BEHAVIOR OF THE BRESLOW-STORER MODEL IN SOME PROSPECTIVE STUDIES

Y.H. Schukken[14], C.E. McCulloch[23], H.N. Erb[1].

BU-925-MA January 1988

[1] Section of Epidemiology, Department of Clinical Sciences

New York State College of Veterinary Medicine

Cornell University, 14853 Ithaca N.Y.

[2] Biometrics Unit

New York State College of Agriculture and Life Sciences

Cornell University, 14853 Ithaca N.Y.

[3] Corresponding author

[4] Present address:

Department of Herd Health and Ambulatory Clinic

University of Utrecht, P.O. Box 80.152

3508 TD, Utrecht, the Netherlands

Summary

Multivariate methods often include implicit assumptions on the relation (additive or multiplicative) between the risk factors in the model. Recently, Breslow and Storer proposed a generalized logistic model in which the relation between the risk factors is estimated by a separate parameter. In this study, the properties of this model were investigated in the environment of a prospective study. Both data from original research and simulated data were used. The Breslow-Storer model showed in a large proportion of cases a decreased ability to estimate the parameters. The problems in estimating the parameters decreased if the sample size increased and if the range of the probabilities increased. However, even in the most favorable design studied, the proportion of non-convergence was still .10. It was concluded that the Breslow-Storer model has limited practical use in small-scale prospective studies.

Introduction

Widely used multivariate analysis techniques to estimate the impact of potential risk factors assume that the effects (e.g., odds ratios) associated with the risk factors combine multiplicatively. These techniques include the log-linear (like the logistic) and proportional hazards models. Models in which the combination of factors is assumed to be additive also exist (1,2). However, there is not always biological reasoning to support the assumption that the risk factors should indeed act in a multiplicative or additive fashion. The models mentioned above are used because of the interpretability of the estimated parameters and the availability of software (1,2). The assumption of multiplication or addition can be relaxed somewhat by adding interaction terms to the models, however the precise nature of the relation of risk factors is not clarified.

Recently, Breslow and Storer (3) proposed a generalized logistic model for case-control studies in which the nature of the relation between risk factors can be estimated from the data. The Breslow-Storer model is given by:

$$\text{logit } (y|x,j) = \alpha_j + \log R(x)$$

where: j indexes the J strata, α indicates the logarithm of the baseline odds and

$$\log R(x) = \begin{cases} \dfrac{(1 + x\beta)^{\theta} - 1}{\theta} & \text{for } \theta \neq 0 \\ \log\ (1 + x\beta) & \text{for } \theta = 0 \end{cases}$$

y = outcome variable 1= event, 0= no-event

x = vector of explanatory variables

β = vector of parameters, estimates of the risk ratio

θ = parameter indicating the nature of the relation between the explanatory variables and logit(y).

Note that for $\theta = 1$ the model reduces to the usual logistic model - indicating that the effect of the risk factors is multiplicative - and for $\theta = 0$ the model yields an additive effect of the risk factors on the odds. For a detailed description of the model see reference 3.

The purpose of this paper is to study the behavior of the model in the setting of a prospective study.

Data

Retained placenta study

A study by Gunnink (4) showed that a decrease in migration activity of peripheral leukocytes and a decreased number of leukocytes in the cotyledons are a sufficient cause for retained placenta (RP) in dairy cows. A decrease in activity of leukocytes also may put an animal at higher risk for infectious diseases. Finally, the risk of dying of an animal may be higher for animals encountering retained placenta, infectious diseases or both. A cohort study was designed to test this hypothesis.

A sample of the female Holstein dairy cows admitted to the New York State College of Veterinary Medicine, Cornell University, between 1/1/1980 and 1/1/1986 was taken. Sampling was done on retained placenta status. Sixty-two RP cases (all available) and 125 non-RP cases were sampled using a computerized random number generator. The risk of dying in the hospital was modelled using a logistic model; abdominal disorder, corticosteroid use, infectious disease and RP were the most important predictors for dying. An interaction between RP and infectious disease turned out not to be significant ($p = .12$), but had, if added to the model, a severe impact on the point estimates of the parameters. A further investigation into the nature of the relation between RP and infectious disease was prompted.

Since the main purpose was to estimate the nature of the relation between RP and infectious disease, the analysis was stratified on the other explanatory variables. Both abdominal disorder and corticosteroid use are binary variables, defining four strata (J=4). The data set is listed in Appendix III.

