

Using lowess to remove systematic trends over time in predictor variables prior to logistic regression with quantile categories[‡]

Craig B. Borkowf^{1,*†}, Paul S. Albert² and Christian C. Abnet¹

¹*National Cancer Institute, Center for Cancer Research, Cancer Prevention Studies Branch, 6116 Executive Blvd., Suite 705, MSC 8314, Bethesda, MD 20892-8314, U.S.A.*

²*National Cancer Institute, Division of Cancer Treatment and Diagnosis, Biometric Research Branch, Executive Plaza North, Room 8136, 6130 Executive Blvd, MSC 7438, Bethesda, MD 20892-7438, U.S.A.*

SUMMARY

In case-control studies one may employ logistic regression to model the relationship between binary responses and continuous predictor variables that have been categorized by the empirical quartiles of the controls. Sometimes, however, systematic trends over time (or drifts) contaminate the laboratory measurements of predictor variables. In this paper we consider the use of locally weighted robust regression (lowess) to estimate and remove these systematic trends when the trends for the cases and controls have a common shape. One can then use the lowess adjusted data in the desired logistic regression model. We illustrate these methods with a case-control study that was designed to assess the risk of oesophageal cancer as a function of the quartile categories of sphinganine levels in the blood serum. Upon examination of the data, it was discovered that the sphinganine laboratory measurements were contaminated by a systematic trend, the magnitude of which depended only on the day of analysis. This trend needed to be removed before performing further analyses of the data. In addition, we present simulations to examine the use of lowess methods to estimate and remove various shapes of trends from contaminated predictor data before constructing logistic regression models with quartile categories. We found that using the trend-contaminated data tends to give attenuated parameter estimates and hence lower significance and power levels than using the uncontaminated data. Conversely, using appropriate lowess methods to adjust the data tends to give nearly unbiased parameter estimates, near nominal significance levels, and improved power. Published in 2003 by John Wiley & Sons, Ltd.

KEY WORDS: cancer; laboratory measurement drift; lowess; measurement error; quantile-category; trend removal

1. INTRODUCTION

In case-control studies we may employ logistic regression to model a binary response, y_i (for example, disease status), as a function of a continuous predictor variable, x_i , where i indexes

*Correspondence to: Craig B. Borkowf, Centers for Disease Control and Prevention, National Center for Infectious Diseases, Division of Viral and Rickettsial Diseases, Influenza Branch, Epidemiology Section, Mail Stop A32, 1600 Clifton Road NE, Atlanta GA 30333, U.S.A.

† E-mail: CBorkowf@cdc.gov

‡ This article is a U.S. Government work and is in the public domain in the U.S.A.

the subjects. In some situations we may categorize this predictor variable by the empirical quartiles of the control measurements, $\{x_i | y_i = 0\}$, to create four new variables, $\{q_j(x_i)\}$, such that $q_j(x_i) = 1$ if measurement x_i falls in the j th quartile category and 0 otherwise ($j = 1, \dots, 4$). Ideally, one would then fit the logistic regression model $\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_j(x_i)$. This approach is particularly desirable because it allows for non-linear relationships between the predictor quartile categories and the response.

Next, let t_i denote the time of measurement for the i th subject. Unfortunately, when the true predictor variable is contaminated by a systematic trend over time (or drift), $\tau(t)$, one cannot use the logistic regression model $\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_j(x_i + \tau(t_i))$ because the definitions of the quartile categories themselves depend on the trend-contaminated data, and thus the trend becomes inextricably tied to the predictor variables; neither can one use the conditional logistic regression model [1] $\text{logit}(P\{Y_i = 1\}) = \alpha_k + \sum_{j=1}^4 \beta_j q_j(x_i + \tau(t_i))$, with strata for the time-unit of analysis $\{\alpha_k\}$, for the same reasons. Rather, one must first estimate and remove this trend before one constructs the quartile categories, and then one may perform the desired logistic regression analyses. Note that similar problems would arise for other non-linear functions of the continuous predictor variable (for example, quadratic terms) and for other non-linear link functions (for example, the probit link).

Consider the following motivating example. The General Population Trial (GPT), conducted in Linxian, China, was a large-scale randomized placebo-controlled trial designed to evaluate the effects of multiple mineral and vitamin supplementation on the risk of oesophageal cancer [2, 3]. Prior to baseline (March 1986), blood samples were drawn from all 29 584 subjects and frozen for future use. After 5.25 years of follow-up (May 1991), the subjects were classified as to case or control status. Subsequently (1997), researchers at the U.S. National Cancer Institute designed a case-control study nested within the larger prospective cohort from the GPT to determine whether sphingolipids, including sphinganine (Sa), are useful predictors of the risk of oesophageal cancer [4]. Sphingolipids have been measured in the blood serum and urine, and their concentrations may serve as biomarkers for exposure to fungal toxins from the consumption of contaminated corn and wheat [5, 6].

In this nested case-control study, the designers stratified the subjects by three age categories (30–50, 51–60 and 61–69 years old at baseline) and two genders (male, female), in order to create six age-by-gender strata. They then selected about 17 oesophageal cancer cases and 34 controls from each of these six strata; thus the cases and controls were frequency matched, but not individually matched. Owing to the loss of samples and other chance occurrences, however, blood samples for only 98 cases and 182 controls were ultimately available for laboratory analysis. In preparation for analysis, the case blood samples were placed in a random order, and then two control samples from the same stratum as each case sample were placed nearby. These samples were subsequently coded to blind their identities and shipped to an off-site laboratory for the measurement of the serum Sa levels by high performance liquid chromatography (HPLC) analysis. Each sample was measured up to twice and the averages of the measurements were calculated.

