J Clin Epidemiol Vol. 44, No. 9, pp. 915–923, 1991
Printed in Great Britain.

0895-4356/91 $3.00 + 0.00

LATENCY AND TIME-DEPENDENT EXPOSURE IN A CASE–CONTROL STUDY

LAWRENCE H. MOULTON[1] and MONIQUE G. LÊ[2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A. and
[2]I.N.S.E.R.M. Unit 287, l'Institut Gustave-Roussy, 94805 Villejuif Cedex, France

(Received in revised form 11 January 1991)

Abstract—Detailed historical data are elicited often from subjects in retrospective studies, yielding time-dependent measures of exposures. Investigation of a hypothesized period of latency can be made by examining disease/exposure relationships in multiple time windows, either along the age or time-before diagnosis axes. We suggest splitting the data into many time intervals and separately fitting regression models to the available data in each interval. Covariances between estimated coefficients from different intervals are empirically estimated, and used for assessing variability of specified functions of the time-specific coefficients. Alternative methods of interval formation and their consequences are discussed. We apply these methods to a French case–control study of oral contraceptive use and cervical cancer incidence, and compare the results to those of standard analyses.

Case–control studies Latent period Longitudinal data analysis Logistic regression Cervical cancer Oral contraceptives

INTRODUCTION

A difficult question in any epidemiologic study is how best to characterize exposure. Often measures such as cumulative exposure or peak exposure during a specified time unit are used. For example, in studies on the effects of smoking, total years smoked, total pack-years smoked (both measures of cumulative exposure), or maximum number of packs smoked in any year may be used in modeling the relationship between smoking and a given disease outcome. However, these summary measures sacrifice information on the patterns of exposures over time that may have etiologic importance.

Going a step further than the uniform weights that yield simple cumulative exposure, many differential weighting schemes have been proposed to accommodate certain types of exposure/disease relationships [1, 2]. These methods presume a strong knowledge of the biological mechanisms involved in causation of the disease in question, and may impose an inappropriate smoothing of the data when this knowledge is not strictly correct. A more exploratory approach has been taken by some workers [3, 4] in which odds ratios (ORs) are calculated for each of many potential exposure periods. To investigate latency and indicate critical time intervals, the ORs are only plotted and inspected. Rothman [5] similarly has proposed repeatedly performing analyses, each time employing as an explanatory variable the exposure that occurred in a separate time interval. Further inference, however, has been hampered by the acknowledged difficulty of accounting for the correlation between the ORs or the interval-specific coefficients, which are based on the same subjects across time.

In this paper, we propose a semi-parametric method of handling this correlation, thereby

permitting statistical inference for relevant functions of time-specific measures of relationship. For illustrative purposes, the approach is applied to a French case–control study [6, 7] carried out to investigate the relationship between cervical cancer incidence and oral contraceptive (OC) use. Cervical cancer is a particularly difficult disease to study with respect to OCs. Invasive cervical cancer results from a series of changes in the cervical epithelium from normal epithelial structure to various grades of squamous dysplasia to carcinoma *in situ*, and thence on to invasive cervical cancer. Oral contraceptives could act at any one stage to enhance progression to the next stage. Thus, arbitrary summarization of cumulative OC use is too restrictive an approach to understanding the possible role of OCs in the malignant process, or to suggest hypotheses about underlying mechanisms.

The French data set

Between 1982 and 1985, 160 cases and 320 controls less than 45 years of age were recruited from seven French medical centers and interviewed. Cases were interviewed within 3 months of histological verification of invasive epidermoid cancer or carcinoma *in situ* of the cervix uteri. Hospital-based controls without malignant disease were: 45% patients with benign thyroid tumor (no hormonal abnomalies), 17% gynecological disorders (no cervical dysplasia), 15% other diseases, and 23% healthy women (regular outpatient screening visits, etc.). No cases or controls were reported to have refused the interview. The controls were matched 2:1 to the cases on age at time of interview (within 2 years), center, and year of interview (within 2 years). The latter two matching criteria would not be expected to be influential, since all subjects were recruited in a 3 year time period, with 88% of them from the Gustave-Roussy Institute. We have ignored accounting for matched stratum membership, thereby incurring at most a slight decrease in observed associations with case–control status [8]. If the data had not been age-matched, we would have to include age as a covariate. Not including it here actually aids interpretation of our left- and right-adjusted analyses, as they will focus on the effect of OC use in a given interval relative to the event time.