Simulation study

Simulated random samples were generated from the Breslow Storer model with $\alpha = -4$, $\beta_1 = \beta_2 = 1$ and $\theta = 1$ (since $\theta = 1$ this is also the logistic model). Defining α at -4 means that the baseline risk is equal to:

$$[1+\exp(-(-4))]^{-1} = .018.$$

The predictor variables were fixed, resembling a prospective study. Three factorial designs were studied. Each design had two explanatory variables, x_1 and x_2. The values of the variables and the ranges of risk generated are listed below.

DESIGNS FOR THE SIMULATION

	Variables		Range of Risk
Design A:	x_1	0,1,2	.018 - .119
	x_2	0,1	
Design B:	x_1	0,1,2,3	.018 - .731
	x_2	0,1,2	
Design C:	x_1	0,1,2,3,4,5	.018 - .993
	x_2	0,1,2,3	

In addition, three sample sizes (48, 96, and 192) were used for each of the three designs. The samples sizes were chosen to resemble small and moderate sized prospective studies.

Methods

The models for the RP study were fit using both SAS proc NLIN (5) and GLIM (6). The SAS program is in Appendix I; the GLIM macro is given in reference 3. Initial parameter estimates for the β - vector were taken from the previously-fitted logistic model and the initial estimate for θ was 1. For the SAS program the Marquardt method of least squares estimation was used. GLIM uses a similar method (Newton-Raphson) based on minimizing the deviance (-2 log likelihood ratio). Both methods give maximum likelihood estimates of the parameters. SAS actually calculates the maximum likelihood estimates for β and θ simultaneously using iteratively reweighted least squares, whereas GLIM calculates the maximum likelihood estimates for β using several values for i, supplied by the data analyst. The θ value with the smallest deviance of the model is then an approximation of the maximum likelihood estimate (3). The simulations were performed in the statistics and

matrix language GAUSS (7) on an IBM PC-XT. Maximum likelihood estimates were calculated using the GAUSS algorithm MAXMUM, which will maximize arbitrary likelihood functions. MAXMUM is an iterative maximization routine which uses a Quasi-Newton algorithm. Iterations always were started at the true values of the parameters. All pseudo-random number generation was performed using built-in GAUSS functions. The two models were compared using the same data sets. Further details are available from the corresponding author.

Three different programs were used for two reasons. Firstly, a wide variety of maximization algorithms were used, insuring that problems in fitting the model were not due solely to inadequate software. Secondly, it demonstrates the different software options available when fitting such a model is appropriate.

Results

The Marquardt least squares approach applied to the RP data did not reach convergence (relative convergence criterion = 10E-8). The least squares estimates of the parameters bounced over a wide range of values but did not settle down after 100 iterations. The iteration values of θ ranged between -324 and 93027. It was decided to restrict the values of the parameters to a range with biological interpretability. The parameters for retained placenta and infectious disease were restricted to values larger than zero, and the exponent was restricted to values between three and minus three. After this restriction, convergence was met after 31 iterations. The estimates of the parameters and the standard errors from the SAS output are given in table 1. Note that standard errors estimated as SAS does will usually be incorrect when the maximum likelihood estimtion falls on the boundary as it did in this case. The ANOVA table and correlation matrix of the parameter estimates from the SAS output are in Appendix II.

The observation that θ was estimated at the upper bound of the restriction (3.000) suggests that without this bound the estimate would go up.

The estimating procedure using the Newton-Raphson method in GLIM showed similar

results: with increasing values for θ the deviance kept getting smaller. The slope of the deviance graph decreased but did not reach zero. The results are shown graphically in Figure 1. A selected number of θ values and the corresponding deviances and estimates of b_1, b_2 are given in Table 2.

The results of the simulations are given in Table 3, which shows the proportion of cases in which the maximum likelihood estimates could not be calculated using the GAUSS package. This occurred in a large percentage of cases for the Breslow-Storer model. The logistic model fared much better.

In Table 4, some values of the estimates corresponding to a typical simulation run (Design B, sample size 96) are reported. There were two typical patterns when the likelihood maximization routine failed to converge: either the estimate of θ grew larger and larger as the estimates of β_1 and β_2 tended towards zero (as in the RP data) or vice versa. In either case, the likelihood was slowly increasing without bound.

