

The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk

Michael Hauptmann^{1,*†‡}, Jay H. Lubin², Philip Rosenberg², Jürgen Wellmann³
and Lothar Kreienbrock⁴

¹*GSF National Research Center for Environment and Health, Institute of Epidemiology, Neuherberg, Germany*

²*National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd., Bethesda, MD 20892, U.S.A.*

³*University of Münster, Institute of Epidemiology and Social Medicine, Domagkstrasse 3, 48129 Münster, Germany*

⁴*Hannover Veterinary School, Institute of Biometry and Epidemiology, Bünteweg 2, 30559 Hannover, Germany*

SUMMARY

To examine the time-dependent effects of exposure histories on disease we use sliding time windows as an exploratory alternative to the analysis of variables like time since last exposure and duration of exposure. The method fits a series of risk models which contain total cumulative exposure and an additional covariate for exposures received during fixed time intervals. Characteristics of the fitted models provide insight into the influence of exposure increments at different times on disease risk. A simulation study is performed to check the validity of the approach. We apply the method to data from a recent German case-control study on smoking and lung cancer risk with about 4300 lung cancer cases and a similar number of controls. The sliding time window approach indicates that the amount of cigarettes smoked from two to 11 years before disease incidence is most predictive of lung cancer incidence. Among different smoking profiles that result in the same lifelong cumulative number of cigarettes smoked, those with a concentration of smoked cigarettes within 20 years before interview bear substantially larger risk than others. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

We present an exploratory method to evaluate the time-dependent effects of exposure histories on a binary disease outcome in a case-control setting. The motivation for such an approach comes from studies of acute exposures, where time since exposure is clearly defined and often has great influence on the risk of disease. The study of Japanese atomic bomb survivors is a prime example [1]. For extended exposures, the definition of time since exposure and its relationship to disease outcome are no longer obvious [2].

*Correspondence to: Michael Hauptmann, National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd., EPS/7089, Bethesda, MD 20892, U.S.A.

†E-mail: hauptmann@nih.gov

‡Now at National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Blvd., EPS/7089, Bethesda, MD 20892, U.S.A.

The proposed method fits a series of risk models which include total cumulative exposure and an additional covariate for exposure received during a fixed time interval. Characteristics of the fitted models provide insight into the influence of exposure increments at different times on disease risk.

The approach has been previously applied in References [3] and [4]. We present a mathematical characterization of the method and extend it by adjusting for cumulative exposure and performing a two-dimensional profile likelihood estimation of the best fitting time window. A simulation study is performed to check the validity of the method.

As an example, we apply the methodology to data from a recent German case-control study on smoking and lung cancer with about 4300 cases and a similar number of controls, an association which has been investigated by various authors [5–7].

2. METHOD: THE SLIDING TIME WINDOW

The method fits a series of models, which include total cumulative exposure and cumulative exposure received within a defined time interval, often referred to as an exposure time window (Muller and Kusiak in Reference [2]). Let y_j denote the disease status of individual j ($j = 1, \dots, n$), and let $x_j(t)$ denote the exposure of the j th individual at time t before interview ($t \in [0, T]$), where T depends on the length of collected exposure histories. Additional covariates $z_j = (z_{1j}, \dots, z_{mj})'$ are used to adjust for confounding.

We sequentially fit models that include cumulative exposure to attained age, A , and cumulative exposure received during a time interval of fixed width k as covariates. Intervals of various width k can be considered. For the time window centred at time c before interview, where $c \in [k/2, T - k/2]$, we fit the model M_c of the form

$$\begin{aligned} \text{logit Pr}(y_j = 1 | z_j, x_j(t), t \in [0, T]) \\ = \alpha_0 + \alpha' z_j + \beta_1 \int_0^{A_j} x_j(t) dt + \beta_2 \int_{c-k/2}^{c+k/2} x_j(t) dt \end{aligned} \quad (1)$$

and compute the likelihood ratio test statistic

$$\text{LR}_c = -2 \log \frac{\max_{\alpha, \beta} L(M_c | \beta_2 = 0)}{\max_{\alpha, \beta} L(M_c)}$$

which compares model (1) to the corresponding ‘null’ model without the time window exposure variable ($\beta_2 = 0$). The value of c is then varied over its range. For fixed c , parameter β_1 represents the increase in the log-odds ratio (OR) per unit exposure, while β_2 represents the additive effect (on a log scale) of a unit exposure that occurred during the specific time window of length k centred at time c . The likelihood ratios between the models with and without the time window, LR_c , can be compared to assess the significance of the additional exposure variable.