For each subject, complete reproductive life histories were determined, spanning the time from first sexual intercourse (FSI) to 3 months before the interview. Yearly data both on full-term pregnancies and fetal losses were recorded, together with detailed information on the contraceptive methods used during each intervening interval. Four women with no history of sexual intercourse (1 case and 3 controls) were excluded from the analyses; results thus will be generalizable at most to an ever-sexually active population.

All of the analyses appearing here include five binary-coded risk factors: FSI before age 18, ever divorced, at least one abortion, greater than 90 g of alcohol consumed per week, and more than 10 cigarettes smoked per day (current status for these latter two). A time-dependent covariate, the cumulative number of live births prior to the start of any interval, also was incorporated into the models.

ANALYSIS VIA TIME-SPECIFIC REGRESSION MODELS

Combining coefficients across time

For each subject, we divide the period from beginning of exposure opportunity until time of interview into consecutive time intervals indexed by $t = 1, \ldots, T$. While the intervals need not be of equal length, interpretation may be simpler if they are.

It might appear reasonable to fit one logistic regression model to all of the data, including T covariates in the model to investigate the relative relationships of exposure in each interval to case–control status. However, there are three drawbacks to this straightforward approach. The first is that one may have to deal with problems of multicollinearity, due to possibly highly correlated exposures across time, or instability from the estimation of too many parameters. This latter consideration will be more important when there are several time dependent variables, each measured at numerous intervals. The second is that the specific effect of a variable at a given time point cannot be ascertained, as its coefficient is adjusted for exposure during all other time points, as well as other covariates' values of other times. Third, there will be much missing data on exposure due to the unequal lengths of the subjects' histories, leading to greatly reduced efficiency. One could presume exposure to be zero outside each person's observed time frame, but this will lead to problems of interpretation regarding zero exposure before birth or, say, 5 years in the future for a given woman.

Instead, we fit a logistic regression model separately to the data corresponding to each interval. The parameter estimates from the intervals then may be combined to yield summary estimates or significance tests, or examined for temporal trends in relationships. Of central importance is the use of the estimated degree of correlation between coefficients from different intervals, so that interpretive difficulties of multiple comparisons or falsely low standard errors are avoided.

In our example of cervical cancer and oral contraceptive use, let y_i denote case–control status for the ith woman (case: $y_i = 1$; control: $y_i = 0$; $i = 1, \ldots, N$). This particular data set contained the predominant method of birth control for each year between FSI and the interview. However, to decrease variability of regression coefficients due to year-to-year variation in OC use, and to ease interpretation of the many sets of regression coefficients, we grouped the data into 5-year intervals.

We consider the case of unconditional analysis; the methods may be extended to conditional (on stratum membership) analysis, but this is not necessary for our example (although it might be slightly more efficient). For the tth time interval, an $N \times p$ matrix X_t carries the covariate values, with vectors x_{it} for each individual. If we fit the model:

$$\text{logit}\{\Pr(y_i = 1 \mid x_{it})\} \equiv \text{logit}(p_{it}) = x_{it}'\beta_t \quad (1)$$

separately for each interval t via maximum likelihood, we obtain T coefficient vector estimates $\hat{\beta}_t$. These estimates in general will be correlated due to within-subject correlation. Thus, to make inferences on functions of the interval-specific coefficients β_t, such correlation must be accounted for. Functions of interest may include, for the jth covariate,

$$\sum_t \beta_{jt}/\mathrm{T}, \quad \max_t(\beta_{jt}),$$

and polynomial coefficient estimates to describe time trends for the β_{jt}. Consistent variance estimates for these quantities may be calculated by using the following empirical covariance estimator for any pair of intervals (r, s):

$$\widehat{\text{Cov}}(\hat{\beta}_r, \hat{\beta}_s) = (X_r' V_r X_r)^{-1} X_r' V_r^{1/2} \times \text{diag}(r_{ir} r_{is}) V_s^{1/2} X_s (X_s' V_s X_s)^{-1}, \quad (2)$$

where

$$V_r = \text{diag}\{\hat{p}_{ir}(1 - \hat{p}_{ir})\}$$

and r_{ir} is the Pearson residual

$$(y_i - \hat{p}_{ir})/\{\hat{p}_{ir}(1 - \hat{p}_{ir})\}^{1/2}.$$

The estimator $\widehat{\text{Cov}}(\hat{\beta}_r, \hat{\beta}_s)$ is an extension of a robust variance estimator [9, 10]; note that the degree of covariation for each subject is estimated via the residual products. Its derivation is outlined in the Appendix.

