion of observations were assigned to the higher level of the covariate. When $HR > 1$, the lower level of the covariate is associated with a smaller likelihood of events. With a smaller number of observations assigned to the lower level, the chance of observing an event in this small sample setting becomes even smaller resulting in a loss of power. Similarly for $HR < 1$, the difference between the empirical and asymptotic power increased with proportion of patients assigned to the lower level of the dichotomous covariate (results not shown).

The type-1 error rates stabilized around the targeted 0.05 level with 15-20 EPV both for the continuous and dichotomous covariates. With smaller number of events, the type 1 error rates were lower than the nominal value for the Wald test and higher than the nominal value for the Score test.

3.2 Models with 2 Covariates

We have also investigated differences in power between asymptotic and the empirical results in models with two covariates correlated at various degrees. The results of these simulations are presented in Figure 3 and in Table 3. The empirical power decreased as the correlation between the primary covariate and the secondary covariate moved from strongly negative to strongly positive. Surprisingly, when there was strong negative correlation between two continuous variables, the empirical power was slightly larger than what was suggested by the asymptotic formulae. When the correlation was not strongly negative, the empirical power was lower than what was suggested by the power formulae as before unless EPV was quite large and the two converged. Type-1 error rate was also inflated when the Cox models included two covariates. When one dichotomous and one continuous covariate were investigated, where the dichotomous covariate was designated as the primary, regardless of the degree of correlation, the power formula always gave higher power than what was empirically observed (Table 3b).

3.3 Effect of Sample Size

It is widely accepted that the power of Cox PH models is driven by the number of events, rather than number of subjects when censoring is present. However as Table 3 and 4 illustrate, a modest increase in power may be achieved in cases where the sample size is larger. For example, in Table 4 below a sample of size 18, 16 events yield an empirical

power of about 57% by the Score Test to detect $HR=2$, while the same number of events result in an empirical power of 62.9% in a sample of size 36. This slight increase in power is a result of the fact that for a fixed number of events, larger sample sizes lead to smaller standard error estimates hence resulting in better precision for the parameter estimate and higher power. Consistent with the statistical literature (e.g., Schoenfeld, 1983; Bradburn et al., 2003), the advantage gained by larger sample sizes is modest compared to advantages gained by larger number of events. Further, the gains by increasing the sample size diminish and eventually disappear as the number of events increase. Overall, the empirical power based on the Score test was higher, and thus closer to the formula-based asymptotic power, compared to the Wald's test when the number of events was small (Table 4).

4. DISCUSSION

Cox PH models are by far the most common approach to exploring associations between various covariates and time to event data where censoring is present. In deciding whether adequate number of events is available the '10 events per variable rule' is often invoked, though this prescription is often inadequate, as shown previously by Peduzzi et al. (1995). Here by using simulation studies, we have shown that even the formulae of Schoenfeld (1983) and Hsieh and Lavori (2000) may underestimate the number of events needed in small sample settings. This is not necessarily surprising given the asymptotic nature of these formulae however our simulations suggest that even in cases with moderately large EPV, the formulae may overestimate the actual power. Overall, the empirical power we observed was lower than the expected power specified in these formulae, especially for dichotomous covariates. Our results indicate that if the intent is to identify covariates which may be associated with an outcome of interest in 1- to 2-covariate Cox models, basing such an analysis on samples with 10 or so events per covariate would likely fail due to lack of power even if a moderate association exists. As Table 4 clearly shows for a dichotomous covariate the power of such an analysis may be as low as 30%. In the cases where more than one covariate is included in the same model, the discrepancy between the asymptotic power and the empirical power is even larger, especially when a high positive correlation exists between the two covariates. Hence we recommend that a simulation study is conducted to estimate the number of events needed to achieve the desired power in small sample settings. The SAS code used to perform the simulations in this manuscript are available upon request to facilitate such explorations.