The approach is equivalent to a profile likelihood estimation of the non-linear parameter c denoting the time window midpoint of the best fitting time window. To avoid the arbitrary selection of window width k , the approach can be extended to a two-dimensional profile likelihood estimation of both the width and the position of the best fitting time window. On the other hand, the window width can be viewed as a smoothing parameter for the series of estimated time window parameters.

Since total cumulative exposure may confound the association between a specific time window exposure and disease, it is included in the model. If one prefers to adjust only for exposure during the time not covered by the time window under consideration, estimates can be easily derived as the sum of $\hat{\beta}_1$ and $\hat{\beta}_2$ of formula (1) for the new time window estimate while leaving the estimate $\hat{\beta}_1$ for the lifelong exposure parameter unchanged. The sum of the two parameters is then interpreted as the exclusive effect of exposure during the time window in comparison to the effect β_1 of exposures received during all other times.

Model (1) is equivalent to the more general model

$$\text{logit Pr}(y_j = 1 | z_j, x_j(t), t \in [0, T]) = \alpha_0 + \alpha' z_j + \beta_1 \int_0^{A_j} w(t) x_j(t) dt$$

where the weight function $w(\cdot)$ is given by

$$w(t) = 1 + \theta I_{[c-k/2, c+k/2]}(t)$$

and θ can be estimated as $\hat{\beta}_2/\hat{\beta}_1$.

3. EXAMPLE: CASE-CONTROL STUDY ON SMOKING AND LUNG CANCER

We apply the method to data from a case-control study carried out from 1990–1996 in Germany [8]. Cases include patients aged 75 years and under with histologically confirmed primary lung cancer. Controls are population-based and frequency-matched to cases on age (within five years), sex, and place of residence (23 regions).

Data on smoking history are obtained by personal interview. Information on the type and amount of tobacco products smoked and on inhalation habits are obtained by intervals of constant smoking habit.

For a cigarette smoker who also smoked cigars, cigarillos or pipes, the tobacco amount equivalent is added to his exposure from cigarettes. After excluding 174 cigars, cigarillos, or pipes only smokers and 43 individuals with incomplete smoking histories, the study population includes 4304 cases and 4526 controls.

Preliminary analysis revealed that a log-linear model for the OR provided a better fit than a linear OR model. Therefore, the former is used throughout the remainder of the text. The odds ratio for a smoker compared to a never-smoker is 18.19 for males (95 per cent CI: [14.00, 23.62]) and 4.71 for females (95 per cent CI: [3.79, 5.87]) adjusted for asbestos exposure and the matching variables. Using total cumulative pack-years smoked (1 pack-year = 365 × 20 cigarettes), the risk of lung cancer, relative to a never-smoker, is generally decreasing with categories of years since quitting smoking (current smoker, 1–4, 5–9, 10+). The odds ratios and corresponding 95 per cent confidence intervals are 12.10 [8.99, 16.29], 15.15 [10.91, 21.03], 8.02 [5.72, 11.25], 4.10 [3.10, 5.43] for males, respectively, and 2.45 [1.64, 3.68], 2.90 [1.68, 4.99], 1.36 [0.75, 2.47], 0.89 [0.61, 1.30] for females.

The slightly higher risk for individuals who recently quit smoking versus current smokers is due to the well-known phenomenon that people tend to stop smoking when they start to feel ill so that current smoking becomes protective compared to having quit recently.