In a situation with little prior information about the mechanism of action or latency period, it may be of interest to perform a global test of whether any relationship exists between the disease and exposure during any time interval. This can be accomplished by employing the procedure of Stram *et al.* [11] as follows: for the jth covariate of interest, we calculate

$$\Pr\left\{\max_t(|\hat{\beta}_{jt}|/\gamma_{jtt}^{1/2}) > \text{largest such observed } z\text{-score}\right\} \quad (3)$$

where γ_{jtt} is the jth diagonal element of the $\mathrm{T} \times \mathrm{T}$ covariance matrix

$$\Gamma_j \equiv \{\widehat{\text{Cov}}(\hat{\beta}_{jr}, \hat{\beta}_{js})\}.$$

This involves integrating over a multivariate normal distribution with mean 0 and covariance matrix Γ_j, corresponding to the null hypothesis of no relationship, but accounting for a given degree of correlation across time intervals.

The statistical significance of a relationship whose strength increases or decreases over time may be assessed by investigating a linear trend in the β_{jt}. For the case of equal interval lengths, the slope of such a trend may be estimated by making the weighted least squares calculation:

$$(C' \Gamma_j^{-1} C)^{-1} C' \Gamma_j^{-1} b_j, \quad (4)$$

where

$$b_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jT})'$$

and

$$C = \begin{bmatrix} 11 \ldots 1 \\ 12 \ldots T \end{bmatrix}',$$

with estimated asymptotic covariance matrix $(C' \Gamma_j^{-1} C)^{-1}$. Similarly, a quadratic term could be added in order to assess the variability of latency for an observed peak in the ORs. Because this approach is not fully parametric, likelihood ratio tests are not possible, and Wald tests (z-scores) must be employed. Note that it

is possible for Γ_j to be non-positive definite; however, for sufficiently large samples this will not often be the case, as it is consistent for the true covariance of the $\hat{\beta}_{jt}$.

Interval construction and missing data

A difficult question is how to align the individual subjects' histories in order to form the time intervals. In our example, the subjects' ages at interview range from 23 to 45, and the time from first sexual intercourse to interview ranges from 3 to 30 years. We consider two kinds of alignment: (1) right-adjustment of the histories by date of interview, and (2) left-adjustment at an age prior to the youngest FSI, say 15 (see Fig. 1). A third method could be left-adjustment to each woman's FSI; since 62% of the women experienced FSI before age 20, this would yield results similar to the second method, but its interpretation would be less useful.

A logistic regression fit to the data from a time interval constructed using the right-adjusted alignment will indicate which variables discriminate between cases and controls at the given number of years before onset of disease. This method has been used by occupational epidemiologists to investigate latency between a given exposure and a disease, and has numerous advantages that are presented elsewhere [3, 4, 12, 13]. If the relationship of OC use to cervical cancer is that of a tumor promotor, this approach is best suited to detecting it, as it emphasizes the years just prior to diagnosis. Note that there will be no data for many of the younger individuals in the early (left side) intervals; the variance of coefficients estimated there will be correspondingly larger. Some of the early intervals may have to be grouped in order to have sufficient data for estimation, although interpretability would be rendered more difficult.

When histories are left-adjusted to a specified age, all subjects have the same age in any resulting time interval. Regression coefficients will enable discrimination of eventual case–control status based on use of OC at any given age. Now, there will be data lacking on the right side of the histories. Thus, more weight will be given to the earlier time intervals, which is what would be best if OC were acting as a tumor initiator. It focuses on age at exposure, rather than time before the appearance of disease.

Data may be missing on an individual for an interval either as a result of a short history or through failure of the investigator to obtain valid information. With the variables used in our example, only the former occurred. However, both types of missing data are handled easily, provided that for the latter case the data are missing completely at random. For any pair of time intervals, only the observations that have complete data for both intervals contribute to estimation of the covariance.