Figure 1 shows a plot of the natural logarithm of the Sa measurements ($\ln\text{-Sa}$) for the controls (circles) and cases (squares) by day of HPLC analysis. We observe several striking features in this plot. First, we note that the HPLC measurements were made on days 1–4 and 19–57. The gap from days 5 to 18 corresponds to a period in which the laboratory was adjusting its HPLC system. Furthermore, several outliers occur on days 23 and 24. Second, the measurements for the cases and controls have an apparent trend over time, contrary to

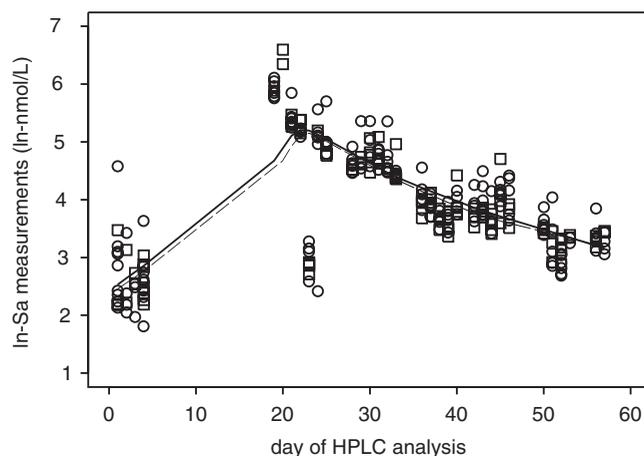


Figure 1. Plot of trend-contaminated ln-Sa measurements versus day of HPLC analysis. This figure shows the scatter plot of the trend-contaminated ln-Sa measurements (in ln-nanomols/litre) for the controls (circles) and cases (squares) by day of HPLC analysis (day 1=9/23/97 to day 57=11/18/97). It also shows the lowess estimated trends using a data-fraction of $f=2/3$ for both the controls (solid line) and cases (dashed line). The means and standard deviations of the ln-Sa data for the controls are 4.00 ± 0.96 and for the cases are 3.87 ± 0.92 , and the empirical quartiles of the controls are 3.36, 3.90 and 4.65 (interquartile range, IQR = 1.29).

the expectation that these measurements should be both independent of time and identically distributed.

We next used locally weighted robust regression [7] (lowess) to estimate the mean trends in the ln-Sa measurements for the cases and controls. Lowess methods provide a non-parametric, robust local smooth of the scatter plot data using a tricubic weight function. The smoothness of the fit increases as the fraction of the data, $f \in [0, 1]$, used to compute the mean at each abscissa value, increases. A smaller data-fraction gives a more local smooth, and thus fits finer features in the data, while a larger data-fraction gives a less local smooth, and thus reveals global features. We used the S-Plus 2000 programming language [8], including the predefined function *lowess*, for all calculations in this paper. Figure 1 also shows the lowess estimated trends using the default data-fraction of $f=2/3$ for both the controls (solid line) and cases (dashed line). The observed trends for the cases and controls are quite similar in location and shape. Both trends rise sharply from day 1 to day 19 and then decrease more gradually to day 57. Note that the analyses that we describe below depend only upon the shapes of curves in the regions where data are present, and not on gaps or areas of interpolation.

The laboratory scientists suggested that the common trend was due to a chemical contaminant that eluted from the HPLC column at the same time as the Sa. They believed that this trend depended only on the day of HPLC analysis and thus was independent of the Sa measurement levels. Unfortunately, the laboratory standards contained such high levels of Sa that they did not detect this trend. Likewise, the quality control samples, constructed from blood samples pooled more recently from a Linxian bloodbank, contained significantly higher levels of Sa than the samples in the case-control study (perhaps due to better preservation or seasonal variation in Sa levels), and hence did not accurately detect this trend either. Thus, there are no independent estimates of the observed trend.

We wished to make inference to the risk of oesophageal cancer as a function of the quartile categories of Sa levels among the population at large. Ideally, we would construct logistic regression models for the risk of oesophageal cancer as a function of the quartile categories defined by Sa data for the controls only, plus other subject covariates. Therefore, in order to be able to perform the desired analyses, we employed lowess methods to estimate and remove the observed trend from the Sa data for the cases and controls themselves.

In Section 2 we develop models for case-control studies in which the laboratory analysis of the predictor variable takes place over significant time periods. We first describe a logistic regression model with the predictor variable defined by the quartile categories of the controls of the uncontaminated (true) predictor data. We then describe models for the trend-contaminated data, the estimation and removal of the trend by lowess methods, and the corresponding logistic regression models for the trend-contaminated data and the lowess adjusted data. In Section 3 we analyse the motivating example in light of these models. Next, in Section 4 we describe a simulation study designed to compare the statistical properties of these models. Finally, in Section 5 we discuss the broader implications of using lowess methods to estimate and remove trends from contaminated data in case-control and other studies.

