STATISTICS IN MEDICINE, VOL. 9, 1501–1515 (1990)

RELATION OF POOLED LOGISTIC REGRESSION TO TIME DEPENDENT COX REGRESSION ANALYSIS: THE FRAMINGHAM HEART STUDY

RALPH B. D'AGOSTINO, MEI-LING LEE AND ALBERT J. BELANGER

Mathematics Department, Boston University, 111 Cummington Street, Boston, MA 02215, U.S.A.

L. ADRIENNE CUPPLES

Boston University, School of Public Health, Epidemiology and Biostatistics, 80 East Concord Street, Boston, MA 02118, U.S.A.

KEAVEN ANDERSON

National Heart, Lung and Blood Institute, Framingham Heart Study, 118 Lincoln Street, Framingham, MA 01701, U.S.A.

AND

WILLIAM B. KANNEL

Boston University, School of Medicine, Preventive Medicine and Epidemiology, 720 Harrison Avenue, Boston, MA 02118, U.S.A.

SUMMARY

A standard analysis of the Framingham Heart Study data is a generalized person-years approach in which risk factors or covariates are measured every two years with a follow-up between these measurement times to observe the occurrence of events such as cardiovascular disease. Observations over multiple intervals are pooled into a single sample and a logistic regression is employed to relate the risk factors to the occurrence of the event. We show that this pooled logistic regression is close to the time dependent covariate Cox regression analysis. Numerical examples covering a variety of sample sizes and proportions of events display the closeness of this relationship in situations typical of the Framingham Study. A proof of the relationship and the necessary conditions are given in the Appendix.

INTRODUCTION

The investigators of the Framingham Heart Study have collected data prospectively every two years since 1948 on a cohort of 5209 subjects, aged 30 to 62 years at the beginning of the study, to examine the relationship of potential risk factors to the development of cardiovascular disease.[1,2] They have measured potential risk factors repeatedly and recorded the times to various disease endpoints. When relating risk factors to disease, the former are regarded as independent variables and the times to diagnosis or death from disease as dependent variables. One question in medical, epidemiological and statistical research is how to use these repeated measures to evaluate the relationship of a risk factor to disease development.

A method developed and used to analyse data from the Framingham Study cohort[3–5] incorporates all repeated observations in a generalized person-exam technique. This method has

0277–6715/90/121501–15$07.50

Received September 1989
Revised May 1990

been called the pooling of repeated observations (PRO) method.[6] Treating each two year examination interval as a mini follow-up study, the PRO method pools observations over all intervals to examine the short-term development of disease. The relation of the pooled observations to the disease endpoints is analysed using a logistic regression model. This analysis is a special case of the Wu and Ware method[7] and is basically a pooled logistic regression.

Another statistical technique which may be applied to this problem is the Cox time dependent covariate regression model.[8, 9] In this model the risk factors are updated at each exam and are the time dependent covariates. The purpose of our paper is to demonstrate the relationship between the PRO method (cross-sectional pooled logistic) and the Cox time dependent covariate regression method. When the intervals between measurements (exams) are short, the probability of an event within an interval is small, and the intercept for the pooled logistic is constant across intervals, the methods are asymptotically equivalent. We demonstrate this numerically with a number of examples and supply the outline of a proof in the Appendix.

The relationship between the Cox model and logistic regression has been discussed previously. Thompson[10] showed that as grouping intervals for the events approach zero the logistic model tends towards the Cox model. He commented on the validity of this observation for time dependent covariates but did not prove it. Abbott[11] showed the relation between a grouped Cox model and the logistic model when the interval lengths approach zero. He did not consider the case of time dependent covariates. Recently Ingram and Kleinman,[12] using simulated data sets and real data sets, compared parameter estimates from the proportional hazards model, the logistic model and a person-time logistic model. The latter is a PRO method procedure, except that in Ingram and Kleinman only fixed covariates were considered. Thus none of the above has considered explicitly the case of time dependent covariates.

SAMPLING DESIGN APPROPRIATE FOR THE COMPARISON

The sampling design employed in cohort follow-up studies such as the Framingham Heart Study includes repeated measurements on risk factors (independent variables) but only one measurement is recorded on outcome (dependent) variables. Typically, the outcome measurement is a survival time – the length of time from entry into the study until an outcome event of interest occurs, such as a myocardial infarction. The investigator is typically interested in initial (or primary) outcomes, or in outcomes that can only happen once (for example, death). This sampling design should be distinguished from other longitudinal studies in which outcomes (that is, the dependent variables) are also measured repeatedly, for example, studies of the effect of air pollution on multiple asthma attacks.[13]

THE POOLED LOGISTIC REGRESSION METHOD

Cupples, D'Agostino, Anderson and Kannel[6] describe extensively pooled logistic regression analysis. Figure 1 illustrates the type of data in our study. We consider a hypothetical study of 1000 persons at risk of a disease. All of these have risk factors measured at time t_0 (or exam 1). We follow them through the first interval of observation. During that time period 40 develop the disease and 10 are lost to follow-up. We remove these 50 from the population at risk. At time t_1 (exam 2) there are 950 subjects on whom risk factors are measured. Of these 25 develop the disease and 5 are lost to follow-up. The remaining 920 have risk factors measured at time t_2 (exam 3), of which 20 develop the disease in the next period and 10 are lost to follow-up. The PRO

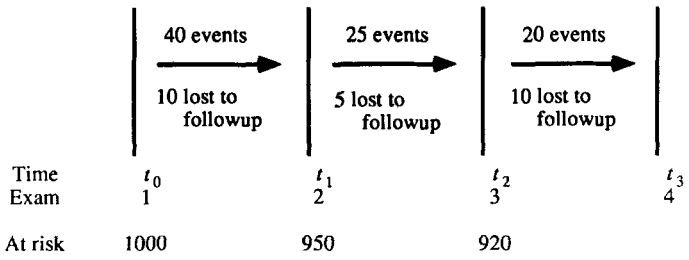

Figure 1. Pooling of repeated observations (PRO)

method pools the subjects at risk in each interval to yield 1000 + 950 + 920 = 2870 person-exams and pools the 40 + 25 + 20 = 85 events. A logistic regression with 2870 observations and 85 events constitutes the pooled logistic regression analysis.