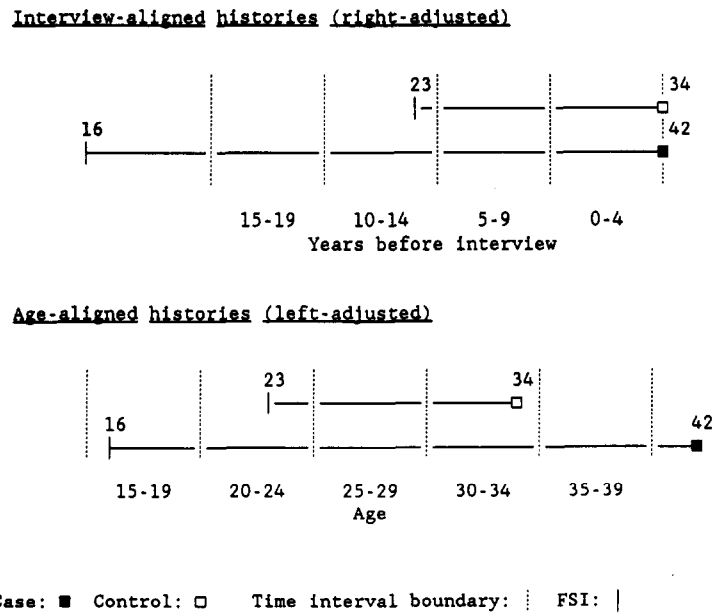

Fig. 1. Alternative time interval constructions for cross-sectional regression fitting. Illustrated for women aged 34–42 at time of interview whose first sexual intercourse occurred at ages 23 and 16, respectively.

Rewriting the middle of (2) to make explicit the subjects' contributions, we use:

$$\widehat{\mathrm{Cov}}(\hat{\beta}_r, \hat{\beta}_s) = f_{rs}(X_r' V_r X_r)^{-1} \times \left[\sum x_{ir} x_{is}'(y_i - \hat{p}_{ir})(y_i - \hat{p}_{is})\right](X_s' V_s X_s)^{-1} \quad (5)$$

where the sum is over all N_{rs} complete pairs of observations, and $f_{rs} = N_r N_s / N_{rs}^2$ scales the interval-specific variances accordingly. The rest of the analysis proceeds as in the preceding section.

APPLICATION TO THE FRENCH DATA SET

Analyses of right-adjusted data

We used only four 5-year intervals because of the sparseness of the data in the fifth interval (20–24 years before interview); none of the women over the age of 40 used ICs before age 20. The interval-specific estimated coefficients for the number of years of OC use within each interval (coded 0, 1, 2, 3, 4, 5) are shown in Fig. 2 with their approximate 95% confidence intervals. There is an observed trend in the coefficients, with corresponding ORs going from 0.87 for the 15–19 years interval to 1.20 for the 0–4 years-before-interview interval. For a woman taking OCs for all 5 previous years, this gives an approximate relative risk of $1.2^5 = 2.4$.

We can assess the overall significance of the OC/cancer relationship by calculating the quantities (3) and (4) and assessing their variability with covariances estimated as in equation (5). First, to motivate the need for accounting for the covariances of coefficients across intervals,

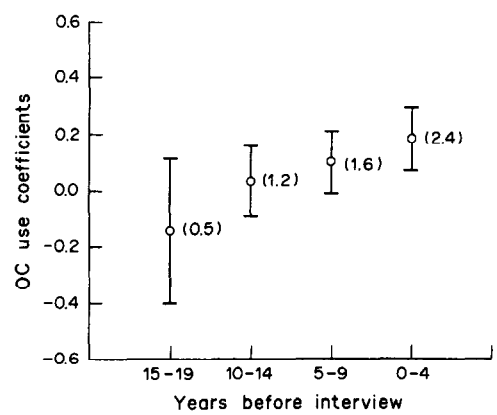

Fig. 2. Confidence intervals for right-adjusted interval-specific oral contraceptive use coefficients. Coefficient values are the log ORs for 1 year of OC use compared to zero; in parentheses are the corresponding odds ratios for 5 years of use.

Table 1. Estimated correlations between coefficients for the OC use variable in each 5-year interval before interview (right-adjusted)

Years before interview	Years before interview			
	15–19	10–14	5–9	0–4
15–19	1			
10–14	0.60	1		
5–9	0.24	0.49	1	
0–4	0.07	0.14	0.51	1

we note there is a high degree of correlation between coefficients from adjacent intervals, which rapidly diminishes with increasing distance in time (Table 1). Our test of the global null hypothesis of no association at any interval is performed by calculating

$$\Pr\left\{\max_t(|\hat{\beta}_{\mathrm{OC},t}|/\text{s.e.}) > 3.12\right\}$$

under the null hypothesis that sampling is from a multivariate normal distribution with zero mean and the correlation matrix of Table 1, where the z-score from the (0–4) interval, $0.178/0.057 = 3.12$, was the largest observed. The resulting p value is 0.007, calculated using algorithm AS 195 [14]. For the right-adjusted coefficients, the weighted estimates of the intercept and slope are -0.154 and 0.086, respectively, with a 95% CI of (0.010, 0.162) for the slope. Thus, the observed upward trend is sufficiently stably estimated to imply that the more recent the use of OCs, the greater the associated OR.

