Pergamon

0895-4356(95)00048-9

J Clin Epidemiol Vol. 48, No. 12, pp. 1503–1510, 1995

0895-4356/95 $9.50+0.00

IMPORTANCE OF EVENTS PER INDEPENDENT VARIABLE IN PROPORTIONAL HAZARDS REGRESSION ANALYSIS II. ACCURACY AND PRECISION OF REGRESSION ESTIMATES

PETER PEDUZZI,[1,2] JOHN CONCATO,[3,4]* ALVAN R. FEINSTEIN,[2,4] and THEODORE R. HOLFORD[2]

[1]Cooperative Studies Program Coordinating Center, West Haven Veterans Affairs Medical Center, West Haven, Connecticut, [2]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut, [3]Medical Service, West Haven Veterans Affairs Medical Center, West Haven, Connecticut, and [4]Department of Medicine (Clinical Epidemiology Unit), Yale University School of Medicine, New Haven, Connecticut

(Received in revised form 8 February 1995)

Abstract—The analytical effect of the number of events per variable (EPV) in a proportional hazards regression analysis was evaluated using Monte Carlo simulation techniques for data from a randomized trial containing 673 patients and 252 deaths, in which seven predictor variables had an original significance level of $p < 0.10$. The 252 deaths and 7 variables correspond to 36 events per variable analyzed in the full data set.

Five hundred simulated analyses were conducted for these seven variables at EPVs of 2, 5, 10, 15, 20, and 25. For each simulation, a random exponential survival time was generated for each of the 673 patients, and the simulated results were compared with their original counterparts. As EPV decreased, the regression coefficients became more biased relative to the true value; the 90% confidence limits about the simulated values did not have a coverage of 90% for the original value; large sample properties did not hold for variance estimates from the proportional hazards model, and the Z statistics used to test the significance of the regression coefficients lost validity under the null hypothesis.

Although a single boundary level for avoiding problems is not easy to choose, the value of EPV = 10 seems most prudent. Below this value for EPV, the results of proportional hazards regression analyses should be interpreted with caution because the statistical model may not be valid.

INTRODUCTION

When the number of outcome events per independent variable (EPV) is "too small" in a multivariable analysis, the results from the fitted regression model may not be accurate or precise. Although this phenomenon has been recognized [1, 2] as a potential limitation of multivariable methods, the magnitude and other characteristics of the problem for proportional hazards models have seldom been documented. This research was done to provide documentary evi-

*Address correspondence and reprint requests to: John Concato, M.D., M.S., M.P.H., Medical Service/111GIM, West Haven VAMC, 950 Campbell Ave., West Haven, CT 06516.

dence via a simulation study, whose rationale and methods were described in Part I of this report. In Part II we compare the results of the simulated proportional hazards analyses and discuss their implications.

METHODS

The methods were completely described in Part I. In brief, the simulations were derived from data for a cohort of 673 patients enrolled in a trial comparing medical and surgical therapy for coronary artery disease. Seven variables that were associated with the outcome (death) at a significance level below 0.10 were selected for analysis. Five hundred simulations were conducted for EPVs of 2, 5, 10, 15, 20, and 25. For each simulation, each of the 673 patients received a randomly generated exponential survival time that preserved the rank order of their baseline covariates, while allowing the truncated survival times to provide the desired EPV. After a proportional hazards analysis was then done for each simulation, the results of the full original group having an EPV of 36 were compared with those from each set of 500 simulations, using the indexes of accuracy, precision, and significance described in Part I.

RESULTS

Original model (EPV = 36)

For each variable in the original model (based on 36 EPV), the prevalence, and the multivariable estimates of the regression coefficients and corresponding standard errors, were presented and discussed in Part I. The values of the seven coefficients ranged from 0.229 to 0.459, corresponding to moderate hazard ratios of 1.26 to 1.58. These results were used as the "gold standard" for comparison with the simulations.