Mathematically the logistic regression model is written:

$$\text{logit}\, q_i(\mathbf{X}(t_{i-1})) = \log\left(\frac{q_i(\mathbf{X}(t_{i-1}))}{1 - q_i(\mathbf{X}(t_{i-1}))_i}\right) = \alpha_i + \gamma_i X_1(t_{i-1}) + \ \ldots\ + \gamma_p X_p(t_{i-1}), \tag{1}$$

where $q_i(\mathbf{X}(t_{i-1}))$ is the conditional probability of observing an event by time t_i given that the individual is event free at time t_{i-1}, and

$$\mathbf{X}(t_{i-1}) = (X_1(t_{i-1}), \ldots, X_p(t_{i-1})) \tag{2}$$

is the vector of risk factors measured at time t_{i-1}. In model (1) the intercept α_i is a function of the time interval between exam $i-1$ and exam i. It is often assumed to be constant over time, in which case α_i in (1) is replaced with α. Cupples, D'Agostino, Anderson and Kannel[6] discuss in detail the assumptions underlying the valid use of this model and extensions of it.

TIME DEPENDENT COVARIATE COX REGRESSION

The covariates in this model are the risk factors recorded at times t_0, t_1 and t_2 (Figure 1). If the observations are grouped into the intervals $[t_{i-1}, t_i)$, and the conditional probability $p_i(\mathbf{X}(t_{i-1}))$ denotes the probability that an individual will survive up to time t_i given survival already up to time t_{i-1}, then

$$p_i(\mathbf{X}(t_{i-1})) = \exp\left\{-\int_{t_{i-1}}^{t_i} h_0(u)\exp[\boldsymbol{\beta}'\mathbf{X}(t_{i-1})]\,du\right\}. \tag{3}$$

Here $h_0(u)$ is the baseline hazard rate and

$$\boldsymbol{\beta}'\mathbf{X}(t_{i-1}) = \beta_1 X_1(t_{i-1}) + \ \ldots\ + \beta_1 X_p(t_{i-1}) \tag{4}$$

is the linear function of the Cox proportional hazard model. See the Appendix for a derivation of (3) and (4).

In the above Cox model the events are grouped into intervals $[t_{i-1}, t_i)$ but not specified as to a time of occurrence within the intervals. This is the grouped Cox model. In the Appendix we prove the asymptotic equivalence of the pooled logistic regression to the grouped Cox model analysis.

In the next section we give numerical comparisons of these two methods and also present for comparison numerical results for the Cox time dependent covariate method where exact times of events are used. We refer to this latter model as the continuous Cox model. The proof in the Appendix can also be extended, under suitable conditions, to this model.

NUMERICAL COMPARISONS

In (1) and (4) we see that both the pooled logistic and time dependent Cox models involve linear functions of the risk factors $\mathbf{X}(t_{i-1})$ measured at time t_{i-1}. For the Framingham Study there are 20 biennial exams. In the following we present results of analyses on risk factor data for exams 3 to 18 and exams 9 to 18, the former covering 15 intervals and a thirty year follow-up and the latter having 9 intervals and an eighteen year follow-up. The objective of these numerical examples is to compare estimated coefficients for the risk factors from the pooled logistic regression and the Cox time dependent covariate regression models. In particular we discuss the percentage differences as measured by

$$R1 = 100\frac{\hat{\gamma} - \hat{\beta}}{\hat{\gamma}}, \tag{5}$$

where $\hat{\gamma}$ and $\hat{\beta}$ are the estimates from the logistic and Cox models, respectively. We also compare the relative risk for the Cox model to the odds ratio of the logistic model. This we measure by

$$R2 = 100\frac{e^{\hat{\gamma}u} - e^{\hat{\beta}u}}{e^{\hat{\gamma}u}}, \tag{6}$$

where u is an appropriate increase in the scale of the risk factor (e.g. an increase of 10 cigarettes per day for the variable number of cigarettes per day or an increase of 10 years for the variable age). While $R2$ actually compares relative risk with the odds ratio, for the sake of brevity we will refer to it as a comparison of relative risks.

To make the comparison over a usefully wide range of situations we have selected examples that vary in the number of subjects, number of events (both as a total and as a fraction of the number of subjects), number of grouping intervals, number of risk factors and significance of risk factors (i.e. including some risk factors that are statistically significant and some which are not). The examples are presented mainly for comparison of the methods; the reader should not view them for definitive substantive interpretation.

Example 1: Coronary Heart Disease (CHD)

Tables I, II and III display the results of examples where the endpoint of interest is coronary heart disease (CHD). In Table I there are 2078 male subjects aged 35 to 64 at exam 3. These are followed for thirty years to exam 18. Three major risk factors (age, systolic blood pressure (SBP) and the number of cigarettes per day (CSM)) are included in the regressions. Of the 2078 subjects, 783 or 37·7 per cent ultimately developed CHD. An examination of the ratios $R1$ and $R2$ shows good agreement between the pooled logistic regression and the two versions of the Cox regression (i.e. with grouping of the endpoint CHD, referred to as Cox grouped, and with the exact times of occurrence of the endpoint, referred to as Cox continuous). Agreements in $R2$ which compares relative risks are better than the direct comparison of coefficients $R1$. The maximum difference of the former is 4·3 per cent and the latter 10·3 per cent. All models agree well with each other.