Analyses of left-adjusted data

Left-adjusting the histories starting at age 15, we were able to fit the logistic models to the five age intervals 15–19, . . . , 35–39; again, estimation for an additional interval was attempted, but there were not enough 40+ women for estimation of all the parameters. In Fig. 3 we see the interval-specific coefficients for OC use follow an upward trend similar to that of the right-adjusted intervals. As before, the coefficient for the age 15–19 interval is highly variable, probably because of the low levels of OC use at those ages. The estimated correlation matrix, shown in Table 2, exhibits the same pattern as seen for the right-adjusted calculation. However, even with an additional interval, the largest z-score is only $0.144 - 0.068 = 2.11$, giving

$$\Pr\left\{\max_t(|\hat{\beta}_{\mathrm{OC},t}|/\text{s.e.}) > 2.11\right\} = 0.149.$$

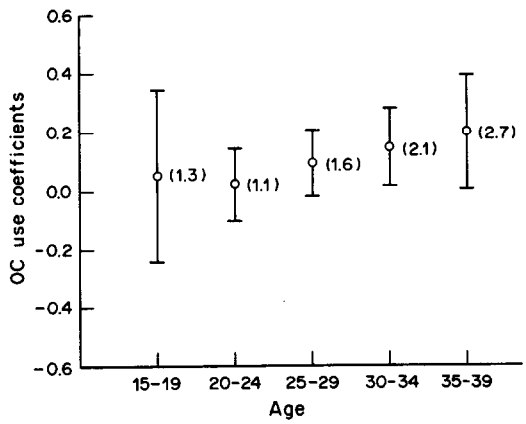

Fig. 3. Confidence intervals for left-adjusted interval-specific oral contraceptive use coefficients. Coefficient values are the log ORs for 1 year of OC use compared to zero; in parentheses are the corresponding odds ratios for 5 years of use.

The z-score for the estimated slope of the five coefficients is $0.052/0.037 = 1.4$, much less significant than for the right-adjusted case.

Cumulative effect of OC use

It could be argued that the above $\beta_{OC,t}$ are increasing due to the cumulative effect of OC use over the years. We investigated this possibility using the right-adjusted interval-specific regression approach. For each interval, instead of using the number of pill-years of use for that interval, we entered the total number of pill-years up to and including the interval. The estimates for this cumulative exposure variable are shown in Table 3. There is still an upward trend as observed with the interval-specific use variables; however, the estimated effects are smaller and less significant for the (5–9) and (0–4) intervals, as the use in those intervals is diluted by the previous intervals. To investigate the separate effects of these two variables adjusting for the other, we fitted interval-specific regressions simultaneously estimating the two coefficients. For the (15–19) interval, the two variables are identical, so only the remaining three regressions were calculated, with results given in Table 4. While the effect of the cumulative use variable almost completely disappeared in the presence of the within-interval use variable, the latter's point estimates and statistical significance increased substantially.

Table 2. Estimated correlations between coefficients for the OC use variable in each 5-year age interval (left-adjusted)

Age interval	Age interval				
	15–19	20–24	25–29	30–34	35–39
15–19	1				
20–24	0.41	1			
25–29	−0.05	0.44	1		
30–34	−0.01	0.08	0.47	1	
35–39	0.04	−0.02	0.15	0.65	1

Table 3. Results for the time-dependent cumulative years of OC use variable

	Years before interview			
	15–19	10–14	5–9	0–4
Coefficient estimate	−0.145	−0.002	0.032	0.056
z-score	−1.10	−0.04	1.00	2.29

Comparison with exposure summarization methods

To give a comparison with the results that would be obtained via more conventional summary exposure-based analyses, we also performed analyses similar to those of Thomas [1], trying out total cumulative exposure, age-at-exposure and interval-from-exposure-to-risk weighting schemes. In these analyses, one logistic regression model is fit using a variable that summarizes each person's exposure across time into a single value.