Simulation results

The proportional hazards models did not always fully converge, as defined in the previous paper, for both the likelihood function and the parameter estimates at all values of EPV. The rates of full convergence were 100% for EPV $\geq$ 10, 99% for EPV = 5, and 80% for EPV = 2. Convergence was difficult to obtain at low EPV because some covariates (congestive heart failure and diabetes) had low prevalence rates, increasing the chance of generating simulated samples of patients in which no deaths occurred when the factor was present. For models that converged, the six panels of Fig. 1 show the frequency distribution of values of the regression coefficients at EPVs of 25, 20, 15, 10, 5, and 2 for a single variable: congestive heart failure (original value of regression coefficient = 0.350). As EPV decreased, the distributions of the regression coefficients became more dispersed with a flatter shape. At EPV of 10 or less, the distributions were distinctly not Gaussian.

The simulations for the other six independent variables produced generally similar patterns, which are not shown here.

Accuracy of regression coefficients. Figure 2 shows that for all seven covariates, the bias of the simulated proportional hazards regression coefficients, relative to the "true" value, tended to increase with decreasing EPV. At an EPV of 10 or greater, the average bias was generally within $\pm 10\%$ of the true value. For EPVs less than 10, the bias increased dramatically for three of the factors. The magnitude of the regression coefficient was overestimated by more than 20% for number of vessels diseased and for New York Heart Association functional class $\geq$ III, and was underestimated by nearly 30% for diabetes.

This problem was further evaluated in Fig. 3, which shows the proportion of simulations in which the percent relative bias in the coefficient exceeded $\pm 100\%$ for each of the seven independent variables. At EPV = 20 or 25, the proportion of simulations with greater than 100% relative bias was less than 0.2 for all variables except congestive heart failure (CHF) and New York Heart Association (NYHA) classification. For each variable, the proportion increased markedly with decreasing EPV; the absolute bias at 2 EPV exceeded 100% in more than half of the simulations, except for left ventricular contraction abnormality (LVC). Thus, extremely distorted estimates of regression coefficients were more likely to occur at low EPV.

Precision of regression coefficients. The "sample variance" (as defined in Part I) of regression coefficients calculated from the 500 simulations is displayed in Fig. 4 for each of the 7 covariates. As expected, the sample variances increased with decreasing EPV because the number of events decreased. The trend for "model variance" (as previously defined) was also similar and is not shown here.

The ratio of mean model variance to sample variance appears in Fig. 5. The departures from values of 1 show that at 2 EPV the large sample properties of the proportional hazards model