The reader should note that the greatest difference in the models is for the variable age. This will be a common feature for all the numerical examples to follow. Age is a suitable variable for

Table I. Comparison of regression coefficients: thirty year follow-up with three major risk factors

Population: males, 35–64 at exam 3, free of CHD
Endpoint: CHD
Exams 3 to 18, 15 grouping intervals
Variables: Age, systolic blood pressure (SBP), number of cigarettes per day (CSM)
Basis: 2078 subjects; 783 events; 21,268 person exams

Pooled logistic

	$\hat{\gamma}$	SE($\hat{\gamma}$)	z	exp($\hat{\gamma}u$)*
Age	0·0427	0·0038	11·21	1·533
SBP	0·0160	0·0015	10·45	1·174
CSM	0·0142	0·0026	5·48	1·153

Cox grouped

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0383	0·0047	8·13	1·467
SBP	0·0152	0·0015	10·30	1·164
CSM	0·0134	0·0025	5·35	1·143

Cox continuous

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0383	0·0047	8·16	1·467
SBP	0·0153	0·0015	10·32	1·165
CSM	0·0142	0·0025	5·71	1·153

Comparison of coefficients†

	Logistic vs. Cox grouped		Logistic vs. Cox cont.		Cox grouped vs. Cox cont.‡	
	$R1$	$R2$	$R1$	$R2$	$R1$	$R2$
Age	10·3	4·3	10·3	4·3	0·0	0·0
SBP	5·0	0·8	4·4	0·7	− 0·7	− 0·1
CSM	5·6	0·8	0·0	0·0	− 6·0	− 0·9

* $u = 10$ for all values.
† $R1 = 100(\hat{\gamma} - \hat{\beta})/\hat{\gamma}$, $R2 = 100(e^{\hat{\gamma}u} - e^{\hat{\beta}u})/e^{\hat{\gamma}u}$.*
‡ Cox regression with grouped data is used as denominator for the comparisons.

the logistic model but may not always be ideal for the Cox models. In the extreme case where all individuals start at the same age, the time dependent variable Age would be perfectly correlated with follow-up time. The effect of age in the Cox models would then be reflected only in the underlying hazard function. In the data for Table I there is variability in the initial ages, and significance for the time dependent variable Age is attained in the Cox models, but not at the level attained in the logistic regression. Part of its effect is still confounded with the follow-up time. The other variables in the model (SBP and CSM) do not have this difficulty and the agreement of the three models for them is much better. We discuss the variable Age further in the discussion and conclusion section.

Another important aspect is the computer time employed. BMDP programs were used for all analyses. The CPU times for the analyses were 90 seconds, 7 seconds and 273 seconds for the pooled logistic, grouped Cox and continuous Cox, respectively. Clearly the last is very time intensive. For the example it does not appear to be any more precise than the other two procedures. The grouped Cox is the most time efficient method.

Table II. Comparison of regression coefficients: eighteen year follow-up with three major risk factors

Population: males, 50–69 years old on exam 9, free of CHD
Endpoint: CHD
Exam 9 to 18, 9 grouping intervals
Variables: Age, systolic blood pressure (SBP), number of cigarettes per day (CSM)
Basis: 1154 subjects, 353 events; 7530 person exams

Pooled logistic

	$\hat{\gamma}$	SE($\hat{\gamma}$)	z	exp($\hat{\gamma}u$)*
Age	0·0272	0·0077	3·51	1·313
SBP	0·0158	0·0024	6·49	1·171
CSM	0·0102	0·0040	2·58	1·107

Cox grouped

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0234	0·0094	2·48	1·264
SBP	0·0151	0·0024	6·43	1·163
CSM	0·0096	0·0038	2·50	1·101

Cox continuous

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0184	0·0095	1·95	1·202
SBP	0·0149	0·0023	6·36	1·161
CSM	0·0108	0·0038	2·83	1·114

Comparison of coefficients†

	Logistic vs Cox grouped		Logistic vs Cox cont.		Cox grouped vs Cox cont.‡	
	$R1$	$R2$	$R1$	$R2$	$R1$	$R2$
Age	14·0	3·7	32·4	8·4	21·4	4·9
SBP	4·4	0·7	5·7	0·9	1·3	0·2
CSM	5·9	0·6	− 5·9	− 0·6	− 12·5	− 1·2

* $u = 10$ for all variables.
† $R1 = 100(\hat{\gamma} - \hat{\beta})/\hat{\gamma}$, $R2 = 100(e^{\hat{\gamma}u} - e^{\hat{\beta}u})/e^{\hat{\gamma}u}$.*
‡ Cox regression with grouped data is used as denominator for comparisons.

The analysis reported in Table II is similar to that in Table I except individuals are followed from exam 9 to 18. There are 1154 male subjects and 353 cases of CHD (30·6 per cent). The agreement among the models is again good, especially if $R2$, the comparison of relative risks, is used. Again, the argument is poorest for Age. There is a 32·4 per cent difference between the pooled logistic regression coefficient and the continuous Cox coefficient. The difference in terms of $R2$ is, however, only 8·4 per cent. In general for this example the pooled logistic and the grouped Cox model agree very well. The maximum difference for $R1$ is 14·0 per cent and for $R2$ 3·7 per cent.