Acknowledgments

We wish to thank Dr. Seok P. Wong of University of Memphis for his contributions to the initial phase of this research. This work was supported in part by NIH grant U01 CA81457 for the Pediatric Brain Tumor Consortium (PBTC) and American Lebanese Syrian Associated Charities. We also would like to thank the Associate Editor and 2 anonymous referees for their insightful comments and suggestions which substantially improved the content and the readability of this manuscript.

REFERENCES

1. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. *Br J Cancer*. Aug 18; 2003 89(4): 605–611. [PubMed: 12915864]
2. Concato J, Peduzzi P, Holfold TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *J Clin Epidemiology*. 1995; 48:1495–501.
3. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society Series*. 1972; B 34:187–220.

4. Hsieh FY, Lavori PW. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Control Clin Trials*. Dec; 2000 21(6):552–60. [PubMed: 11146149]
5. Latouche A, Porcher R, Chevret S. Sample size formula for proportional hazards modelling of competing risks. *Stat Med*. Nov 15; 2004 23(21):3263–74. [PubMed: 15490425]
6. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiology*. 1995; 48:1503–10.
7. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983; 39:499–503. [PubMed: 6354290]
8. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiology*. Mar 15; 2007 165(6):710–8. Epub 2006 Dec 20.

Figure-1.

Comparison of Power between Hsieh and Lavori's (2000) formula and simulations targeting hazard ratios of 2 and 0.4 using the Wald and Score tests based on a uniformly distributed continuous covariate. The vertical dotted lines represent the sample sizes, namely 12, 18, 24, 30, 36, and 45.

Figure-2.
Effect of imbalance in the levels of the dichotomous covariate targeting hazard ratio of 2.0.
The vertical dotted lines represent the sample sizes, namely, 60, 90, and 120.

Figure-3. Asymptotic and Empirical power from Cox models with two continuous covariates correlated at various degrees, targeting Hazard Ratio of 2.0. The vertical dotted lines represent the sample sizes, namely 30, 60, and 90.

Table-1

Simulated scenarios

Scenario	Distribution	EPV
Single Continuous Covariate	Uniform(0,3)	5 to 45
	Normal(1.5, 0.75)	
Single Dichotomous Covariate	Bernoulli(0.50)	10 to 90
	Bernoulli(0.80) *	
	Bernoulli(0.30) *	
	Bernoulli(0.15) *	
Two Continuous Covariates	Normal variables correlated at -0.8, -0.4, 0, 0.4, and 0.8	60 to 120
One Binary and One Continuous Covariate	Bernoulli(0.5) and normal variable with point biserial correlation of 0.73, 0.45, and 0.73	60 to 120

* Generated to investigate the impact of imbalance in the levels of the binary covariate on power.

Table-2

Number of Events based on Schoenfeld's (1983) and Hsieh and Lavori's (2000) formulae, for one continuous, one dichotomous covariate, and two covariate cases, vs. empirical estimates with $\alpha=0.05$ and power=0.80

Covariate(s)	Targeting HR=2.0		Targeting HR=0.4	
	Formula	Empirical Estimates	Formula	Empirical Estimates
One Cont. Variable	18	24	10	14
One Binary Variable	52	67	30	39
Two Cont. Vars. ($\rho=-0.8$)	21	16	12	12
Two Cont. Vars. ($\rho=-0.4$)	16	23	10	15
Two Cont. Vars. ($\rho=0.0$)	18	26	10	16
Two Cont. Vars. ($\rho=0.4$)	21	30	12	18
Two Cont. Vars. ($\rho=0.8$)	48	67	28	38
Binary & Cont. Var. ($\rho^*=-0.73$)	103	>>120	59	78
Binary & Cont. Var. ($\rho^*=-0.45$)	65	103	37	51
Binary & Cont. Var. ($\rho^*=-0.73$)	103	>>120	59	79