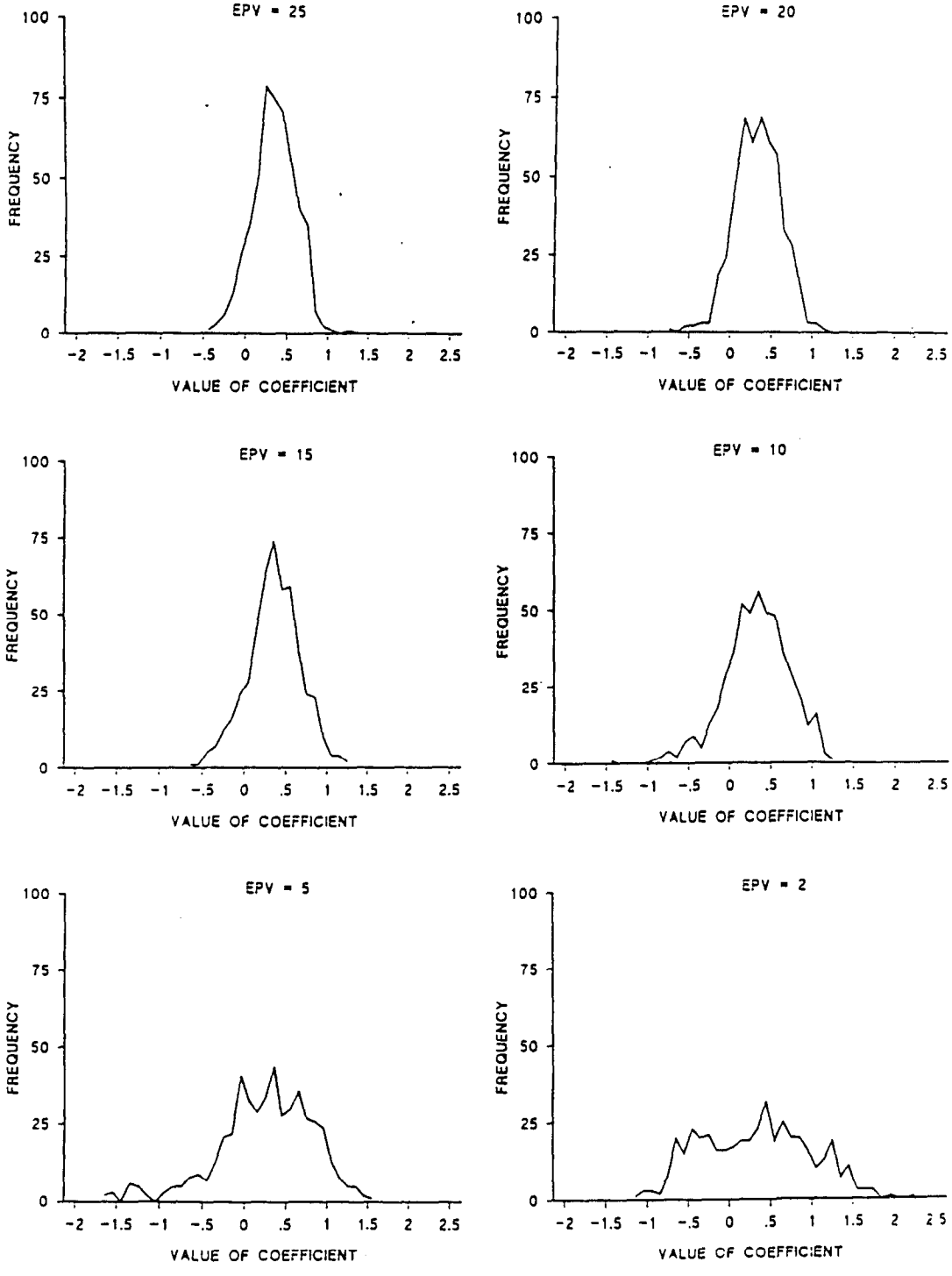

Fig. 1. Number of events per variable, and distribution of simulated regression coefficients for variable "congestive heart failure."

may not hold. At 2 EPV, the variance was overestimated by more than 20% for diabetes and by nearly 80% for congestive heart failure; and was underestimated by 20% for number of vessels diseased. All other ratios were within ±10% of the expected value of 1.

Significance testing of regression coefficients. Figure 6 shows the proportion of simulations in which the true value (point estimate) of the coefficient was included in the 90% confidence interval about the estimated value. This proportion or "coverage" was generally greater

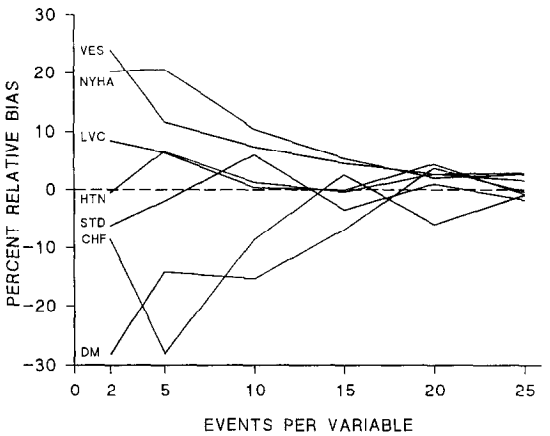

Fig. 2. Number of events per variable, and average percent relative bias. Abbreviations for variables: CHF, history of congestive heart failure; DM, history of diabetes mellitus; HTN, history of hypertension; LVC, presence of a left ventricular contraction abnormality; NYHA, New York Heart Association Functional class III or IV; STD, ST depression in the resting baseline electrocardiogram; VES, number of coronary vessels with significant lesions.