The analysis in Table III uses the same subjects as in Table II. In Table III six major risk factors are included. These include the previous three (Age, SBP and CSM) plus serum cholesterol (SCL), metropolitan relative weight (MRW) and glucose intolerance (GLI). All six except for GLI are continuous variables. GLI is a dichotomous variable: 1 for present and 0 for absent. The measurement of these variables is reported elsewhere.[3]

Table III. Comparison of regression coefficients: eighteen year follow-up with six major risk factors

Population: males, 50–69 years old on exam 9, free of CHD
Endpoint: CHD
Exams 9 to 18, 9 grouping intervals
Variables: Age, systolic blood pressure (SBP), cigarettes per day (CSM), total serum cholesterol (SCL), metropolitan relative weight (MRW), glucose intolerance (GLI)
Basis: 1154 subjects; 353 events; 7530 person exams

Pooled logistic

	$\hat{\gamma}$	SE($\hat{\gamma}$)	z	exp($\hat{\gamma}u$)*
Age	0·0312	0·0080	3·91	1·366
SBP	0·0139	0·0025	5·56	1·149
CSM	0·0119	0·0040	2·99	1·126
SCL	0·0041	0·0013	3·07	1·085
MRW	0·0061	0·0034	1·78	1·063
GLI	0·1310	0·0690	1·90	1·140

Cox grouped

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0269	0·0096	2·80	1·309
SBP	0·0134	0·0024	5·53	1·143
CSM	0·0112	0·0039	2·89	1·119
SCL	0·0040	0·0013	3·12	1·083
MRW	0·0054	0·0033	1·64	1·056
GLI	0·1130	0·0669	1·69	1·120

Cox continuous

	$\hat{\beta}$	SE ($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0215	0·0096	2·23	1·240
SBP	0·0132	0·0024	5·46	1·141
CSM	0·0121	0·0038	3·19	1·129
SCL	0·0040	0·0013	3·12	1·083
MRW	0·0050	0·0033	1·50	1·051
GLI	0·1121	0·0667	1·68	1·119

Comparison of coefficients†

	Logistic vs Cox grouped		Logistic vs Cox cont.		Cox grouped vs Cox cont.‡	
	$R1$	$R2$	$R1$	$R2$	$R1$	$R2$
Age	13·8	4·2	31·1	9·2	20·1	5·3
SBP	3·6	0·5	5·0	0·7	1·5	0·2
CSM	5·9	0·7	− 1·7	− 0·2	− 8·0	− 0·9
SCL	2·4	0·2	2·4	0·2	0·0	0·0
MRW	11·5	0·7	18·0	1·1	7·4	0·5
GLI	13·7	1·8	14·4	1·9	0·8	0·1

* $u = 10$ for all variables except SCL where $u = 20$ and GLI where $u = 1$.
† $R1 = 100(\hat{\gamma} - \hat{\beta})/\hat{\gamma}$, $R2 = 100(e^{\hat{\gamma}u} - e^{\hat{\beta}u})/e^{\hat{\gamma}u}$.*
‡ Cox regression with grouped data is used as denominator for the comparisons.

The agreement among the models is good. Age and GLI have differences of 13·8 and 13·7 per cent, respectively, for the $R1$ comparison of the pooled logistic and group Cox regression coefficients. In terms of $R2$, the relative risk comparison, the differences are only 4·2 and 1·8 per

Table IV. Comparison of regression coefficients: eighteen year follow-up with one non-significant variable and endpoint with low frequency

Population: males 50–59 years old on exam 9, free of lung cancer
Endpoint: lung cancer
Exams 9 to 18, 9 grouping intervals
Variables: Age, ststolic blood pressure (SBP), number of cigarettes per day (CSM)
Basis: 1333 subjects; 66 events; 9360 person exams

Pooled logistic

	$\hat{\gamma}$	SE($\hat{\gamma}$)	z	exp($\hat{\gamma}u$)*
Age	0·0865	0·0185	4·67	2·375
SBP	− 0·0075	0·0065	− 1·15	0·928
CSM	0·0426	0·0077	5·54	1·531

Cox grouped

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0591	0·0229	2·58	1·806
SBP	− 0·0075	0·0066	− 1·14	0·928
CSM	0·0427	0·0076	5·65	1·533

Cox continuous

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0638	0·0218	2·92	1·893
SBP	− 0·0058	0·0062	− 0·94	0·944
CSM	0·0477	0·0070	6·79	1·611

Comparison of coefficients†

	Logistic vs Cox grouped		Logistic vs Cox cont.		Cox grouped vs Cox cont.‡	
	*R*1	*R*2	*R*1	*R*2	*R*1	*R*2
Age	31·7	24·0	26·2	20·3	− 8·0	− 4·8
SBP	0·0	0·0	22·7	− 1·7	22·7	− 1·7
CSM	− 0·2	− 0·1	− 12·0	− 5·2	− 11·7	− 5·1

* $u = 10$ for all variables.
† $R1 = 100(\hat{\gamma} - \hat{\beta})/\hat{\gamma}$, $R2 = 100(e^{\hat{\gamma}u} - e^{\hat{\beta}u})/e^{\hat{\gamma}u}$.*
‡ Cox regression with grouped data is used as denominator for the comparison.

cent, respectively. For the logistic regression and continuous Cox the maximum difference for *R*2 is 9·2 per cent for Age, and it is much smaller for the other variables.

A feature of the three methods is that all agree on the importance of all the variables. The clearly significant variables (Age, SBP, CSM and SCL) are highly significant for all models. The marginally significant variables (MRW and GLI) are marginal for all models. There is also generally good agreement of standard errors of regression coefficients. Except for Age, the three methods produce almost the same estimated standard errors.

Example 2: Lung cancer

The analysis in Table IV examines the coefficients relating risk factors to lung cancer incidence. The number of males free of cancer on exam 9 was 1333. Of these 66 (5·0 per cent) developed lung cancer over the next eighteen years or over the span of the next nine exams. This example illustrates the use of the methods for a situation where there is a very small incidence rate (5 per

Table V. Comparison of regression coefficients: eighteen year follow-up with small sample and non-significant variable

Population: males, 50–69 years old on exam 9, free of CHD
Endpoint: CVD
Exams 9 to 18, 9 grouping intervals
Variables: Age, systolic blood pressure (SBP), number of cigarettes per day (CSM)
Basis: 107 subjects; 50 events, 643 person exams

Pooled logistic

	$\hat{\gamma}$	SE($\hat{\gamma}$)	z	exp($\hat{\gamma}u$)*
Age	0·0308	0·0205	1·50	1·361
SBP	0·0157	0·0065	2·43	1·170
CSM	0·0175	0·0117	1·50	1·191