* Point Biserial Correlation

Table-3

Asymptotic and empirical power estimates for HR=2 for the primary covariate in a bivariable model with two continuous (a) and a continuous and a dichotomous covariate (b):

		a)					
		Rho=-0.80		Rho=0.0		Rho=0.80	
No. of Events	N	Empirical Power	Power Formula	Empirical Power	Power Formula	Empirical Power	Power Formula
20	30	85.1		67.3		33.3	
	60	87.7	78.4	69.7	85.1	34.5	48.6
	90	88.4		70.6		35.7	
	120	89.0		70.8		35.5	
30	30	92.6		78.2		41.6	
	60	96.6	90.9	85.5	95.0	46.6	62.8
	90	96.9		86.1		47.8	
	120	97.3		85.9		48.1	
40	60	99.0		93.5		57.1	
	90	99.2	96.4	94.2	98.4	58.9	73.7
	120	99.4		94.0		59.6	
50	60	99.6		96.8		67.0	
	90	99.8	98.6	97.6	99.5	68.3	81.6
	120	99.9		97.3		68.8	
60	60	99.8		98.0		71.8	
	90	100.0	99.5	99.1	99.9	76.0	87.4
	120	100.0		98.9		75.8	
70	90	100.0	99.8	99.8	100.0	81.8	91.4
	120	100.0		99.5		81.9	
80	90	100.0	99.9	99.9	100.0	86.1	94.3
	120	100.0		99.9		86.5	

b)

No. of Events N	N	Rho [*] = -0.73		Rho [*] = 0.45		Rho [*] = 0.73	
		Empirical Power	Power Formula	Empirical Power	Power Formula	Empirical Power	Power Formula
20	60	19.4		23.4		16.0	
	90	19.3	29.2	22.9	39.8	16.0	29.2
	120	19.4		22.6		16.1	
30	60	26.5		36.0		24.7	
	90	27.3	38.1	35.8	52.1	23.1	38.1
	120	27.4		36.0		23.8	
40	60	33.5		48.1		33.4	
	90	34.0	46.2	46.9	62.4	30.6	46.2
	120	33.8		47.0		31.1	
50	60	39.6		57.2		40.8	
	90	40.8	53.5	56.1	70.8	38.7	53.5
	120	41.0		56.8		37.8	
60	60	44.1		60.8		44.1	
	90	46.9	60.0	64.7	77.5	46.0	60.0
	120	47.1		64.9		44.7	
70	90	52.4		72.0		53.0	
	120	52.7	65.8	72.6	82.9	51.0	65.8
80	90	57.9		77.4		58.3	
	120	58.3	70.8	78.7	87.0	57.8	70.8

* Biserial Correlation between the primary binary covariate and secondary continuous covariate

Table-4

Effect of Sample Size on Empirical Power in the Single Covariate case Targeting HR=2

EPV	N	Continuous Covariate		Dichotomous Covariate (Balanced)			
		Score Test	Wald Test	Hsieh and Lavori (2000)	Score Test	Wald Test	Schoenfeld (1983)
16	18	57	54.3	77.5	26.6	23.1	39.8
	24	61.9	59.9		28.5	25.3	
	30	63	60.8		28.1	25	
	36	62.9	60.9		27.8	24.7	
	45	63.1	61.1		27.4	24	
18	18	57.7	55.2	81.6	26.9	23.5	43.1
	24	66.5	64.5		31.5	28.8	
	30	68.1	66.5		31.5	28.6	
	36	67.8	66		30.4	27.7	
	45	68.1	66.4		30	27.1	
24	24	72.9	71.4	90.2	36	33.5	52.1
	30	79.2	78.1		39.5	37.4	
	36	79.3	78.4		38.2	36.2	
	45	80	79.2		38	36	
	45				60	58.8	
45	60			75.2	62.7	61.7	75.2
	75				62.5	61.3	
	90				62.8	61.7	
	60				68.4	67.7	
	75				68.7	67.9	
52	90			80.4	69.9	69	80.4
	60				72.1	71.4	
	75				75	74.4	
60	90			85.1	75.6	74.9	85.1
	75				80.6	80.2	
	90				81.9	81.3	
70	75			89.5			89.5
	90						