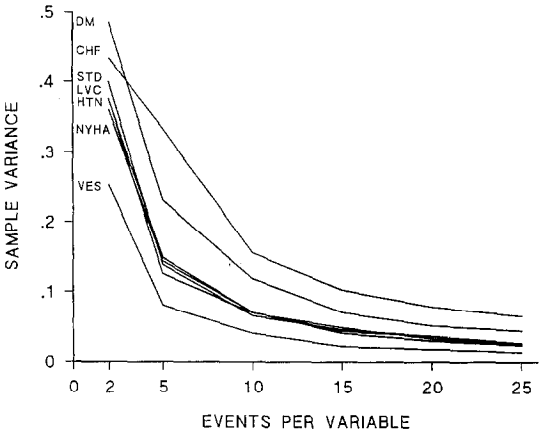

Fig. 4. Sample variance of the simulated regression coefficients. Abbreviations are as indicated in Fig. 2.

than 90% at low EPV; and at 2 EPV the coverage was almost 95% for diabetes and congestive heart failure.

The proportion of simulations in which the Z statistic exceeded the critical positive value, 1.28 (a counterpart of 90% power), is displayed in Fig. 7. The values at EPV = 36 correspond to the original model. The power to detect significant effects decreased with decreasing EPV, and at an EPV of 2 the curves began to converge at levels well below 50% power.

Figure 8 shows the proportion of simulations

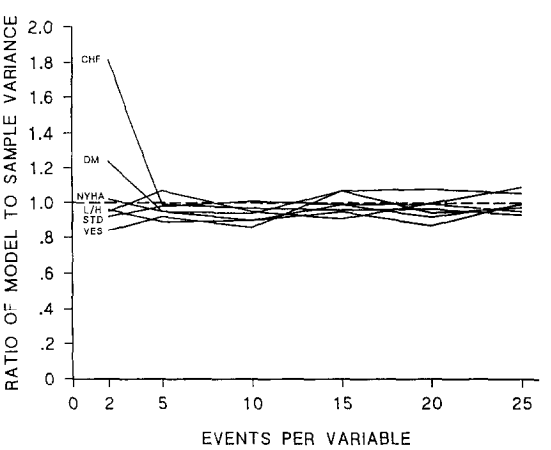

Fig. 5. Ratio of "model variance" (mean of variance terms) to "sample variance" (variance of regression coefficients), as defined in text. Abbreviations are as indicated in Fig. 2.

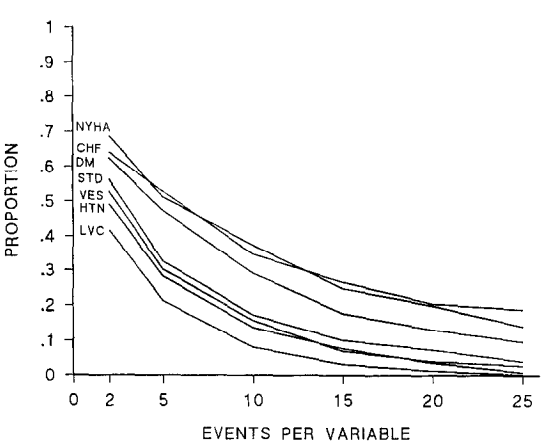

Fig. 3. Number of events per variable and proportion of simulations in which the percent relative bias exceeded ±100%. Abbreviations are as indicated in Fig. 2.