Cox grouped

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0099	0·0245	0·41	1·104
SBP	0·0163	0·0061	2·65	1·177
CSM	0·0164	0·0110	1·50	1·178

Cox continuous

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0075	0·0240	0·31	1·078
SBP	0·0158	0·0059	2·67	1·171
CSM	0·0186	0·0106	1·75	1·204

Comparison of coefficients†

	Logistic vs Cox grouped		Logistic vs Cox cont.		Cox grouped vs Cox cont.‡	
	$R1$	$R2$	$R1$	$R2$	$R1$	$R2$
Age§	67·9	18·9	75·6	20·8	24·2	2·4
SBP	− 3·8	− 0·6	− 0·6	− 0·1	3·1	0·5
CSM	6·3	1·1	− 6·3	− 1·1	− 13·4	− 2·2

* $u = 10$ for all variables.
† $R1 = 100(\hat{\gamma} - \hat{\beta})/\hat{\gamma}$, $R2 = 100(e^{\hat{\gamma}} - e^{\hat{\beta}u})/e^{\hat{\gamma}u}$.*
‡ Cox regression with grouped data is used as denominator for the comparison.
§ Age is non-significant in analysis; differences are due to small denominators in ratios.

cent spread over eighteen years) and the inclusion of a non-significant variable (systolic blood pressure) and a highly significant variable (cigarette smoking). An examination of the data in Table IV shows that all three models agree in declaration of significance. Age and cigarette smoking are highly significant, and systolic blood pressure has a negative non-significant relation. There is a 20 to 30 per cent disagreement between the pooled logistic and Cox model coefficients for the variable Age. Again, as with the previous example, all three models are in good agreement for declaration of significance and estimation of the magnitude of importance of the risk factors.

Example 3: Cardiovascular disease (CVD) on small sample

The analysis in Table V is designed to strain the agreement among the methods. We selected an 8 per cent random sample of males on exam 9 who were free of cardiovascular disease (CVD). CVD includes CHD as well as stroke, congestive heart failure and intermittent claudication. The

sample consisted of only 107 males, of whom 50 (46·7 per cent) developed CVD over the eighteen year follow-up between exams 9 and 18. The variables Age, systolic blood pressure (SBP) and number of cigarettes smoked per day (CSM) were included in the analysis.

The agreement of the models for the variables SBP and CSM is very good. At most there is a 13 per cent disagreement in coefficients and a 2·2 per cent disagreement of relative risks. The variable Age is non-significant in all three models. There is disagreement between the pooled logistic and both of the Cox models for the regression coefficients of the magnitude 68 to 75 per cent. For the comparison of the relative risks the differences are much smaller, being only of magnitude 19 to 21 per cent. The reason for the differences of regression coefficients is probably due to the problem of Age mentioned above and the lack of significance for Age in all models. Age is an important risk factor for CVD and the lack of significance is related to the small sample size. In large samples it would be significant.

Example 4: Stroke

The analysis in Table VI uses stroke as the endpoint, and again includes a relatively small sample ($n = 334$) and some insignificant variables. In addition there are a small number of events (44 or 13·2 per cent). For this example the agreement among the models is extremely good. All three models agree on the significance of the variable Age. The coefficients for Age are within 11 per cent of each other. The relative risks are within 9 per cent. All three methods lead to marginal significance for systolic blood pressure. The $R1$ and $R2$ comparisons have maximum values of 10 and 1 per cent, respectively. Finally for all three models the cigarette variable CSM is not significant. With more data all three did declare statistical significance for CSM. The insignificance of the present regressions is due to the small sample and low number of events.

DISCUSSION AND CONCLUSION

In the Appendix we outline a proof demonstrating the asymptotic equivalence of the pooled logistic regression and the grouped Cox regression with time dependent variables. Necessary conditions for this equivalence are short intervals for grouping of the outcome events, small probability of an event in the intervals, and equal intercepts of the pooled logistic for each interval. The last assumption is actually not required in the proof. It is necessary only to produce equal probability estimates of an event for the pooled logistic and Cox models. In the previous section we presented some real examples covering a variety of sample sizes, frequency of events, length of follow-up, number of grouping intervals and significance of variables. In these and many other examples examined there was good agreement of the pooled logistic and Cox models, both with grouped or continuous recording of the time of the events. In all cases all three models usually agreed on the declaration of significance of the variables. The magnitude of the importance (e.g. relative risk) was also in good agreement, in fact in closer agreement than the magnitude of the coefficients.

The largest differences occurred for the variable Age when it was not statistically significant in any of the models. The Age variable is the most constant of all variables across exams; the difference from one exam to the next is only the addition of a constant two years, the time between exams. As mentioned above, the time dependent variable Age may not always be appropriate in the time dependent models. In the extreme case where all individuals start at the same age, the time dependent variable Age is perfectly correlated with follow-up time and the effect of age in the Cox models is reflected only in the underlying hazard functions. If the age distribution at time zero had a large non-zero variance, as is often the case in our examples, its use as a time

Table VI. Comparison of regression coefficients: eighteen year follow-up with small sample and non-significant variables

Population: males, 50–69 years old on exam 9, free of stroke
Endpoint: stroke
Exams 9 to 18, 9 grouping intervals
Variables: Age, systolic blood pressure (SBP), number of cigarettes per day (CSM)
Basis: 334 subjects; 44 events; 2254 person exams

Pooled logistic

	$\hat{\gamma}$	SE($\hat{\gamma}$)	z	exp($\hat{\gamma}u$)*
Age	0·0866	0·0229	3·79	2·377
SBP	0·0087	0·0071	1·23	1·091
CSM	0·0100	0·0115	0·87	1·105

Cox grouped

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0775	0·0272	2·85	2·171
SBP	0·0084	0·0071	1·19	1·088
CSM	0·0104	0·0114	0·91	1·110