Discussion

The large standard errors of the parameters and the small slope of the deviance indicate that no stable estimate of θ could be calculated in the RP data. This signifies that it is hard if not impossible to make any inferences on the biologic interrelation between retained placenta and infectious diseases and their effect on the death risk in the hospital. It seems more likely that the relation is multiplicative or supra-multiplicative rather than additive or sub-additive.

It is reassuring to see that the point estimates of the θ parameters at $\beta = 3$ are equivalent in the SAS Proc NLIN and the GLIM estimation methods.

The results of the simulation show that use of the Breslow-Storer model caused problems in estimating the parameters. In the situation with a design spanning the widest range of risk (Design C) and a moderate sample size (96), it was not possible to calculate the estimates in a third of the cases. With smaller sample sizes and more restricted ranges of risks the results were worse. The logistic model, in comparison, gave estimates in nearly all

the studied situations. Only for the design with the smallest range of risk (Design A) and the smallest number of observations did the logistic model fail to give convergence for a significant proportion of cases. Perhaps more importantly, the estimates of the Breslow-Storer model, when convergence was reached were often farther from the true values than the logistic model. Also, graphical analyses (as advocated in Breslow and Storer, 1985) of the situations when convergence was not reached would have suggested supra-multiplicative or sub-additive models for data generated from an multiplicative model.

The theoretical advantages of developing models which potentially explain more of the data under consideration are in some instances outweighed by difficulties in estimating the parameters. The addition of an extra parameter (θ) to the logistic model may yield uninterpretable results or no results at all. Especially estimating coefficients of interaction-like terms (like θ) do require quite a large sample size.

A regression model is linear if the prediction equation is a linear function of the parameters and the predictor variables. Most ordinary least square models are of the form :

$$Y = a + b_1X_1 + b_2X_2$$

and are therefore linear models. The logistic model is a non-linear model, because the probability of an event is modelled as

$$Pr(Y=1) = [1 + \exp\text{-}(a + b_1X_1 + b_2X_2)]^{-1} \quad .$$

However, with a transformation of the outcome variable this model can be linearized as :

$$\ln(Pr(Y=1)/1\text{-}Pr(Y=1)) = a + b_1X_1 + b_2X_2 \quad .$$

Hence, the logistic model is defined as a linearizable or a generalized linear model. The Breslow-Storer model cannot be linearized and is therefore a true non-linear model.

It is possible to estimate the degree of non-linearity on a scale developed by Bates and Watts (8) however, this estimate depends on both the degree of non-linearity of the model and the range of the supplied X-values. With the increase of non-linearity in the model, the chance that convergence will not be met in certain data sets also increases.

As shown by examples in the Breslow-Storer paper, in some instances there will be a finite maximum likelihood estimate of θ, while in our data there was not. One of the

differences between our data and the data from these studies is the sampling scheme. Our data are sampled on predictors, while the data in the Breslow-Storer paper were sampled on outcome. In a case-control study there are by the nature of the design a high proportion of cases per stratum. The data points will be scattered over a wide range of the outcome variable, restricting the estimates to a limited area. In a prospective study, the probability of disease is small even in the high-risk categories. This means that the observations are relatively close together in a small area of the outcome variable. The results of the analysis from the original data set and the simulations indicate that the Breslow-Storer model has problems in estimating the parameters from this limited information. Figure 2 shows the Breslow-Storer model with α set at -4. It can be seen that the lines are virtually identical for different values of the exponent at low values of the outcome variable.

Another difference between our study and the examples in reference 3 is the sample size. Our study showed that with small sample sizes the Breslow-Storer model is difficult to fit and can give misleading results. However, with larger sample sizes (such as in the examples of Breslow and Storer; 3) estimation is likely to be easier. This was evident with the improvement in estimation with increasing sample size (Table 3).

Sampling on outcome rather than on predictor will increase the range over which the data are measured and will ease the fitting of a mathematical model. Since the Breslow-Storer model is very non-linear it is likely that it will be most useful in case-control studies with moderate to large sample size.

References

1. Storer, B.E. Wacholder, S.Breslow,N.E. Maximum likelihood fitting of general risk models to stratified data. *Applied Statistics* 32: 172-81 , 1983.

2. Barlow, W.E. General relative risk models in stratified epidemiological studies. *Applied Statistics* 34: 246-57, 1985.