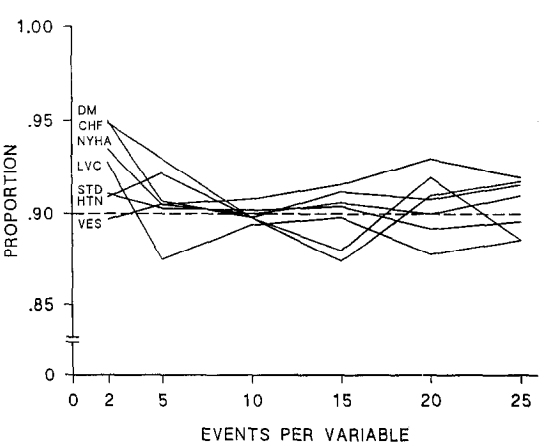

Fig. 6. Proportion of simulations in which the 90% confidence interval about the simulated regression coefficient includes the true value. Abbreviations are as indicated in Fig. 2.

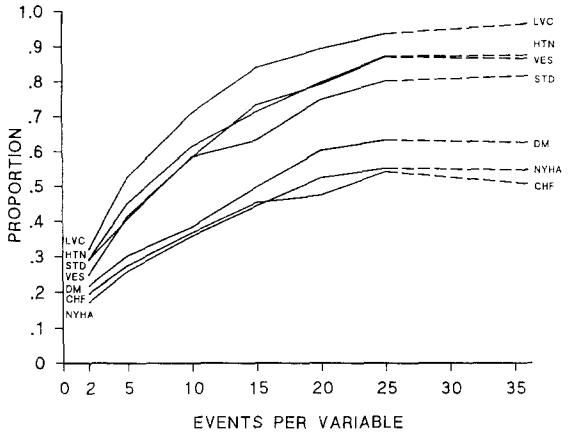

Fig. 7. Proportion of simulations in which the Z statistic (coefficient/standard error) exceeds the standard normal deviate of 1.28 for 90% power. The dashes represent the actual power for the full model with 36 events per variable. Abbreviations are as indicated in Fig. 2.

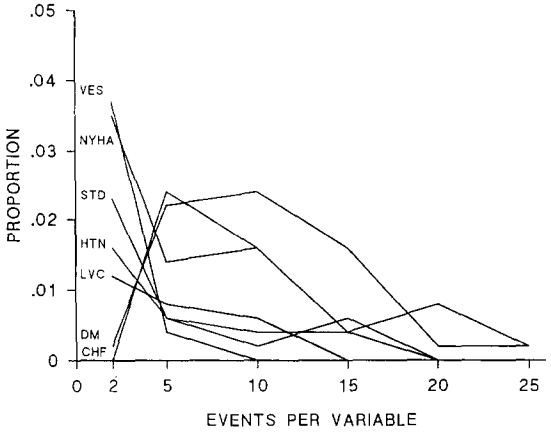

Fig. 8. Proportion of simulations in which the Z statistic was less than -1.28. Abbreviations are as indicated in Fig. 2.

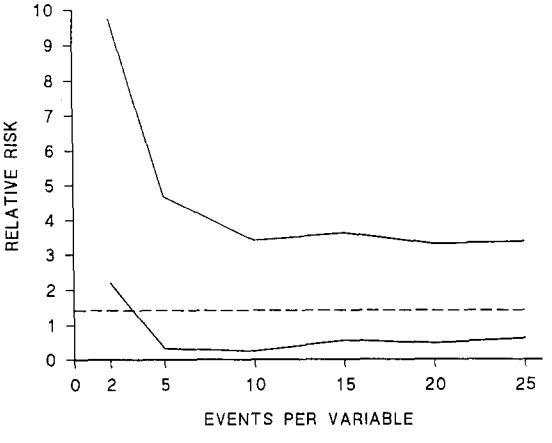

Fig. 9. Minimum and maximum values of relative risk that were significant at the 10% level (two-tail) for the variable "congestive heart failure."

in which paradoxical fitting would occur because the Z statistic was less than -1.28. Although the proportions were low for all covariates, the chance of paradoxical fitting was greatest at low EPV.