Cox continuous

	$\hat{\beta}$	SE($\hat{\beta}$)	z	exp($\hat{\beta}u$)*
Age	0·0781	0·0277	2·82	2·184
SBP	0·0092	0·0071	1·30	1·096
CSM	0·0110	0·0113	0·97	1·116

Comparison of coefficients†

	Logistic vs Cox grouped		Logistic vs Cox cont.		Cox vs Cox grouped	Cox cont.‡
	$R1$	$R2$	$R1$	$R2$	$R1$	$R2$
Age	10·5	8·7	9·8	8·1	− 0·8	− 0·6
SBP§	3·4	0·3	− 5·7	− 0·5	− 9·5	− 0·7
CSM§	− 4·0	− 0·4	− 10·0	− 1·0	− 5·8	− 0·5

* $u = 10$ for all variables.
† $R1 = 100(\hat{\gamma} - \hat{\beta})/\hat{\gamma}$, $R2 = 100(e^{\hat{\gamma}u} - e^{\hat{\beta}u})/e^{\hat{\gamma}u}$.*
‡ Cox regression with grouped data is used as denominator for comparisons.
§ Systolic blood pressure and smoking are not significant in any of the models.

dependent variable in the Cox models would be more meaningful and its effect less confounded with follow-up time. The logistic model handles age in a better fashion. Even in the extreme case of zero variability in age, the use of the time dependent variable Age in the logistic model would be meaningful and appropriate. The user of these techniques clearly must consider the nature and appropriateness of variables for analysis. Variables such as Age which may be highly correlated with follow-up time may be better employed in the Cox model as non-time dependent variables or simply recognized as highly correlated with follow-up time. On the other hand, variables such as systolic blood pressure and cigarette smoking, which are not highly correlated with follow-up time and for which changes can have important effects even for short term follow-up, are quite appropriate for Cox time dependent analysis. It is for these latter types of variable that our comparisons of the pooled logistic model and the Cox models were best. Even variables such as Age, if they have large variability such as in some of our examples, show reasonable agreement across models. We believe that the asymptotic agreement of the models appears to be applicable

for a variety of situations such as displayed above. With careful attention to assumptions and the nature of the variables involved, the relation among these models can be of use to the researcher in selecting and interpreting analyses.

APPENDIX

In this appendix we show that under appropriate assumptions, regression coefficients, likelihood function, and a score test for testing $\boldsymbol{\beta} = \mathbf{0}$ based on a grouped Cox regression model with covariates updated at the beginning of each interval are close to those based on a logistic model.

Regression coefficients

Let $h(t)$ be the instantaneous hazard rate at time t and let $S(t)$ denote the probability of being event-free up to time t. It is well known that, in a continuous model, $\log S(t) = -\int_0^t h(u)\mathrm{d}u$. Hence it can easily be shown that for any $\Delta t > 0$, one has

$$\frac{S(t+\Delta t)}{S(t)} = \exp\left[-\int_t^{t+\Delta t} h(u)\mathrm{d}u\right].$$

Let $\mathbf{X}(t)' = (X_1(t), X_2(t), \ldots, X_p(t))$ be the covariates vector associated with an individual at time t. In a Cox regression model with time dependent covariates, the hazard rate at time t is a function of covariate $\mathbf{X}(t)$ such that

$$h(t \mid \mathbf{X}(t)) = h_0(t)\exp[\boldsymbol{\beta}'\mathbf{X}(t)],$$

where $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$, β_j is the regression coefficient associated with the jth risk factor X_j, and $h_0(t)$ is the baseline hazard rate. Assume that observations are grouped into k intervals and let $I_i = [t_{i-1}, t_i)$ denote the ith interval. Then the conditional probability, denoted by $p_i(\mathbf{X}(t))$, that an individual will survive up to time t_i given that he or she has survived to time t_{i-1}, is a function of $\mathbf{X}(t)$, with t varying in the interval I_i, such that

$$p_i(\mathbf{X}(t)) = \frac{S(t_i \mid \mathbf{X}(t_i))}{S(t_{i-1} \mid \mathbf{X}(t_{i-1}))} = \exp\left\{-\int_{t_{i-1}}^{t_i} h_0(u)\exp[\boldsymbol{\beta}'\mathbf{X}(u)]\mathrm{d}u\right\}. \tag{7}$$

To simplify the derivation procedures, denote $H_i(\mathbf{X}(t)) = \int_{t_{i-1}}^{t_i} h_0(u)\exp[\boldsymbol{\beta}'\mathbf{X}(u)]\mathrm{d}u$. Then when the value of $H_i(\mathbf{X}(t))$ is small, for t varying in the interval I_i, the conditional probability $p_i(\mathbf{X}(t))$ can be approximated by its Taylor expansion,

$$\begin{aligned} p_i(\mathbf{X}(t)) &= \exp(-H_i(\mathbf{X}(t))) \\ &= 1 - H_i(\mathbf{X}(t)) + \frac{H_i(\mathbf{X}(t))^2}{2!} - \frac{H_i(\mathbf{X}(t))^3}{3!} + \frac{H_i(\mathbf{X}(t))^4}{4!} - \cdots \\ &= 1 - H_i(\mathbf{X}(t)) + o(H_i(\mathbf{X}(t))) \\ &\approx 1 - H_i(\mathbf{X}(t)). \end{aligned} \tag{8}$$

If the covariate vector $\mathbf{X}$ is measured at the beginning of each interval only, then $\mathbf{X}(t) = \mathbf{X}(t_{i-1})$ for any t in interval, I_i; hence

$$\begin{aligned} H_i(\mathbf{X}(t_{i-1})) &= \int_{t_{i-1}}^{t_i} h_0(u)\exp[\boldsymbol{\beta}'\mathbf{X}(t_{i-1})]\mathrm{d}u \\ &= \exp\left\{\left[\log\int_{t_{i-1}}^{t_i} h_0(u)\mathrm{d}u\right] + \boldsymbol{\beta}'\mathbf{X}(t_{i-1}),\right\}. \end{aligned} \tag{9}$$

Thus, by equations (8), the conditional probability $p_i(\mathbf{X}(t_{i-1}))$ in a grouped proportional hazard model can be approximated by $1 - H_i(\mathbf{X}(t_{i-1}))$. Note that the requirement of small values of $H_i(\mathbf{X}(t_{i-1}))$ can be justified by assuming a small probability of an event occurrence in short intervals.