2. MODELS FOR PREDICTOR VARIABLES IN CASE-CONTROL STUDIES

2.1. A model for the uncontaminated (true) data

We now describe a model for the analysis of a case-control study by logistic regression with quartile categories of a continuous predictor variable measured over time. First, let n_0 denote the number of controls, let n_1 denote the number of cases, and let $n = n_0 + n_1$. We also assume that there is no matching between the cases and controls. Next, let y_i denote the response variable, coded as $y_i = 0$ or 1 if the i th subject is a control or case, respectively ($i = 1, \dots, n$).

Now, let x_i^{true} denote the uncontaminated (true) continuous predictor measurement for the i th subject. For simplicity, we will not consider any other covariates. Next, using the indicator function $I\{\cdot\}$, let $\hat{F}^{\text{true}}(x) = n_0^{-1} \sum_{i=1}^n I\{x_i^{\text{true}} \leq x, y_i = 0\}$ denote the empirical distribution function of the uncontaminated data for the controls, $\{x_i^{\text{true}} \mid y_i = 0\}$. Then, let u_α^{true} denote the empirical α -quantile of the uncontaminated data for the controls, namely

$$u_\alpha^{\text{true}} = \inf\{x \mid \hat{F}^{\text{true}}(x) \geq \alpha\} \quad (1)$$

By convention, set $u_0^{\text{true}} = -\infty$ and $u_1^{\text{true}} = +\infty$. In turn, we can define the indicator variables for the four quartile categories ($j = 1, \dots, 4$)

$$q_{ij}^{\text{true}} = 1 \quad \text{if } u_{(j-1)/4}^{\text{true}} < x_i^{\text{true}} \leq u_{j/4}^{\text{true}} \text{ for the } i\text{th subject and } 0 \text{ otherwise} \quad (2)$$

Note that quantile categories are invariant to monotonic increasing (almost everywhere) transformations of the underlying continuous data. Now, let $\text{logit}(p) = \log_e(p/(1-p))$ denote the logistic link function and let $P\{Y_i = 1\}$ denote the probability that the i th subject is a case. We first consider the following logistic regression equation:

$$\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_{ij}^{\text{true}} \quad (3)$$

By convention, $\beta_1=0$. Thus, each of the β_j should be interpreted with reference to the first quartile category. For example, this model implies that the odds of being a case are $\exp(\beta_4)$ times as large for individuals with the x -predictor in the fourth quartile category as for those with the x -predictor in the first quartile category. The likelihood ratio test (LRT) for the inclusion of the quartile categories tests the null hypothesis that $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ against the alternative that $H_1 : \beta_j \neq 0$ for at least one j ($j = 2, 3, 4$) based on three degrees of freedom (d.f.).

Finally, let d denote the number of days (or other time-unit) over which the measurement analyses of the predictor variable take place, and let $t_i \in [0, 1]$ denote the fraction of the total time at which the i th subject's measurement analysis occurs. Ideally, the same number of controls and cases should be analysed on each day in order to avoid confounding between case-control status and time of analysis.

2.2. Models for the trend-contaminated data and the lowess adjusted data

We now consider models for the construction of the trend-contaminated data and the use of lowess methods to estimate and remove trends in order to compute the lowess adjusted data. We also define logistic regression models corresponding to (3) for each of these variables. Let $\tau(t)$ denote a generic trend function of the time-fraction t on the closed interval $[0, 1]$. We assume that this trend function is completely independent of the true data. The addition of this systematic trend to the true data $\{x_i^{\text{true}}\}$ creates the trend-contaminated data

$$x_i^{\text{trend}} = x_i^{\text{true}} + \tau(t_i) \tag{4}$$

Next, we can use standard lowess methods with data-fraction f to estimate the trend from the contaminated data, $\{x_i^{\text{trend}}\}$. First, we can simply estimate the trend from the measurements for the controls only, $\{x_i^{\text{trend}} | y_i = 0\}$. Second, when the numbers of cases and controls are (nearly) balanced on each day of analysis, we can estimate the trend from the measurements for the cases and controls combined, $\{x_i^{\text{trend}}\}$, that is, using the complete data set. This second method essentially produces a weighted average of the respective lowess curves for the cases and controls. More generally, when the numbers of cases and controls are unbalanced, one can use advanced versions of lowess methods, such as those given by the S-Plus function *loess* [8], with an indicator variable for case-control status, to construct a weighted estimate of the common trend. The lowess estimated trends from these data are denoted by $\hat{\tau}_c(t | f)$ (c = controls only) and $\hat{\tau}_b(t | f)$ (b = both cases and controls), respectively. In turn, we obtain the following two new variables by removing the trend:

$$x_i^{\text{lowc}} = x_i^{\text{trend}} - \hat{\tau}_c(t_i | f) = x_i^{\text{true}} + \tau(t_i) - \hat{\tau}_c(t_i | f) \tag{5}$$

and

$$x_i^{\text{lowb}} = x_i^{\text{trend}} - \hat{\tau}_b(t_i | f) = x_i^{\text{true}} + \tau(t_i) - \hat{\tau}_b(t_i | f) \tag{6}$$

Note that for a satisfactory estimator $\hat{\tau}(t | f)$, the curve $\tau(t) - \hat{\tau}(t | f)$ should be nearly constant over time with unknown mean near $-\bar{x}_i^{\text{true}}$, and hence the lowess adjusted data $\{x_i^{\text{lowc}}\}$ and $\{x_i^{\text{lowb}}\}$ should have means near zero. Thus, the process of trend estimation and removal loses information about the absolute values of the uncontaminated measurements. Nevertheless, this process ideally recovers information about the relative differences between the uncontaminated measurements, which is sufficient in order to construct the quartile categories of the data.