3. Breslow, N.E. Storer, B.E. General relative risk functions for case-control studies. *American Journal of Epidemiology* 122: 149-62, 1985.

4. Gunnink, J.W. Retained placenta and leukocyte activity. *The Veterinary Quarterly* 6: 49-59, 1984.

5. *SAS Institute Inc. SAS User's Guide: Statistics, Version* 5 *Edition.* Cary, NC: SAS Institute Inc., 1985.

6. Baker, R.J., Nelder, J.A. *The GLIM System Release* 3, Numerical Algorithms Group, Oxford, 1978.

7. *GAUSS manual, Version* 1.49*B*, Aptech Systems Inc., Kent WA, 1986.

8. Bates, D.M., Watts, D.G. Relative curvature measures of non linearity. *Journal of the Royal Statistical Society,* (B)42: 1-25, 1980.

List of abbreviations

df Degrees of freedom

RP Retained placenta

SS Sum of Squares

Appendices

I. SAS Proc NLIN program

II. Anova table and correlation matrix of the parameters

III. RP data set

Appendix I

SAS Program

```
OPTIONS LS=80;
TITLE 'FITTING THE BRESLOW-STORER MODEL TO R.P. COHORTS';
DATA;
INFILE RPCOH;                          reading the data from file
INPUT S1 S2 S3 X1 X2 Y;
PROC NLIN METHOD = MARQUARDT MAXITER=100 NOHALVE SIGSQ=1;
PARAMETERS A1 = 0
           A2 = 0
           A3 = 0                      initial parameter values
           A4 = 0
           B1 = .931
           B2 = 2.81
            L = 1;
BOUNDS -3<L<3, B1>0, B2>0;             restricting the parameter ranges
E=EXP(-1*(A1+A2*S1+A3*S2+A4*S3+((((1+B1*X1+B2*X2)**L)-1)/L)));
M=((1+B1*X1+B2*X2)**(L-1));
K=(-1*((1/(L**2))*((1+B1*X1+B2*X2)**L))
+((1/L)*(LOG(1+B1*X1+B2*X2))*((1+B1*X1+B2*X2)**L))+(L**-2));
P=1/(1+E);
D=((1+E)**-2);
MODEL Y = P;                           model specification
W=1/(P*(1-P));                         binomial weight
_WEIGHT_=W;
DER.A1=D*E;
DER.A2=D*E*S2;
DER.A3=D*E*S3;                         specification of the partial
DER.A4=D*E*S4;                         derivatives
DER.B1=D*E*M*X1;
DER.B2=D*E*M*X2;
DER.L=D*E*K;
```

Appendix II

ANOVA table of restricted MLE, R.P. Study

Source	df	SS	Weighted SS
Regression	7	150.844	21.549
Residual	209	220.191	1.05
Uncorrected total	216	371.035	
(Corrected total)	215	322.868	

Asymptotic correlation matrix (strata parameters, n=4, excluded)

	b_1	b_2	exp
b1	1	.9693	-.9794
b2	.9693	1	-.9964
exp	-.9794	-.9964	1

Appendix III

RP DATA SET

Stratum	RP	ID	N	R
1	0	0	68	4
1	1	0	21	1
1	0	1	12	8
1	1	1	12	10
2	0	0	7	0
2	1	0	6	1
2	0	1	3	2
2	1	1	1	0
3	0	0	10	2
3	1	0	1	0
3	0	1	7	3
3	1	1	5	5
4	0	0	10	2
4	1	0	11	0
4	0	1	9	3
4	1	1	4	3

Stratum 1 = no abdominal disorder, no corticosteroid use

2 = abdominal disorder, no corticosteroid use

3 = no abdominal disorder, corticosteroid use

4 = abdominal disorder, corticosteroid use

RP = retained placenta (1=yes, 0=no)

ID = infectious disease (1=yes, 0=no)