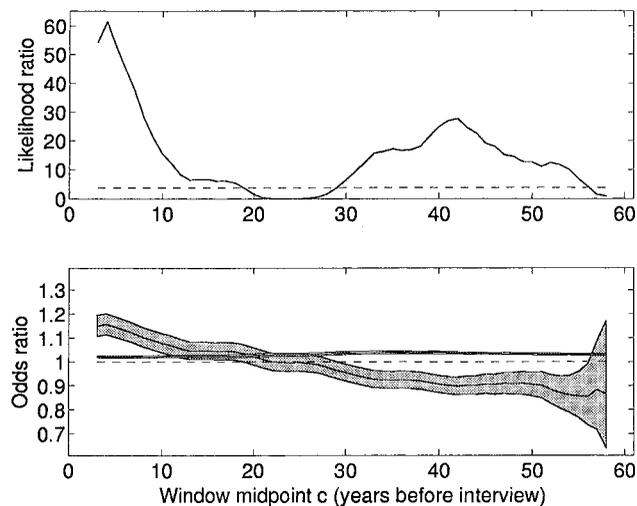


Figure 1. Results of the sliding five-year time window analysis for $n = 7094$ males. The upper panel shows the likelihood ratio (solid line, maximum at $\hat{c} = 4$) and the $\chi^2_{1,0.95}$ -quantile (dashed line). The lower panel shows the odds ratios per unit pack-year with pointwise 95 per cent confidence bands ($\exp(\hat{\beta}_1)$, narrow band) and the multiplicative effect of the time window pack-years ($\exp(\hat{\beta}_2)$, wide band), and a reference line at unity (dashed line).

4. RESULTS

For this example, time before interview is taken as discrete, and $x_j(t)$ denotes the number of pack-years smoked by the j th individual ($j = 1, \dots, 8830$) during year t before interview ($t = 1, \dots, 75$). Because of the discrete time scale, integration in model (1) is replaced by summation. All analyses are adjusted for the matching variables and for time since quitting smoking in years (non-smoker, current smoker, 1–4, 5–9, 10+). Analyses restricted to males only are additionally adjusted for asbestos exposure (ever/never).

Figures 1 and 2 present the results for men and women, respectively, for model (1) using a fixed window of width five years ($k = 5$). The upper panel shows the likelihood ratio test statistic LR_c that compares the fit of the null model with the adjustment variables and total pack-years of smoking with the model that additionally includes the cumulative pack-years smoked during the time window centred at c years prior to age at interview. The lower panel shows the pointwise 95 per cent confidence intervals of the estimated odds ratios for one cumulative pack-year (narrow band) and for the cumulative time window pack-years (wide band).

Figure 1 shows that for men the maximum of LR_c occurs for the time window from two to six years before interview, that is, $\hat{c} = 4$. The odds ratio for the corresponding time window indicates a significantly positive extra effect, after adjustment for lifelong exposure. In contrast, the local maximum for the likelihood ratio occurring at 42 years before interview corresponds to a negative effect for pack-years smoked many years ago relative to the overall effect of total cumulative pack-years.

The time window parameter shows an antagonistic behaviour with a positive effect for exposures received within the last 20 years and a negative effect for exposures received more than 20 years