Next, we can use the variables $\{x_i^{\text{trend}}\}$, $\{x_i^{\text{lowc}}\}$ and $\{x_i^{\text{lowb}}\}$ to define the empirical quartiles $\{u_{j/4}^{\text{trend}}\}$, $\{u_{j/4}^{\text{lowc}}\}$ and $\{u_{j/4}^{\text{lowb}}\}$, respectively, as in equation (1). In turn, we can use these empirical quartiles to define the indicator variables for the four quartile categories ($j = 1, \dots, 4$) $\{q_{ij}^{\text{trend}}\}$, $\{q_{ij}^{\text{lowc}}\}$ and $\{q_{ij}^{\text{lowb}}\}$, respectively, as in equation (2). Finally, corresponding to the fundamental logistic regression (3), we can use each of these three new sets of variables to write

$$\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_{ij}^{\text{trend}} \quad (7)$$

$$\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_{ij}^{\text{lowc}} \quad (8)$$

and

$$\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_{ij}^{\text{lowb}} \quad (9)$$

Again, by convention, $\beta_1 = 0$. We expect that the use of equation (7), with the trend-contaminated data, will give estimates of $\{\beta_j\}$ that are attenuated towards zero. Hence, the LRT will have low power for testing the null hypothesis of $H_0: \beta_2 = \beta_3 = \beta_4 = 0$. Conversely, we hope that the use of equations (8) and (9), with the lowess adjusted data, will give less biased estimates of $\{\beta_j\}$, which will in turn give near nominal significance levels and improved power.

3. ANALYSIS OF THE OESOPHAGEAL CANCER CASE-CONTROL STUDY

3.1. Estimation and removal of the observed trend in the Sa measurements

In the analysis of the motivating example, we examined the nature of the common systematic trend in the Sa measurements for the cases and controls in order to estimate and remove it. We first constructed a series of regression models to determine whether the observed trend in the ln-Sa measurements might be due to an unlucky randomization of the sample orders and thus explained by a covariate confounded with the day of HPLC analysis. In these regression models, we regarded the ln-Sa measurements as the response and the age and gender stratification variables and other patient covariates available from the GPT (such as smoking, drinking, cholesterol and nutritional intervention treatment group) as predictor variables. None of these variables explained the systematic trend over time in the ln-Sa measurements.

Next, we used lowess methods with a data-fraction of $f = 2/3$ to estimate the systematic trend in the ln-Sa data. We also experimented with other possible data-fractions and obtained comparable results. To maximize the efficiency of estimation, we used both the cases and controls combined to estimate the common trend. We then subtracted the lowess estimate of the trend from the contaminated ln-Sa data to obtain the lowess adjusted ln-Sa data.

Figure 2 shows a plot of the lowess adjusted ln-Sa measurements for the controls (circles) and cases (squares) by day of HPLC analysis. It also shows the trends in the lowess adjusted data, estimated by lowess using a data-fraction of $f = 2/3$, for both the controls (solid line) and cases (dashed line). We observe that both trends are approximately horizontal, as expected, with means near zero. Recall that while the lowess adjusted data loses information about

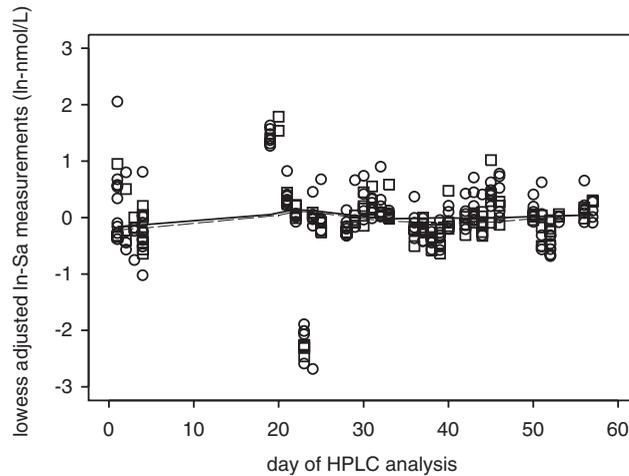


Figure 2. Plot of lowess adjusted ln-Sa measurements versus day of HPLC analysis. This figure shows the scatter plot of the lowess adjusted ln-Sa measurements (in ln-nanomols/litre) for the controls (circles) and cases (squares) by day of HPLC analysis (day 1 = 9/23/97 to day 57 = 11/18/97). It also shows the lowess estimated trends using a data-fraction of $f = 2/3$ for both the controls (solid line) and cases (dashed line). The means and standard deviations of the lowess adjusted ln-Sa data for the controls are 0.020 ± 0.600 and for the cases are -0.091 ± 0.605 , and the empirical quartiles of the controls are $-0.228, 0.002$ and 0.266 (IQR = 0.494).

the absolute ln-Sa measurements, it ideally recovers information about the relative differences between these measurements.

3.2. Analysis of the lowess adjusted ln-Sa data

As mentioned above, we wished to make inference to the risk of oesophageal cancer as a function of the quartile categories of Sa levels among the population at large. To do so, we need to estimate the population quartiles using the empirical quartiles of the measurements for the controls only. We now compare the results below for logistic regression models that use the empirical quartile categories created from the trend-contaminated ln-Sa data and from the lowess adjusted ln-Sa data.