The maximum time from FSI to interview was 30 years, so six 5-year right-adjusted intervals cover the periods of potential OC use of all the women. The constructed summary exposure variables used are described in Table 5.

The first of each set of weights corresponds to the interval 25–30 years before interview, the last to the 0–5 years-before-interview interval. The measures are calculated as the weighted sum of the numbers of pill-years of use within each 5-year interval. The latter two sums were transformed to "direct effect" [1] variables,

Table 4. Results for the simultaneous inclusion of the cumulative years of OC use and interval-specific use variables

Variable		Years before interview		
		10–14	5–9	0–4
Cumulative OC use	Coefficient estimate	−0.188	−0.036	0.003
	z-score	−1.32	−0.66	0.10
Interval-specific use	Coefficient estimate	0.172	0.241	0.328
	z-score	1.55	2.67	3.69

Table 5. Description of summary exposure variables

Name	Weights	Meaning
UW	(1, 1, 1, 1, 1, 1)	Uniform; usual cumulative exposure
IW	(0, 1, 2, 3, 4, 5)	Increasing; "age-at-exposure"
DW	(5, 4, 3, 2, 1, 0)	Decreasing; "interval from exposure to risk"

IWD = IW/UW and DWD = DW/UW, to examine the effects of use patterns as distinct from the amount used. Table 6 shows the results of including these variables, both with and without UW, in the model. They are in accordance with those obtained from the approach of estimating and combining interval-specific regression coefficients, with exposures (OC use) in the years just prior to interview being far more implicated than either early years of use or the cumulative amount used. As above, once the recent use is accounted for, adding cumulative use explains virtually no more of the deviance of the model.

Comparison with a multiple interval, single regression approach

As what may be seen as an extension of a proposal of Rothman [5], in one logistic regression we simultaneously included variables for the numbers of years of OC use within each 5-year interval. Thus, using right-adjustment of the data, there were 4 terms included for OC use; also used were 4 terms for number of children born prior to each interval, the other variable on which we had time-dependent information. The sample size was reduced from 476 to 303, the youngest women being deleted because of lack of information on OC use before FSI. In this data set, assuming the OC exposure to be zero in the intervals before FSI (keeping the sample size at 476) yielded virtually the same results. Table 7 shows the set of coefficients from our interval-specific regressions (already displayed in Fig. 1) as well as those from the multiple exposure interval terms in the single regression. Neither the consistent trend nor the statistical significance of the former analysis is retained in the latter. The disparity is greater when the single regression method is applied to the left-aligned data, with the largest z-score being only 1.1. In the situation, the missing data appear at the right end, where more OC use and the strongest relationships were found.

Table 6. Results of fitting five models with summary OC use variables

Wald test p-values			
Summary variables in model			
UW	IWD	DWD	−2 log-likelihood
0.024			528.07
	0.001		521.70
		0.987	533.19
0.797	0.012		521.64
0.015		0.373	527.26

DISCUSSION

Although the random quantities in a retrospective design are the exposure variables, case–control status may be treated as the response variable in a prospective model framework [15, 16]. With retrospective longitudinal data, this response is fixed across time. Yet the detailed exposure histories may induce variable degrees of serial correlation across time, as is the case with prospective data.

There have been many recent advances in the analysis of repeated measures of binary outcome variables with time-dependent covariates [9, 11, 17, 18]. The approaches of Moulton and Zeger [9] and Stram *et al.* [11] are most relevant here, as they both permit modeling of the overall regression relationship as well as focus on strength-of-association at any

Table 7. Comparison of OC use coefficients from right-adjusted interval specific regressions and a single regression with variables for each interval

	Years before interview			
	15–19	10–14	5–9	0–4
Number of observations	303	435	471	476
Interval specific regressions				
Estimate	−0.145	0.031	0.098	0.178
z-score	−1.10	0.48	1.72	3.12
Single regression				
Estimate	−0.232	0.095	−0.059	0.215
z-score	−1.49	0.98	−0.58	2.32

specific point in time. This latter aspect is in the same spirit of the time-specific analyses of ORs [3], serially additive expected doses [12], etiologic fractions [4], and retrospective excess exposure fractions [13] investigated by other workers. However, the advantages of our proposed methodology are the ability to handle across-time correlation of exposure measures and the ease of adjusting for other, possibly time-dependent, covariates.