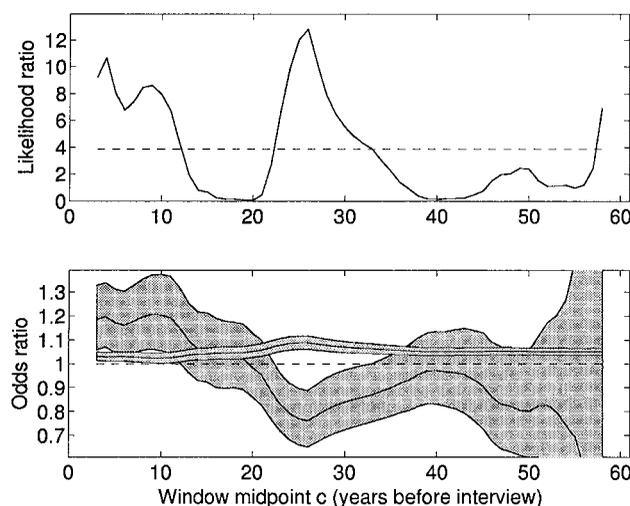


Figure 2. Results of the sliding five-year time window analysis for $n = 1736$ females. The upper panel shows the likelihood ratio (solid line, maximum at $\hat{c} = 25$) and the $\chi^2_{1,0.95}$ -quantile (dashed line). The lower panel shows the odds ratios per unit pack-year with pointwise 95 per cent confidence bands ($\exp(\hat{\beta}_1)$, narrow band) and the multiplicative effect of the time window pack-years ($\exp(\hat{\beta}_2)$, wide band), and a reference line at unity (dashed line).

ago. This antagonism will always occur with total pack-years whenever there is a positive time window effect for some segment of the exposure history, since there must be a counter-balancing negative effect for some other segment. The relatively small influence from estimating the time window parameter β_2 on β_1 is due to the greater amount of data used to estimate the β_1 value and thus its greater stability.

For women, Figure 2 shows that the best fit corresponding to a positive contribution of time window exposure is about four years before interview with a second local maximum at nine years before interview (corresponding to time windows from two to six years before interview and from seven to 11 years before interview, respectively). The pattern in females is essentially similar to that of men, but with greater uncertainty in the estimates for women due to smaller sample size. The odds ratio for lifelong exposure again remains rather constant for the different time windows.

Figures 1 and 2 indicate that a specific time window model may result in an odds ratio less than one, that is, less than that of a non-smoker, for a smoker who smokes a high percentage of his total number of cigarettes within a short time period more than 20 years ago. Such extreme profiles are not present in the underlying data, and the estimated odds ratios are highly variable and would have wide confidence limits.

The likelihood based on model (1) can be maximized in both c and k . The maximum value of the likelihood function corresponding to a positive time window parameter estimate is reached at the one-year time window two years before interview for males as well as for females. The 'best' fitting time windows with an associated negative time window parameter estimate are from six to 56 years before interview and from 24 to 25 years before interview for males and females, respectively. The results are consistent with the one-dimensional approach.

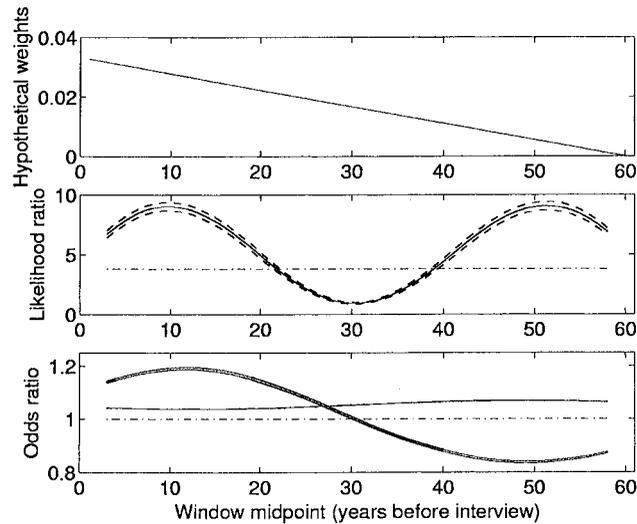


Figure 3. Simulation study results of the sliding five-year time window analysis based on 1000 generated case-control studies with 1000 observations (half cases and controls). Data were generated as described in the text using linearly decreasing weights (shown in the upper panel). The middle panel shows the mean likelihood ratio (solid line) of the replications and approximate pointwise normal 95 per cent confidence intervals (dashed lines). The $\chi^2_{1,0.95}$ -quantile (dashed-dotted line) is also shown. The lower panel shows the pointwise 95 per cent confidence bands for the simulated odds ratios per unit pack-year ($\exp(\hat{\beta}_1)$, narrow band) and the multiplicative effect of the time window pack-years ($\exp(\hat{\beta}_2)$, wide band).