First, in addition to the notation of the previous section, we define the following indicator variables for each subject ($i = 1, \dots, n$) for the covariates for age, gender, smoking and drinking, all determined at baseline. Let $a_{ij} = 1$ if the i th subject's age is between (i) 30 and 50 inclusive, (ii) 51 and 60 inclusive, and (iii) 61 to 69 inclusive for $j = 1, 2, 3$, respectively, and 0 otherwise; let $g_i = 1$ if the i th subject's gender is female, and 0 if male; let $s_i = 1$ if the i th subject smoked for a total of 6 or more months at any time, and 0 otherwise; and let $d_i = 1$ if the i th subject drank any alcoholic beverage during the previous 12 months, and 0 otherwise.

Using the trend-contaminated ln-Sa data $\{x_i^{\text{trend}}\}$, we fit the following logistic regression model:

$$\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_{ij}^{\text{trend}} + \beta_5 a_{i2} + \beta_6 a_{i3} + \beta_7 g_i + \beta_8 s_i + \beta_9 d_i \quad (10)$$

Table I. Parameter estimates, standard errors and odds ratios for the model of the risk of oesophageal cancer as a function of the quartile categories computed from the trend-contaminated ln-Sa data.

Variable	Symbol	β_j	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	p -value	OR	95 per cent CI for OR
Intercept	1	β_0	-0.616	0.339	0.07		
Quartile 2	q_{i2}	β_2	0.278	0.343	0.42	1.320	(0.674, 2.586)
Quartile 3	q_{i3}	β_3	-0.252	0.376	0.50	0.778	(0.372, 1.624)
Quartile 4	q_{i4}	β_4	-0.192	0.366	0.60	0.826	(0.403, 1.691)
Age [51,60]	a_{i2}	β_5	-0.037	0.313	0.91	0.964	(0.522, 1.780)
Age [61,69]	a_{i3}	β_6	-0.112	0.313	0.72	0.894	(0.484, 1.651)
Gender (f)	g_i	β_7	-0.487	0.405	0.23	0.614	(0.278, 1.359)
Smoking	s_i	β_8	0.630	0.414	0.13	1.877	(0.833, 4.229)
Drinking	d_i	β_9	0.356	0.345	0.30	1.427	(0.725, 2.807)

This table shows the parameters $\{\beta_j\}$ for model (1), and the corresponding estimates $\{\hat{\beta}_j\}$, standard errors ($\{SE(\hat{\beta}_j)\}$), p -values, odds ratios (ORs) $\{\exp(\hat{\beta}_j)\}$ and 95 per cent confidence intervals (CIs) for the ORs of the form $\{\exp[\hat{\beta}_j \pm 1.96SE(\hat{\beta}_j)]\}$. The LRT for the inclusion of the quartile categories gives $G^2 = 2.68$ on 3 degrees of freedom (p -value = 0.44).

Table I shows the parameter estimates, standard errors, p -values, odds ratios (ORs) and 95 per cent confidence intervals (CIs) for ORs for the terms in model (10). Thus, $\exp(\beta_j)$ gives the OR for the risk of oesophageal cancer for the j th quartile category of the ln-Sa measurements compared to the first quartile category ($j = 2, 3, 4$). The ORs for the quartile categories are close to one and thus provide no evidence that Sa exposure increases the risk of oesophageal cancer (all p -values ≥ 0.42). Also, the LRT for the inclusion of the quartile categories is not significant at the 5 per cent level ($G^2 = 2.68$, 3 d.f., p -value = 0.44). Furthermore, none of the covariates is a significant predictor of the risk of oesophageal cancer.

By comparison, using the lowess adjusted ln-Sa data $\{x_i^{\text{lowb}}\}$, we fit the following logistic regression model:

$$\text{logit}(P\{Y_i = 1\}) = \beta_0 + \sum_{j=1}^4 \beta_j q_{ij}^{\text{lowb}} + \beta_5 a_{i2} + \beta_6 a_{i3} + \beta_7 g_i + \beta_8 s_i + \beta_9 d_i \quad (11)$$

Table II shows the parameter estimates, standard errors, p -values, ORs and 95 per cent CIs for ORs for the terms in model (11). As before, the ORs for the second and third quartile categories are very close to one (both p -values ≥ 0.95), while the OR for the fourth quartile category has decreased to 0.556 (p -value = 0.12), opposite the expected direction, but still insignificant. Next, the LRT for the inclusion of the quartile categories is still not significant at the 5 per cent level ($G^2 = 3.38$, 3 d.f., p -value = 0.34), although the p -value has decreased, as expected. Furthermore, none of the covariates is a significant predictor either. Thus, we conclude from this study that Sa is not a useful predictor for oesophageal cancer. Indeed, the fact that the original trends for the cases and controls are virtually identical in location supports this conclusion (Figures 1 and 2).

As mentioned in the introduction, we were concerned about the effects of the apparent outliers and gap on the analysis of the data. We therefore repeated the above analyses with and without the nine outliers on days 23 and 24 and on various subsets of the data. In particular, we examined the subsets of data after the gap (days 19–57) and after the both the

Table II. Parameter estimates, standard errors and odds ratios for the model of the risk of oesophageal cancer as a function of the quartile categories computed from the lowess adjusted ln-Sa data.