With our semi-parametric approach, we have explored the relationship OC use and risk of cervical cancer in a particular data set. Slightly stronger positive associations were observed for the right-adjusted data, which indicated that most recent use of OCs discriminated best between cases and controls. In the left-adjusted data, the interval-specific coefficients for OC use followed a similar upward trend, with the exception of the age 15–19 interval. The variance of that interval's coefficient was high; it is also possible that OC use in the teenage years is serving as a proxy for having more sexual partners, a known risk factor for cervical cancer [19]. In our data set, this factor was not available, and we could not directly control for it. However, it may partially be accounted for by our having included age at FSI, which has been seen to be correlated with number of partners [20].

Our results are consistent with the hypothesis that the role of OC use in cervical cancer incidence is either that of a tumor promotor, which typically has short latent periods between exposure and appearance of disease, or that of an accelerator of changes in the cervical epithelium from severe squamous dysplasia to carcinoma *in situ*, and from carcinoma *in situ* to invasive carcinoma. Of course, this interpretation depends to some degree on assuming that the quality of the retrospectively-collected data does not deteriorate too badly as one recedes in time. Other studies of OCs and cervical cancer have reported mixed results, possibly due to the "negative" studies' lack of total OC pill-years of use and low power to detect an OC/cancer relationship [21]. In the "positive studies," the relationship between OC use and risk of cervical cancer was often interpreted as a non-causal relationship since this cancer is not a hormono-dependent cancer. Many authors suggested that the observed association reflected only some uncontrolled confounders.

In conclusion, we have shown how case–control studies with exposure information available for many points in time may be analyzed in a manner similar to what can be done for a prospective longitudinal study. In the retrospective setting, greater care must be given to the interpretation of coefficients, which will depend on how the life histories are aligned and on the mixture of fixed and time-dependent covariates. In our analyses, for example, the coefficient for OC use calculated for the 15–19 age interval is adjusted for, among other variables, eventual smoking status at the time of investigation. Ideally, it would be best to have complete time-dependent data for each relevant variable, a situation that is readily and efficiently handled via our approach.

The interval-specific regression approach, while offering substantial exploratory possibilities, still enables confirmatory analysis, in that the variability of quantities of interest may be assessed. Thus it has many of the attractive features of the investigative approaches favored by some epidemiologists [3, 4, 5, 12, 13], as well as the capabilities of the weighted summary exposure approach to handle numerous potential confounding variables and choose between alternative models on a statistical basis. It renders feasible the statistical assessment of a purported or observed latent period. In addition, use of time-dependent covariates can account for differential exposure patterns by period and cohort; however, residual cohort effects may continue to be confounded with age [22].

The approaches presented above also could be extended to the analyses of matched data sets, which may be necessary when the matching variables have strong relationships with both exposure and disease. Another useful modification would be incorporation of a method for restricting or smoothing the estimated covariance matrices Γ_j to diminish the possibility of a non-positive definite matrix. More complex multi-stage models perhaps could be handled by our interval-specific approach. However, such more detailed investigations are better performed on larger sets of data, and with more knowledge about the biological nature of the disease process, as opposed to earlier, more exploratory studies.

Acknowledgements—This research was made possible by a grant from the National Cancer Institute, USA, and l'Institut National de la Santé et de la Recherche Médicale, France. The authors thank the reviewers for their useful comments.