One can derive similar approximations for a logistic regression model as follows. Let $q_i(\mathbf{X}(t_{i-1})) = 1 - p_i(\mathbf{X}(t_{i-1}))$ denote the conditional probability of observing an event by time t_i, given that the individual is event free at time t_{i-1}. Then the logistic regression model yields

$$\text{logit}\, q_i(\mathbf{X}(t_{i-1})) = \log\left(\frac{q_i(\mathbf{X}(t_{i-1}))}{1 - q_i(\mathbf{X}(t_{i-1}))}\right) = \alpha_i + \boldsymbol{\Gamma}'\mathbf{X}(t_{i-1}), \tag{10}$$

where $\boldsymbol{\Gamma}' = (\gamma_1, \ldots, \gamma_p)$, γ_i is the coefficient associated with risk factor $X_i(t_{i-1})$, and α_i is the effect due to the ith interval. Note that an equivalent form of the logistic regression model is given by

$$p_i(\mathbf{X}(t_{i-1})) = \frac{1}{1 + \exp[\alpha_i + \boldsymbol{\Gamma}'\mathbf{X}(t_{i-1})]}. \tag{11}$$

To simplify notation, denote $G_i(\mathbf{X}(t_{i-1})) = \exp[\alpha_i + \boldsymbol{\Gamma}'\mathbf{X}(t_{i-1})]$. Then for small values of $G_i(\mathbf{X}(t_{i-1}))$ one has the Taylor expansion

$$\begin{aligned} p_i(\mathbf{X}(t_{i-1})) &= \frac{1}{1 + G_i(\mathbf{X}(t_{i-1}))} \\ &= 1 - G_i(\mathbf{X}(t_{i-1})) + G_i(\mathbf{X}(t_{i-1}))^2 - G_i(\mathbf{X}(t_{i-1}))^3 \ldots \\ &= 1 - G_i(\mathbf{X}(t_{i-1})) + o(G_i(\mathbf{X}(t_{i-1}))) \\ &\approx 1 - G_i(\mathbf{X}(t_{i-1})). \end{aligned} \tag{12}$$

Again, the requirement of small values of $G_i(\mathbf{X}(t_{i-1}))$ can be justified by assuming that the probability of an event occurring in a short interval is small. Hence, in a logistic model, $p_i(\mathbf{X}(t_{i-1}))$ can be approximated by $1 - G_i(\mathbf{X}(t_{i-1}))$.

We can now compare the approximation results obtained in equations (8), (9) and (12). Notice that both functions $H_i(\mathbf{X}(t_{i-1}))$ and $G_i(\mathbf{X}(t_{i-1}))$ are log-linear in terms of $\mathbf{X}(t_{i-1})$; therefore the two corresponding little o functions, namely $o(H_i(\mathbf{X}(t_{i-1}))$ and $o(G_i(\mathbf{X}(t_{i-1}))$, are of same order. Hence $1 - H_i(\mathbf{X}(t_{i-1})) \approx 1 - G_i(\mathbf{X}(t_{i-1}))$. Comparing the component coefficients of the covariate vector $\mathbf{X}$ in $H_i(\mathbf{X}(t_{i-1}))$ and $G_i(\mathbf{X}(t_{i-1}))$, one may conclude that γ_j and β_j are approximately equal for each j, and therefore $\boldsymbol{\Gamma} \approx \boldsymbol{\beta}$. Similarly, $\alpha_i \approx \log \int_{t_{i-1}}^{t_i} h_0(u)\mathrm{d}u$.

In practice, regression coefficients of both the logistic model and the grouped Cox regression model are computed iteratively using the Newton–Raphson procedure, and their corresponding covariance matrices are computed as a by-product.

Likelihood functions

Assume that censoring takes place at the end of each interval only, and assume that there are no tied observations. Let R_i denote the risk set at time t_{i-1}, D_i denote the set of individuals in whom events are observed in interval $I_i = [t_{i-1}, t_i)$, and C_i the set of individuals censored in I_i. Applying approximation results derived in the previous section, it can be shown that when interval lengths approach zero and the probability of an event occurring in a short interval is small, the likelihood function for a grouped Cox regression model, with time dependent covariates updated at the beginning of each interval, is approximately equal to that of a logistic regression model.

Using the notation given above, the likelihood function of a grouped proportional hazard model is given by

$$\prod_{i=1}^{k}\left\{\prod_{l\in D_i}(1-\exp(-H_i(\mathbf{X}_l(t_{i-1}))))\prod_{m\in R_i-D_i}\exp(-H_i(\mathbf{X}_m(t_{i-1})))\right\} \quad \text{by (7)}$$

$$\approx \prod_{i=1}^{k}\left\{\prod_{l\in D_i}H_i(\mathbf{X}_l(t_{i-1}))\prod_{m\in R_i-D_i}(1-H_i(\mathbf{X}_m(t_{i-1})))\right\} \quad \text{by (8) and (9)}$$

$$\approx \prod_{i=1}^{k}\left\{\prod_{l\in D_i}G_i(\mathbf{X}_l(t_{i-1}))\prod_{m\in R_i-D_i}(1-G_i(\mathbf{X}_m(t_{i-1})))\right\} \quad \text{as earlier}$$