5. SIMULATION STUDY

A simulation study is performed to check the validity of the method and its robustness with respect to uncertainties in exposure assessment.

Hypothetical smoking profiles are generated for up to 60 years prior to interview following the German smoking data. A weighted cumulative exposure is calculated using different time-dependent weights: constant; linearly increasing and decreasing; triangle shape; trapezoidal shape. The response variable is generated using the weighted cumulative exposure within a linear logistic regression model by sampling a synthetic retrospective study from a prospective study according to Reference [9].

This procedure results in simulated case-control studies based on five scenarios represented by different hypothetical weights. For each scenario, 1000 case-control studies with 500 cases and 500 controls each are generated. For details of the simulation study design refer to the Appendix.

Figures 3 and 4 show the results of the simulation study for two given hypothetical weights (upper panels). The approximate normal pointwise 95 per cent confidence bands of the mean time window odds ratio from 1000 replications (wide band in the lower panel) are in good accordance with the given weights and indicate a small simulation variability. The likelihood ratio (middle panel) yields high values when the absolute value of the time window parameter estimate is largest, representing the significance of the corresponding time window exposure with respect to the null model containing total cumulative pack-years only.

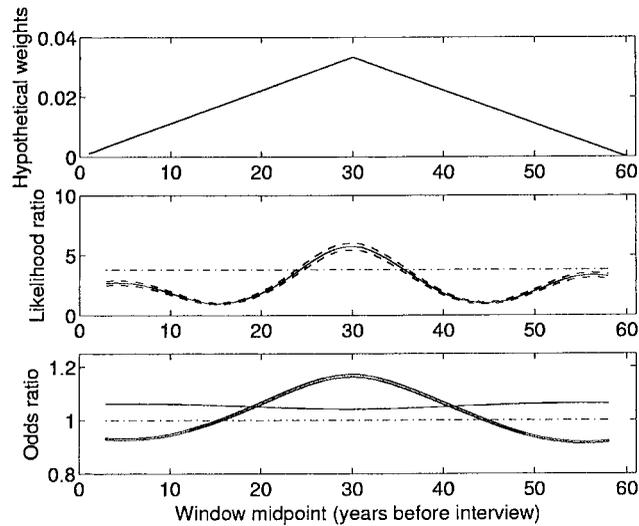


Figure 4. Simulation study results of the sliding five-year time window analysis based on 1000 generated case-control studies with 1000 observations (half cases and controls). Data were generated as described in the text using triangle shaped weights (shown in the upper panel). The middle panel shows the mean likelihood ratio (solid line) of the replications and approximate pointwise normal 95 per cent confidence intervals (dashed lines). The $\chi^2_{1,0.95}$ -quantile (dashed-dotted line) is also shown. The lower panel shows the pointwise 95 per cent confidence bands for the simulated odds ratios per unit pack-year ($\exp(\hat{\beta}_1)$, narrow band) and the multiplicative effect of the time window pack-years ($\exp(\hat{\beta}_2)$, wide band).

Next we evaluate the effects of measurement error. A problem with retrospective collection of exposure histories is that exposures many years ago may be subject to greater measurement error than more recent exposures. We examine the sensitivity of the method to time-dependent uncertainties in exposure assessment, with greater uncertainty occurring earlier in exposure histories.

Instead of the generated exposure profiles $x(t)$, $t = 1, \dots, 60$, we use the erroneous profiles

$$\tilde{x}(t) = x(t)e(t)$$

where the time-dependent multiplicative measurement error $e(t)$ is uniformly distributed with expectation one and with variance increasing with time t before interview.

For all hypothetical weights used for data generation and different sizes of error, the erroneous profiles yield similar results to the true profiles. Also we observe only a very small amount of attenuation of parameter estimates. Therefore, no illustrations are shown.

In a second approach, we assume that the probability that a person who smoked ℓ years ago reports not having smoked at that time increases with ℓ independent of disease status.