Variable	Symbol	β_j	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$	p -value	OR	95 per cent CI for OR
Intercept	1	β_0	-0.528	0.326	0.11		
Quartile 2	q_{i2}	β_2	0.014	0.343	0.97	1.014	(0.518, 1.989)
Quartile 3	q_{i3}	β_3	-0.021	0.349	0.95	0.979	(0.494, 1.942)
Quartile 4	q_{i4}	β_4	-0.587	0.380	0.12	0.556	(0.264, 1.172)
Age [51,60]	a_{i2}	β_5	-0.052	0.312	0.87	0.949	(0.515, 1.751)
Age [61,69]	a_{i3}	β_6	-0.129	0.315	0.68	0.879	(0.474, 1.629)
Gender (f)	g_i	β_7	-0.431	0.403	0.29	0.650	(0.295, 1.434)
Smoking	s_i	β_8	0.639	0.414	0.12	1.895	(0.842, 4.268)
Drinking	d_i	β_9	0.319	0.343	0.35	1.375	(0.702, 2.693)

This table shows the parameters $\{\beta_j\}$ for model (2), and the corresponding estimates $\{\hat{\beta}_j\}$, standard errors ($\{SE(\hat{\beta}_j)\}$), p -values, odds ratios (ORs) $\{\exp(\hat{\beta}_j)\}$ and 95 per cent confidence intervals (CIs) for the ORs of the form $\{\exp[\hat{\beta}_j \pm 1.96SE(\hat{\beta}_j)]\}$. The LRT for the inclusion of the quartile categories gives $G^2 = 3.38$ on 3 degrees of freedom (p -value = 0.34).

gap and outliers (days 25–57). The results of these additional analyses were comparable to those for the entire data set, which is not surprising since the main results of this study are negative. It also suggests that lowess methods provide a sufficiently local smooth to the data so as not to be significantly affected by the odd features mentioned above. Since we had no *a priori* reasons to exclude any data, we have reported the analyses on the entire data set in this paper.

Also, in addition to the analyses reported here, we performed analyses with models that included other covariates and interactions (particularly the age-by-gender interaction). None of these additional models yielded significant results either.

4. DESIGN AND RESULTS OF A SIMULATION STUDY

4.1. Parameter settings, generation of data and shapes of trends

We were concerned that the contaminating trend in the laboratory measurements had perhaps obscured meaningful differences between the cases and controls, and we wished to determine whether lowess methods could indeed recover these differences if they existed. We also recognize that the standard errors in the logistic regression equations (8) and (9) do not account for the variation due to the estimation of the trend, which may potentially cause spuriously high rejection rates. We therefore designed a simulation study to compare the use of logistic regression with predictor variables defined by the quartile categories of the control data only. Recall that the four logistic regression models defined in Section 2 use the uncontaminated data (3), the trend-contaminated data (7), the lowess adjusted data computed using the controls only (8), and the lowess adjusted data computed using both the cases and controls combined (9). We considered sample sizes of (a) $n_0 = n_1 = 100$, (b) $n_0 = n_1 = 200$ and (c) $n_0 = 200$ and $n_1 = 100$. In the motivating example, $n_0 = 182$ and $n_1 = 98$, which is closest to the last case. We then set $y_i = 0$ for the first n_0 observations and $y_i = 1$ for the last n_1 observations.

Next, we simulated the values of $\{x_i^{\text{true}}\}$ as follows. If $y_i = 0$, we simulated an observation x_i^{true} from the standard normal distribution (mean 0, variance 1), whereas if $y_i = 1$, we simulated an observation x_i^{true} from a normal distribution with mean $\mu = 0.0, 0.25$ or 0.35 and variance 1. Thus, the x -predictors are independent and identically distributed (i.i.d.) for the controls and cases, separately.

We set $d = 50$ in all simulations and randomly assigned n_0/d controls and n_1/d cases to each day of analysis. Thus, either two or four cases and controls were analysed on each day. The time-fractions of analysis, t_i , were set in the middle of each day, so the analyses for days $j = 1, 2, 3, \dots, 50$ occur at times $t = (j - 0.5)/d = 0.01, 0.03, 0.05, \dots, 0.99$.

In addition, we considered two models for the shape of the added trend, a step function and a triangular function. The step function corresponds to a sudden perturbation, whereas the triangular function allows for a change in the direction of the trend. One can define a step function with magnitude η and jump at $t = \zeta$ as

$$\tau_{\text{step}}(t | \eta, \zeta) = \eta I\{\zeta < t \leq 1\} \quad (12)$$

Similarly, one can define a triangular function with magnitude η and peak at $t = \zeta$ as

$$\tau_{\text{tri}}(t | \eta, \zeta) = \eta(t/\zeta)I\{0 \leq t \leq \zeta\} + \eta[(1-t)/(1-\zeta)]I\{\zeta < t \leq 1\} \quad (13)$$

Note that when $\zeta = 0$ or 1 , the triangular function simply gives a sloped line. The magnitudes of these functions are relative to the standard deviation of the uncontaminated data, which is 1. We then constructed the values of $\{x_i^{\text{trend}}\}$ by equation (4).