$$= \prod_{i=1}^{k}\left\{\prod_{l\in D_i}\frac{G_i(\mathbf{X}_l(t_{i-1}))}{1-G_i(\mathbf{X}_l(t_{i-1}))}\prod_{m\in R_i}[1-G_i(\mathbf{X}_m(t_{i-1}))]\right\}$$

$$\approx \prod_{i=1}^{k}\left\{\prod_{l\in D_i}G_i(\mathbf{X}_l(t_{i-1}))\prod_{m\in R_i}\frac{1}{1+G_i(\mathbf{X}_m(t_{i-1}))}\right\} \quad \text{by (12)}$$

$$= \prod_{i=1}^{k}\left\{\prod_{l\in D_i}\exp[\alpha_i+\boldsymbol{\beta}'\mathbf{X}_l(t_{i-1})]\prod_{m\in R_i}\frac{1}{1+\exp\{\alpha_i+\boldsymbol{\beta}'\mathbf{X}_m(t_{i-1})\}}\right\}, \tag{13}$$

which is the likelihood function for a logistic regression model.

Note that if we further assume that there is no interval effect (i.e. $\alpha_i = \alpha$ for all i), then the last equation in (13) can be considered as the likelihood function for a pooled cross-sectional logistic model when short term prediction is of interest.

For cases with tied observations or with censoring taking place other than at the end of each interval, modifications can be made as by Thompson.[10]

Score test statistics for testing $\boldsymbol{\beta} = \mathbf{0}$

By equations (7) and (9), we can write

$$p_i(\mathbf{X}(t_{i-1})) = \left\{\exp\left(-\int_{t_{i-1}}^{t_i}h_0(u)\mathrm{d}u\right)\right\}^{\exp[\boldsymbol{\beta}'\mathbf{X}(t_{i-1})]}$$

$$= p_i(0)^{\exp[\boldsymbol{\beta}'\mathbf{X}(t_{i-1})]}, \tag{14}$$

where $p_i(\mathbf{0}) = \exp\left(-\int_{t_{i-1}}^{t_i}h_0(u)\mathrm{d}u\right)$. Values of $p_i(\mathbf{0})$ lie between 0 and 1; therefore a re-parameterization is considered to remove this range restriction. Define new parameters $\theta_i = \log(-\log p_i(\mathbf{0}))$, $i = 1, \ldots, k$. Then, from equations (13) and (14), the log-likelihood function of a grouped Cox regression model with time dependent covariates can be written as

$$\log L(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \sum_{i=1}^{k}\left\{\sum_{l\in D_i}\log[\exp\{\exp(\theta_i+\boldsymbol{\beta}'\mathbf{X}_l(t_{i-1}))\}-1]-\sum_{l\in R_i}\exp(\theta_i+\boldsymbol{\beta}'\mathbf{X}_l(t_{i-1}))\right\}. \tag{15}$$

The above log-likelihood function is of the same form as that of a grouped Cox regression model with fixed covariates. Hence one can follow Prentice and Gloeckler's[14] procedure to derive a partial score test. It can be shown that the elements involved in computing the partial score test do not differ greatly from those for the logistic regression model when interval lengths are short and the event considered is a rare event. The derivation is straightforward but tedious and hence is omitted.

ACKNOWLEDGEMENTS

The authors wish to thank the referees for their extensive comments which have greatly improved the paper. Research funded in part by National Heart, Lung and Blood Institute grant 1-RO1-HL-40423-02.

REFERENCES

1. Dawber, T. R. *The Framingham Study: the Epidemiology of Atherosclerotic Disease*, Harvard University Press, Cambridge, Massachusetts, 1980.
2. D'Agostino, R. B. and Kannel, W. B. 'Epidemiological background and design: the Framingham Study', in *Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions*, American Statistical Association, Alexandria, Virginia, 1989, pp. 707–718.
3. Shurtleff, D. 'Section 30: some characteristics related to the incidence of cardiovascular disease and death: Framingham Study 18-year follow-up', in Kannel, W. B. and Gordon, T. (eds), *The Framingham Study: an Epidemiological Investigation of Cardiovascular Disease*, U.S. Government Printing Office, DHEW publication (NIH)74-599, 1974.
4. Kahn, H. A. and Dawber, T. R. 'The development of coronary heart disease in relation to sequential biennial measures of cholesterol in the Framingham Study', *Journal of Chronic Diseases*, **19**, 611–620 (1966).
5. Schatzkin, A., Cupples, L. A., Heeren, T., Morelork, S. and Kannel, W. B. 'Sudden death in the Framingham Heart Study: differences in incidence and risk factors by sex and coronary disease status', *American Journal of Epidemiology*, **120**, 888–899 (1984).
6. Cupples, L. A., D'Agostino, R. B., Anderson, K. and Kannel, W. B. 'Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study', *Statistics in Medicine*, **7**, 205–218 (1988).
7. Wu, M. and Ware, J. H. 'On the use of repeated measurements in regression analysis with dichotomous response', *Biometrics*, **35**, 513–521 (1979).
8. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
9. Cox, D. R. and Oakes, D. *Analysis of Survival Data*, Chapman and Hall, New York, 1984.
10. Thompson, Jr. W. A. 'On the treatment of grouped observations in life studies', *Biometrics*, **33**, 463–470 (1977).
11. Abbott, R. D. 'Logistic regression in survival analysis', *American Journal of Epidemiology*, **121**, 465–471 (1985).
12. Ingram, D. D. and Kleinman, J. C. 'Empirical comparisons of proportional hazards and logistic regression models', *Statistics in Medicine*, **8**, 525–538 (1989).
13. Korn, E. L. and Whittemore, A. S. 'Methods for analyzing panel studies of acute health effects of air pollution', *Biometrics*, **35**, 795–802 (1979).
14. Prentice, R. L. and Gloeckler, L. A. 'Regressions analysis of grouped survival data with application to breast cancer data', *Biometrics*, **34**, 57–67 (1978).