The minimum and maximum relative risks (hazard ratios) that were "significant" at the 10% level (two-tail) are displayed in Fig. 9 for the variable congestive heart failure. The range of significant relative risks was reasonably constant from 25 to 10 EPV, and increased markedly at levels of 5 and 2 EPV. At 2 EPV, the maximum significant relative risk was nearly 10 compared with the "gold standard" value of 1.42. Similar trends for the other six variables are not shown.

The panels of Fig. 10 show the distribution of the Z statistic for the coefficient of congestive heart failure. The distributions are derived from the simulations conducted under the global null hypothesis of no effect of the covariates, and were all significantly different from a Gaussian distribution at the 0.05 level, according to the Kolmogorov D statistic [3]. The distributions also tended to narrow with decreasing EPV. In contrast, for coefficients of the other six independent variables (not shown), significant departures from normality occurred only at EPV of less than 10.

Table 1 displays the number and percentage of simulations for each variable found to be significant at the 0.10 level under the null hypothesis of no covariate effect. The global type I error was calculated as the total number of variables found to be significant over all the simulations, divided by the total number of variables evaluated (i.e., 7 times the number of analyzable simulations, up to 500). As with the primary set of simulations, the proportional hazards models did not always converge for both the likelihood function and the parameter estimates. The convergence rates for the simulations under the null hypothesis of no covariate effect were 100% for EPV $\geq$ 10, 90% for EPV = 5, and 60% for EPV = 2. The type I error decreased from 9.7% for 25 EPV to 7.2% for 2 EPV, indicating that the Z statistic became overly conservative at low EPV (i.e., the nominal p value was less than 0.10).

DISCUSSION

The simulation studies demonstrate some of the problems that arise for a too-small number of events per variable in proportional hazards analysis. As EPV decreases, the bias of the re-