REFERENCES

1. Thomas DC. Statistical methods for analyzing effects of temporal patterns of exposure on cancer risks. **Scand J Work Environ Health** 1983; 9: 353–366.
2. Thomas DC. Models for exposure–time–response relationships with applications to cancer epidemiology. **Ann Rev Public Health** 1988; 9: 451–482.
3. Smith AH, Checkoway H, Goldsmith DF, Wolf PH, Tyroler HA. An analytic procedure for investigation of cancer latency in matched case control studies illustrated with occupational data. **Am J Epidemiol** 1978; 108: 226.
4. Goldsmith DF. Calculating cancer latency using data from a nested case–control study of prostatic cancer. **J Chron Dis** 1987; 40 (Suppl. 2): 119S–123S.
5. Rothman KJ. Induction and latent periods. **Am J Epidemiol** 1981; 114: 253–259.
6. Lê MG, Bachelot A, Doyon F, Kramar A. Etude de la relation entre contraception orale et cancer du sein et du col utérin: Résultats préliminaires d'une enquête française. **Contraception–Fertilité–Sexualité** 1985; 13: 553–558.
7. Lê MG, Bachelot A, Doyon F, Kramar A, Hill C. Oral contraceptive use and breast or cervical cancer: preliminary results of a French case–control study. In: Wolff J-P, Scott JS, Eds. **Hormones and Sexual Factors in Human Cancer Aetiology.** Amsterdam: Elsevier; 1984; 139–147.
8. Breslow NE, Day NE. **Statistical Methods in Cancer Research,** Vol. I. Lyon: International Agency for Research on Cancer: 1980.
9. Moulton LH, Zeger SL. Analyzing repeated measures on generalized linear models via the bootstrap. **Biometrics** 1989; 45: 381–394.
10. Royall RM. Model robust confidence intervals using maximum likelihood estimators. **Int Stat Rev** 1986; 54: 221–226.
11. Stram DO, Wei LJ, Ware JH. Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. **J Am Stat Assoc** 1988; 83: 631–637.
12. Smith AH, Waxweiler RJ, Tyroler HA. Epidemiologic investigation of occupational carcinogenesis using a serially additive expected dose model. **Am J Epidemiol** 1980; 112: 787–797.
13. Smith AH. Looking backwards from outcome to exposure to assess cancer latency. **J Chron Dis** 1987; 40, (Suppl. 2): 113S–117S.
14. Schervish MJ. Algorithm AS 195. Multivariate normal probabilities with error bound. **Appl Stat** 1984; 33: 81–94.
15. Prentice RL, Pyke R. Logistic disease incidence models and case–control studies. **Biometrika** 1979; 66: 403–411.
16. Breslow N, Powers W. Are there two logistic regressions for retrospective studies? **Biometrics** 1978; 34: 100–105.
17. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. **Biometrics** 1984; 40: 961–971.
18. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. **Biometrika** 1986; 73: 13–22.
19. Terris M, Wilson F, Smith H, Sprung E, Nelson JH. The relationship of coitus to carcinoma of the cervix. **Am J Public Health** 1967; 57: 840–847.
20. Slattery ML, Overall JC, Abbott TM, French TK, Robison LM, Gardner J. Sexual activity, contraception, genital infections, and cervical cancer: Support for a sexually transmitted disease hypothesis. **Am J Epidemiol** 1987; 130: 248–258.
21. Prentice RL, Thomas DB. On the epidemiology of oral contraceptives and disease. **Adv Cancer Res** 1987; 49: 285–401.
22. Greenland S. Interpreting time-related trends in effect estimates. **J Chron Dis** 1987; 40 (Suppl. 2): 17S–24S.

APPENDIX

Deviation of the Covariance Estimator of Equation (2)

Let β_r, β_s be the true coefficient vectors at a pair (r, s) of time points. Taylor series expansions of the respective maximum likelihood estimators about the true values yield:

$$\hat{\beta}_r - \beta_r \approx -(X_r' V_r X_r)^{-1} X_r' V_r^{1/2} \mathbf{r}_r|_{\beta_r},$$

$$\hat{\beta}_s - \beta_s \approx -(X_s' V_s X_s)^{-1} X_s' V_r^{1/2} \mathbf{r}_s|_{\beta_s},$$

where $\mathbf{r}_r$, $\mathbf{r}_s$ are the Pearson residual vectors. Then we have:

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_r, \hat{\beta}_s) &= (X_r' V_r X_r)^{-1} X_r' V_r^{1/2} \\ &\quad \times \mathrm{Cov}(\mathbf{r}_r, \mathbf{r}_s) V_s^{1/2} X_s (X_s' V_s X_s)^{-1} \\ &= (X_r' V_r X_r)^{-1} X_r' V_r^{1/2} E(\mathbf{r}_r \mathbf{r}_s) V_s^{1/2} X_s (X_s' V_s X_s)^{-1},\end{aligned}$$

which may be estimated consistently by substitution of the sample quantities y_i, $\hat{p}_{ir}$, $\hat{p}_{is}$ to arrive at

$$\begin{aligned}\widehat{\mathrm{Cov}}(\hat{\beta}_r, \hat{\beta}_s) &= (X_r' V_r X_r)^{-1} X_r' V_r^{1/2} \\ &\quad \times \mathrm{diag}(r_{ir} r_{is}) V_s^{1/2} X_s (X_s' V_s X_s)^{-1}.\end{aligned}$$

The central diagonal matrix reflects the assumption of independence between subjects, but not across time within subject.