Because the x -predictor is assumed to be i.i.d. given case-control status, the addition, estimation and removal of the three trends $\tau(t), \pm\tau(t) \pm \delta$ and $\pm\tau(1-t) \pm \delta$, where δ is any constant displacement, will have the same impact on the estimation of $\{\beta_j\}$ ($j = 2, 3, 4$), although clearly the value of β_0 will depend on $\pm\delta$. Thus, without loss of generality, we only consider below the addition of trends with positive magnitudes and with peaks or jumps on the interval $[0.5, 1]$. Specifically, we consider magnitudes $\eta = 1$ and 2 for both the step and triangular functions. For the step function, we consider jumps at $\zeta = 0.5$ and 0.75 , while for the triangular function, we consider peaks at $\zeta = 0.5, 0.75$ and 1.0 (a slope). Inferences for other situations can be made by due alteration of details. By comparison, in the motivating example, the trend appears to have a triangular shape with a peak at about $\zeta = 19/57 = 0.33$ and a magnitude of about $\eta = 2$ (although it could also be modelled by a simple slope on days 19–57 if the initial days and gap were excluded).

Next, we used standard lowess methods for trend estimation given by the S-Plus function *lowess* (described above) with data-fractions of $f = 1/3, 1/2$ and $2/3$. We then computed the lowess adjusted data $\{x_i^{\text{lowc}}\}$ and $\{x_i^{\text{lowb}}\}$ using equations (5) and (6), respectively.

4.2. Numerical results for the simulation study

Tables III, IV and V show the simulated rejection rates for the LRT for the inclusion of the quartile categories at the 10 per cent significance level, $G^2 > \chi_{3,0.9}^2 = 6.25$, for sample sizes of (a) $n_0 = n_1 = 100$, (b) $n_0 = n_1 = 200$, and (c) $n_0 = 200$ and $n_1 = 100$, respectively. The last eight columns of these tables correspond to the following models and methods of analysis: column 1, equation (3) with the uncontaminated data $\{x_i^{\text{true}}\}$; column 2, equation (7) with the trend-contaminated data $\{x_i^{\text{trend}}\}$; columns 3–5, equation (8) with the lowess adjusted data using the controls only $\{x_i^{\text{lowc}}\}$ and data-fractions of $f = 1/3, 1/2$ and $2/3$; columns 6–8,

Table III. Simulated rejection rates for the LRT for the inclusion of the quartile categories for various underlying distributions, trends and methods of analysis for sample sizes of $n_0 = n_1 = 100$ at the 10 per cent significance level.

μ	Trend parameters			x_i^{true}	x_i^{trend}	Method of analysis					
	η	τ	ζ			$\{x_i^{\text{lowc}}\}$ with f			$\{x_i^{\text{lowb}}\}$ with f		
						1/3	1/2	2/3	1/3	1/2	2/3
0.00	1	Step	0.50	0.105	0.086	0.121	0.111	0.106	0.103	0.100	0.100
		Step	0.75	0.099	0.092	0.121	0.109	0.105	0.104	0.103	0.102
		Triangular	0.50	0.104	0.097	0.117	0.107	0.102	0.105	0.100	0.102
		Triangular	0.75	0.099	0.093	0.116	0.107	0.106	0.099	0.099	0.097
0.00	2	Slope	1.00	0.098	0.098	0.118	0.109	0.106	0.101	0.103	0.101
		Step	0.50	0.101	0.063	0.116	0.100	0.096	0.095	0.089	0.089
		Step	0.75	0.098	0.065	0.119	0.102	0.099	0.099	0.095	0.093
		Triangular	0.50	0.096	0.077	0.115	0.107	0.103	0.102	0.101	0.101
0.00	2	Triangular	0.75	0.103	0.081	0.119	0.107	0.099	0.105	0.099	0.097
		Slope	1.00	0.101	0.082	0.119	0.111	0.106	0.104	0.103	0.105
		Step	0.50	0.362	0.293	0.374	0.359	0.357	0.365	0.356	0.354
		Step	0.75	0.367	0.311	0.376	0.368	0.359	0.370	0.364	0.360
0.25	1	Triangular	0.50	0.370	0.340	0.380	0.373	0.369	0.377	0.376	0.368
		Triangular	0.75	0.358	0.338	0.378	0.369	0.366	0.372	0.366	0.365
		Slope	1.00	0.371	0.350	0.389	0.377	0.373	0.382	0.381	0.378
		Step	0.50	0.358	0.152	0.357	0.337	0.322	0.347	0.339	0.318
0.25	2	Step	0.75	0.358	0.215	0.357	0.335	0.323	0.346	0.338	0.317
		Triangular	0.50	0.371	0.278	0.393	0.382	0.374	0.382	0.379	0.373
		Triangular	0.75	0.355	0.271	0.379	0.363	0.355	0.364	0.361	0.350
		Slope	1.00	0.357	0.272	0.380	0.363	0.364	0.369	0.367	0.365
0.35	1	Step	0.50	0.591	0.498	0.598	0.586	0.577	0.593	0.586	0.579
		Step	0.75	0.586	0.525	0.594	0.579	0.576	0.590	0.585	0.579
		Triangular	0.50	0.587	0.560	0.593	0.585	0.591	0.596	0.596	0.592
		Triangular	0.75	0.588	0.564	0.607	0.594	0.590	0.604	0.596	0.593
0.35	2	Slope	1.00	0.588	0.554	0.602	0.596	0.594	0.603	0.598	0.599
		Step	0.50	0.582	0.274	0.570	0.555	0.534	0.572	0.553	0.534
		Step	0.75	0.588	0.387	0.578	0.556	0.544	0.575	0.557	0.543
		Triangular	0.50	0.591	0.468	0.605	0.595	0.592	0.599	0.595	0.588
0.35	2	Triangular	0.75	0.593	0.467	0.597	0.587	0.574	0.598	0.589	0.575
		Slope	1.00	0.585	0.463	0.594	0.588	0.587	0.593	0.593	0.593