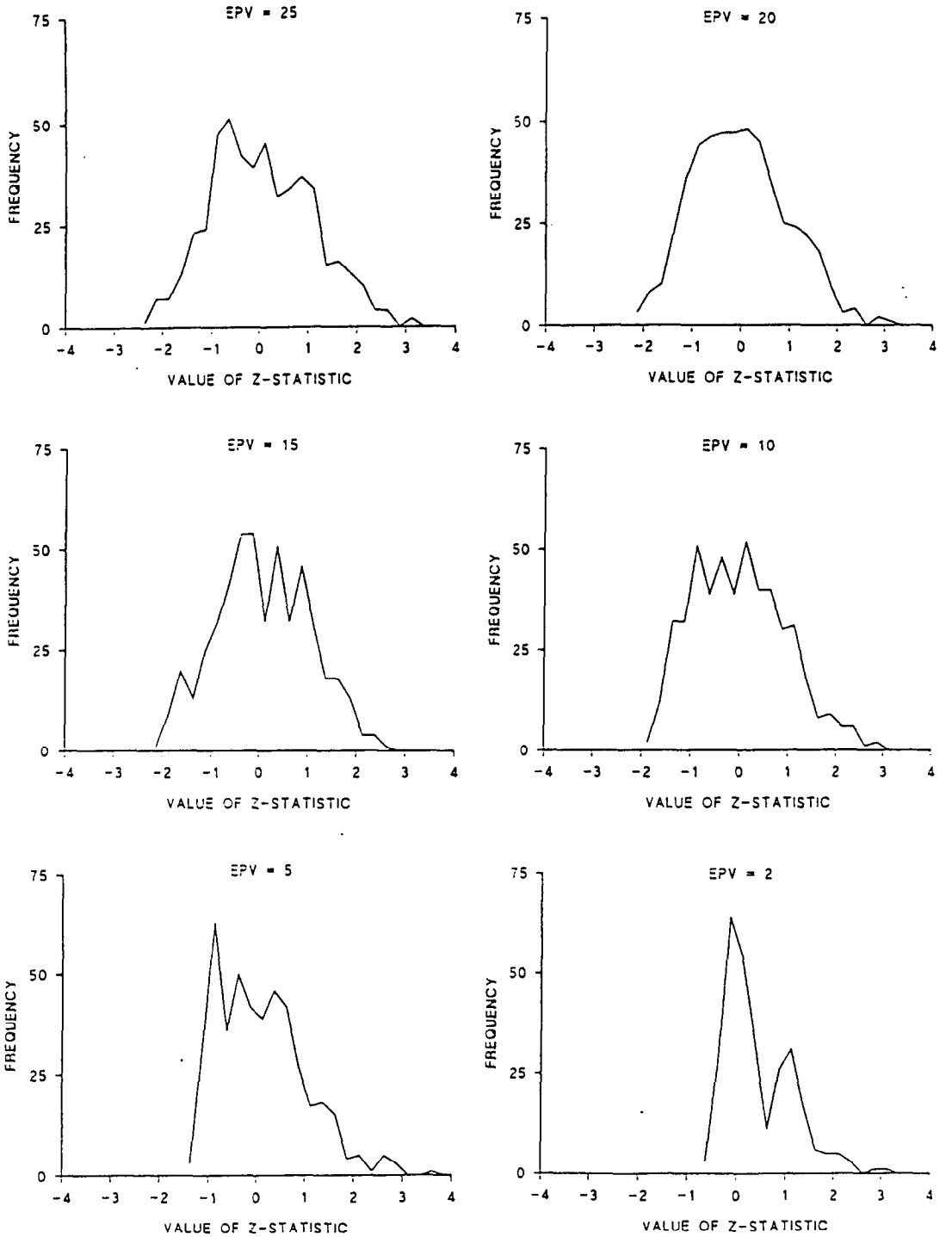

Fig. 10. Distribution of the Z statistic for congestive heart failure under the null hypothesis that the covariate has no effect with outcome.

gression coefficient increases to produce both overestimation and underestimation of the true effect. At an EPV of 2, the bias was substantial, often exceeding 100% in either direction so that the coefficient was either twice as big or half as small as the true value. As EPV decreased, the distributions of the regression coefficients departed from normality, increasing the chance of falsely extreme values.

The impact of low EPV on variance and power was not surprising. The variance of the proportional hazards model is inversely related to the

Table 1. Number and percentage of occasions in which variable was significant ($p < 0.10$) under null hypothesis of no covariate effects

	EPV = 25		EPV = 20		EPV = 15		EPV = 10		EPV = 5		EPV = 2	
Variable[a]	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
Converged[b]	500		500		500		498		449		288	
STD	51	10.2	51	10.2	48	9.6	48	9.6	36	8.0	18	6.3
HTN	44	8.8	44	8.8	43	8.6	53	10.6	44	9.8	22	7.6
NYHA	47	9.4	47	9.4	47	9.4	41	8.2	30	6.7	24	8.3
CHF	45	9.0	45	9.0	45	9.0	31	6.2	25	5.6	15	5.2
DM	48	9.6	48	9.6	48	9.6	46	9.2	39	8.7	22	7.6
VES	47	9.4	47	9.4	43	8.6	49	9.8	44	9.8	26	9.0
LVC	58	11.6	58	11.6	53	10.6	46	9.2	44	9.8	18	6.3
Overall[c]	340	9.7	340	9.7	327	9.3	314	9.0	262	8.3	145	7.2

[a]Abbreviations defined in Patient Population and Study Variables (Part I).
[b]Number of proportional hazards analyses that converged out of 500 samples.
[c]*p* Value for overall was calculated as (number of variables significant)/(total number of variables evaluated = 7 × number of samples that converged).

number of events [4] and can be expected to increase when EPV becomes low. An important and surprising finding, however, is that the large sample properties of proportional hazards model variance may not hold at low EPV, where the values of variance were both overestimated (up to 80%) and underestimated (up to 20%).

The power to detect significant effects also decreased with decreasing EPV, leading to problems of underfitting. Significant effects could also be identified in the wrong direction (i.e., negative instead of positive association with outcome), but such paradoxical fittings were infrequent.