This type of measurement error causes a remarkable bias of the time window parameter estimate towards the null. The bias increases with increasing probability of erroneously reporting not to have smoked at certain times. The simulation also shows that such an error results in large standard errors of estimated coefficients so that none of the time windows yields a significant contribution to the model, not even time windows with high weights assigned. No illustrations are shown.

Both measurement error simulations suggest that the patterns observed in Figures 1 and 2 are not likely to be a result of measurement errors of these types.

6. DISCUSSION

For the analysis of individual exposure histories, the sliding time window approach presented here is an alternative to the analysis of time since last exposure and pack-years. It evaluates the contribution to risk of exposures in increments of prior years.

In general, the approach is not limited to logistic regression models nor is it limited to case-control studies, and it can be performed with standard statistical software.

Using the sliding time window approach, we find that for a fixed number of pack-years the amount smoked less than 20 years previously is mostly responsible for an increased risk, with special emphasis on the time windows covering the interval from two to six years before interview. Cigarettes smoked within this time result in a higher risk than cigarettes smoked more than 20 years before interview.

Quitting smoking means no exposure during the last years before interview. Since all analyses are adjusted for time since smoking cessation, the decreasing effect of time windows with time before interview cannot be ascribed to the well known decline of risk with time since quitting smoking. Moreover, the analysis suggests that the effect of duration of smoking given a certain amount of cigarettes smoked depends on the period of exposure.

An extension of the approach is possible by using different time scales, namely time since exposure (as done here), age at exposure, or calendar year. However, these variables are highly correlated, so that interpretation of an effect may be problematic. As to the example presented, we can argue as follows. Data were collected retrospectively within a few years, so a major effect of calendar year can be excluded. We performed the sliding time window analysis stratified for attained age on the time since exposure scale. Since the resulting likelihood ratio pattern was similar among strata, the effects found above can be ascribed to time since exposure.

The simulation study shows the method's ability to find given time patterns of exposure weights. The sensitivity analysis suggests that the sliding time window approach is rather robust against a time-dependent multiplicative random measurement error that may corrupt the collected exposure profiles.

In our situation, a substantial systematic error in retrospective assessment of exposure histories causes biased time window parameter estimates with large variances throughout the time scale. In contrast to this, results based on the real data show a clear and consistent pattern over age and sex strata with significant estimates for certain time periods. Therefore, it is unlikely that our results are the consequence of such an error.

In summary, the sliding time window approach adds new information about the influence of temporal patterns of smoking habits on lung cancer risk. The approach is easy to implement using standard software. However, the method is exploratory in nature, and therefore cannot replace careful modelling of a dose-response relationship.

APPENDIX: SIMULATION STUDY DESIGN

Hypothetical smoking profiles are generated for up to 60 years prior to interview following the German smoking data. With probability 0.25, an individual is considered to be a lifelong non-smoker.

With probability 0.75, the number of pack-years smoked during the year prior to interview is taken from the following distribution

$$x_i(1) = \max(0, 0.47 + 0.3672 v_i), \quad i = 1, \dots, n \quad (\text{A1})$$

where $v_i \sim N(0, 1)$. The truncated normal distribution is chosen so that the expectation of its untruncated counterpart equals the mean yearly exposure rate in the German smoking data and that the probability of zero exposure is 0.1.

Changes in smoking rates over time are modelled as follows. With a 50 per cent chance of change in 10 years we define the probability p_t of a change in smoking rate at year t before interview such that

$$0.5 = (1 - p_t)^{10}.$$

If the subject has changed smoking rate, then we resample a new smoking rate from distribution (A1).

Given time-dependent weights $w(t)$, we generate the response using the weight function model

$$\text{logit } \Pr(y_j = 1 | x_j(t), t \in [0, T]) = \beta_0 + \beta_1 \sum_{t=1}^T w(t) x_i(t) \quad (\text{A2})$$

with given parameter values β_0 and β_1 .

We use several simple weight functions $w(t)$: constant; linearly increasing and decreasing; triangle shape; trapezoidal shape. They are all standardized so that their integral on $[0, T]$ is unity.