N = size of cohort

R = # deaths

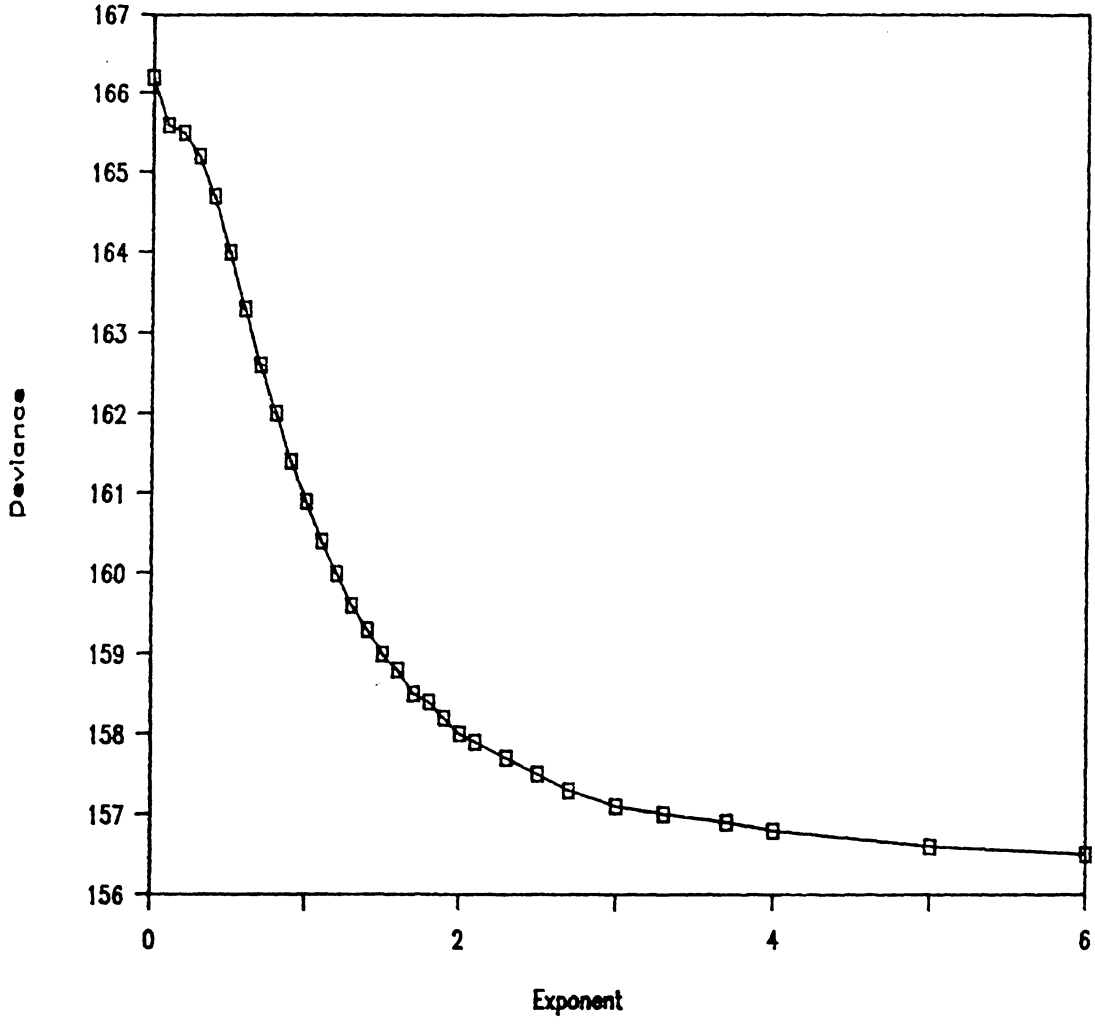

Figure 1. Exponent (θ) of the Breslow-Storer model versus the Deviance, RP data

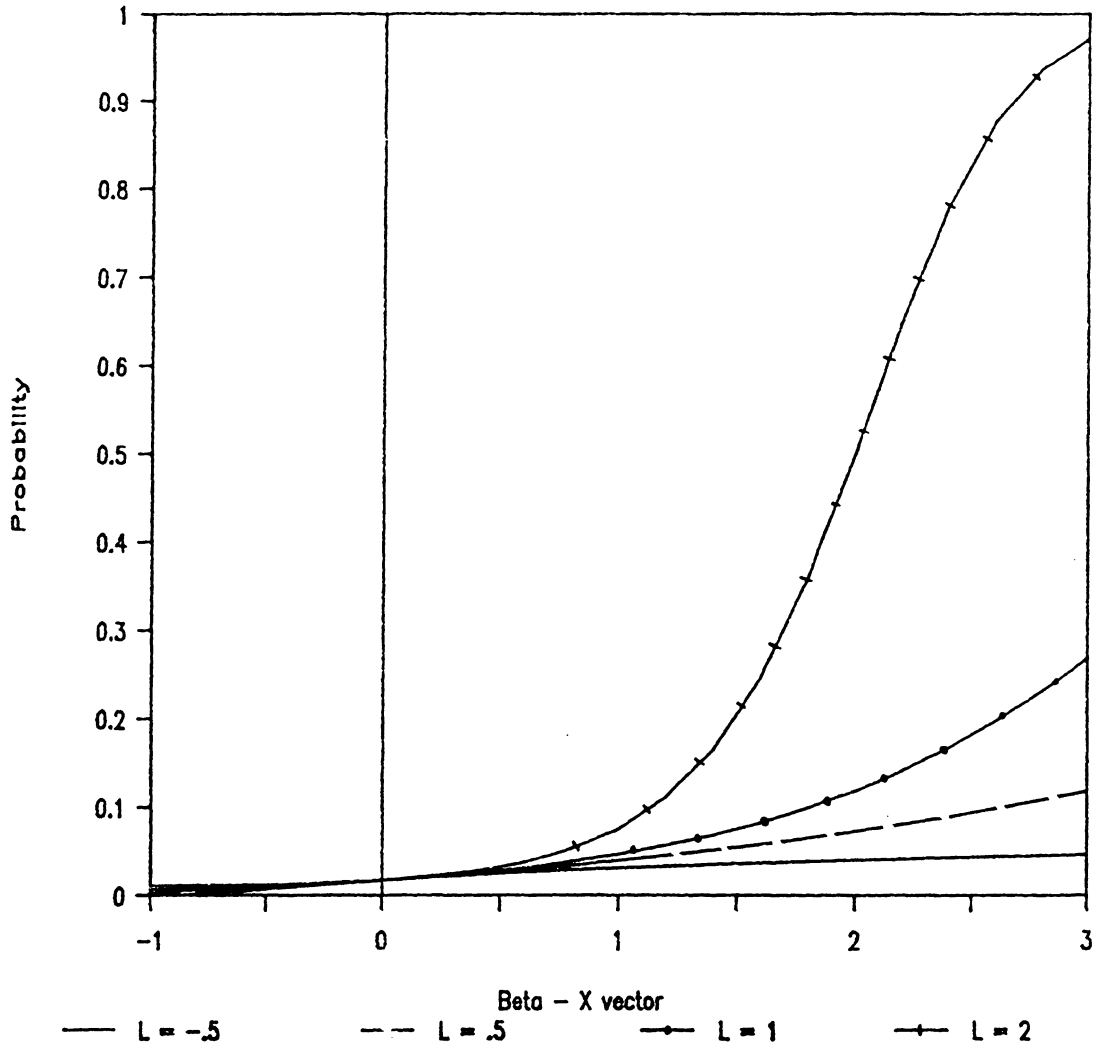

Figure 2. Graph of the Breslow-Storer model with several values for the exponent ($\theta = L$), α set at -4

Table 1. Weighted least square estimates of the parameters (SAS) from RP Study

Parameter	Estimate	Standard error
b_1[a]	.3299	.5705
b_2[b]	1.0361	1.2023
exponent	3.0000	3.9754

[a]parameter for X_1, indicating retained placenta

[b]parameter for X_2, indicating infectious disease

Table 2. Deviances and parameter estimates (GLIM) from RP Study

θ	Deviance(df=210)	b_1[a]	b_2[b]
0	166.2	-.8494	15.74
.1	165.6	.0419	10.61
.5	164	.9917	4.919
1	160.9	.9306	2.808
1.5	159	.6888	1.914
2	158.1	.5194	1.490
3	157.1	.3299	1.036
4	156.8	.2571	.8633
6	156.5	.1429	.5731

[a]parameter for X_1, indicating retained placenta

[b]parameter for X_2, indicating infectious disease

Table 3. Results of the simulation study (GAUSS). Proportion of cases not converging

	Design		
Sample size	A	B	C
48	.5[a] 1[b] (10)[c]	0 .67 (30)	0 .5 (10)
96	.033 .9 (30)	0 .533 (30)	0 .36 (50)
192	0 .8 (10)	0 .42 (50)	0 .1 (10)

[a]Logistic model

[b]Breslow-Storer model

[c]number of replications