Low EPV also produced problems in significance testing. The coverage of the calculated 90% confidence intervals about the estimated values was not 90%, with a suggestion of wider-than-expected limits at low levels of EPV. Perhaps the most striking finding at very low EPV was that the *Z* statistic (or Wald statistic) was not valid under the null hypothesis. At low EPV, the statistic did not have a Gaussian distribution, and the narrow distribution led to an overly conservative test in which the null hypothesis was rejected less often than the stated significance level. (We did not evaluate the effects of EPV for the two other analogous statistics—likelihood ratio and score—in the proportional hazards model.)

An additional and fundamental problem at low EPV was the frequent inability of the proportional hazards model to converge when coefficients were calculated by the maximum likelihood method. In some instances, the likelihood function converged but not the regression coefficients, and at low EPV, neither entity converged, indicating that the number of outcome events was inadequate.

The simulations in this research could address only a limited number of issues. As an initial investigation, we examined a single data set containing six binary variables and one ordinal variable. We assumed that the results with the original seven-variable model were a gold standard. The prevalence of a positive value in the binary variables ranged from 7 to 59%, and lower prevalence increased the susceptibility to EPV problems (e.g., for the variable "congestive heart failure"). The evaluated variables had only a moderate association with outcome; whether strength of associations modifies the impact of EPV remains to be clarified. Overfitting could not be evaluated because we did not include any "insignificant" variables. Thus, the observed results may differ from other data sets that include continuous variables, a wider range of prevalence, and stronger or nonsignificant associations.

We did not address the impact of interaction terms or the effect of sequential selection of variables in forward and backward "stepping" procedures. Proportional hazards analyses including these features may be more prone to the cited problems, such as inflation of type I error or overfitting. Thus, the current study is probably a conservative demonstration of problems, inherent with low EPV, that will occur in other data sets and with other analytical strategies.

Finally, by varying the number of outcome events in the simulations (keeping the number of independent variables constant), the question of whether results at each EPV would be similar if the number of independent variables were varied (with a constant number of events) cannot be answered. In this population with initial 36 EPV = 252 deaths/7 variables, many additional independent variables would be required—but were

unavailable in our data set—to obtain desired values of EPV (e.g., 2 EPV = 252 deaths/126 variables). An alternative strategy would be to vary both the number of events and independent variables (e.g., 2 EPV = 8 deaths/4 variables). Both of these approaches, however, would create a new problem of defining gold standard results for models with a diverse number of independent variables (e.g., what are the "true" values of regression coefficients for 7-, 126-, and 4-variable simulations?). Accordingly, a general question remains concerning whether "too few" outcome events, or "too many" independent variables, could affect proportional hazards analysis in addition to (or regardless of) the EPV ratio itself.

The research findings have important implications for interpreting published results of studies having relatively few outcome events per variable analyzed. The accuracy, precision, and significance of the coefficients estimated by the proportional hazards method will become untrustworthy when EPV is too low. At EPV $\leq$ 10, the regression coefficients become increasingly biased; confidence intervals may not have the proper coverage; the test statistics may not be valid for the model; the loss of power to identify important associations may lead to underfitting; and the frequency of paradoxical associations may increase. Our results can also be used in clinical investigation to guide the parsimonious selection of independent variables and to gauge the adequacy of the number of outcome events.

In conclusion, the conduct and reporting of proportional hazards analysis, and perhaps other multivariable methods, should include recognition of the importance of the number of events per independent variable. The results should be cautiously interpreted in studies having fewer than 10 events per variable analyzed.

Acknowledgment—This research was supported by the Department of Veterans Affairs Cooperative Studies Program of the Veterans Health Services and Research Administration.

REFERENCES

1. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. **Ann Intern Med** 1993; 118: 201–210.
2. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. **Cancer Treat Rep** 1985; 69: 1071–1077.
3. Stephens MA. EDF statistics for goodness of fit and some comparisons. **J Am Stat Assoc** 1974; 69: 730–737.
4. Tsiatis A. A large sample study of Cox's regression model. **Ann Stat** 1981; 9: 93–108.