A value for β_1 is chosen from the case-control data. Since the time-weighted exposure $\sum w(t)x(t)$ is related to exposure rate, we substitute the log odds ratio estimated from the empirical data, $\log \widehat{\text{OR}} = 3.163$, for β_1 .

For case-control data, β_0 is the log odds of being a case for a never-smoker, that is, at $x(t) = 0$ for all $t = 1, \dots, T$. Since we have half cases and half controls in our study, we set the probability of being a case at mean exposure to one half, that is

$$\Pr(y = 1 | \bar{x}) = 0.5 \iff \beta_0 = -\beta_1 \bar{x}.$$

With a mean exposure rate of $\bar{x} = 0.47$ pack-years, we have $\beta_0 = -1.49$.

The response generation imitates a synthetic retrospective study sampled from a prospective study according to Reference [9]. To generate cases, we perform a Bernoulli experiment for each candidate profile with probability of success $\Pr(y_j = 1 | x_j(t), t \in [0, T])$ of model (3). If the experiment is successful, the candidate profile is added to our simulated data set as a case profile. If not, we take the next candidate profile. For controls, a candidate profile enters the data set, if the experiment fails, and is rejected otherwise.

The generated exposure profiles are corrupted in two ways to imitate uncertainties in exposure assessment. First, instead of the generated exposure profiles $x(t)$, $t = 1, \dots, 60$, we use the erroneous profiles

$$\tilde{x}(t) = x(t)e(t)$$

where the time-dependent multiplicative measurement error $e(t)$ is uniformly distributed according to

$$e(t) \sim \text{U} \left[1 - \frac{t(1-p)}{60}, 1 + \frac{t(1-p)}{60} \right].$$

The measurement error distribution is 'linearly increasing' in time t from $U[1, 1]$ at interview to $U[p, 2 - p]$ at 60 years before interview for $p = 0, 0.25, 0.5, 0.75, 1$.

Secondly, we assume that a person who smoked ℓ years ago reports not having smoked at that time with increasing probability as ℓ increases. Therefore, the erroneous profiles

$$\tilde{x}(t) = x(t)e(t)$$

are used, where $e(t)$ is a Bernoulli variate with probability of success $tp/60$ for $p = 0.33, 0.5, 0.75$.

REFERENCES

1. Shimizu Y, Kato H, Schull WJ. Studies of the mortality of a-bomb survivors. 9. Mortality, 1950-85: Part 2. Cancer mortality based on the recently revised doses (DS86). *Radiation Research* 1990; **121**:120-141.
2. United States National Academy of Sciences, National Research Council. *Committee on Biological Effects of Ionizing Radiation: Health Risks on Radon and other Internally Deposited Alpha Emitters*. National Academy Press, Washington D.C., 1988.
3. Finkelstein MM. Use of time windows to investigate lung cancer latency intervals at an Ontario steel plant. *American Journal of Industrial Medicine* 1991; **19**:229-235.
4. Finkelstein MM. Clinical measures, smoking, radon exposure, and risk of lung cancer in uranium miners. *Occupational and Environmental Medicine* 1996; **53**:697-702.
5. Doll R, Peto R. Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong nonsmokers. *Journal of Epidemiology and Community Health* 1987; **32**:303-313.
6. Lubin JH, Blot WJ, Berrino F, Flamant R, Gillis CR, Kunze M, Schmaehl D, Visco G. Modifying risk of developing lung cancer by changing habits of cigarette smoking. *British Medical Journal* 1984; **288**:1953-1956.
7. Becher H, Jöckel KH, Timm J, Wichmann HE, Drescher K. Smoking cessation and nonsmoking intervals: effect of different smoking patterns on lung cancer risk. *Cancer Causes and Control* 1991; **2**:381-387.
8. Kreienbrock L, Wichmann HE, Gerken M, Heinrich J, Götze H-J, Kreuzer M, Keller G. The German radon project - feasibility of methods and first results. *Radiation Protection Dosimetry* 1992; **45**:643-649.
9. Mantel N. Synthetic retrospective studies and related topics. *Biometrics* 1973; **29**:479-486.