This table shows the simulated rejection rates for the LRT for the inclusion of the quartile categories for sample sizes of $n_0 = n_1 = 100$ at the 10 per cent significance level. The columns denote the eight methods of analysis described in the text. The rows denote the means of the cases (μ), the shapes of the trends (τ , step or triangular/slope), the magnitudes of the trends (η), and the locations of the jump for step trends and the peak for triangular/slope trends (ζ). Note that the same simulated $M = 10\,000$ data sets were used to calculate the rejection rates for all entries in the same row.

equation (9) with the lowess adjusted data using both the case and controls $\{x_i^{\text{lowb}}\}$ and data-fractions of $f = 1/3, 1/2$ and $2/3$. The row headings in these tables indicate the mean of the cases ($\mu = 0.0, 0.25, 0.35$; recall that the controls have mean zero), the magnitude of the trend ($\eta = 1, 2$), and the shape of the trend (step with jumps at $\zeta = 0.5$ and 0.75 , triangular with peaks at $\zeta = 0.5, 0.75$ and 1.0).

When the cases and controls have the same means ($\mu = 0$), we observe that using the uncontaminated data $\{x_i^{\text{true}}\}$ gives rejection rates near the nominal value of 0.1. By comparison,

Table IV. Simulated rejection rates for the LRT for the inclusion of the quartile categories for various underlying distributions, trends and methods of analysis for sample sizes of $n_0 = n_1 = 200$ at the 10 per cent significance level.

μ	Trend parameters			x_i^{true}	x_i^{trend}	Method of analysis					
	η	τ	ζ			$\{x_i^{\text{lowc}}\}$ with f			$\{x_i^{\text{lowb}}\}$ with f		
						1/3	1/2	2/3	1/3	1/2	2/3
0.00	1	Step	0.50	0.101	0.085	0.108	0.102	0.098	0.098	0.102	0.102
		Step	0.75	0.105	0.091	0.110	0.104	0.103	0.104	0.102	0.102
		Triangular	0.50	0.099	0.091	0.109	0.101	0.099	0.101	0.101	0.101
		Triangular	0.75	0.099	0.095	0.111	0.108	0.104	0.104	0.103	0.099
		Slope	1.00	0.101	0.093	0.110	0.102	0.104	0.102	0.101	0.102
		Slope	1.00	0.101	0.093	0.110	0.102	0.104	0.102	0.101	0.102
0.00	2	Step	0.50	0.097	0.058	0.101	0.098	0.091	0.092	0.091	0.086
		Step	0.75	0.097	0.069	0.099	0.091	0.089	0.089	0.093	0.086
		Triangular	0.50	0.102	0.082	0.110	0.105	0.101	0.104	0.102	0.100
		Triangular	0.75	0.099	0.082	0.113	0.104	0.097	0.097	0.102	0.095
		Slope	1.00	0.104	0.080	0.111	0.109	0.107	0.108	0.107	0.109
		Slope	1.00	0.104	0.080	0.111	0.109	0.107	0.108	0.107	0.109
0.25	1	Step	0.50	0.593	0.498	0.588	0.579	0.577	0.594	0.585	0.579
		Step	0.75	0.605	0.538	0.605	0.600	0.594	0.606	0.596	0.591
		Triangular	0.50	0.594	0.564	0.599	0.596	0.594	0.601	0.602	0.594
		Triangular	0.75	0.597	0.565	0.598	0.598	0.596	0.599	0.598	0.596
		Slope	1.00	0.595	0.572	0.607	0.601	0.602	0.604	0.603	0.599
		Slope	1.00	0.595	0.572	0.607	0.601	0.602	0.604	0.603	0.599
0.25	2	Step	0.50	0.599	0.283	0.579	0.566	0.545	0.580	0.568	0.543
		Step	0.75	0.601	0.395	0.589	0.573	0.559	0.584	0.568	0.560
		Triangular	0.50	0.595	0.472	0.603	0.596	0.594	0.599	0.598	0.594
		Triangular	0.75	0.597	0.472	0.599	0.589	0.581	0.601	0.590	0.579
		Slope	1.00	0.588	0.478	0.600	0.593	0.593	0.597	0.598	0.596
		Slope	1.00	0.588	0.478	0.600	0.593	0.593	0.597	0.598	0.596
0.35	1	Step	0.50	0.861	0.793	0.857	0.850	0.850	0.859	0.851	0.852
		Step	0.75	0.865	0.819	0.863	0.860	0.858	0.866	0.865	0.864
		Triangular	0.50	0.861	0.839	0.863	0.861	0.862	0.865	0.858	0.859
		Triangular	0.75	0.858	0.837	0.860	0.859	0.858	0.862	0.858	0.857
		Slope	1.00	0.860	0.838	0.858	0.855	0.857	0.861	0.859	0.859
		Slope	1.00	0.860	0.838	0.858	0.855	0.857	0.861	0.859	0.859
0.35	2	Step	0.50	0.858	0.554	0.846	0.838	0.826	0.846	0.837	0.831
		Step	0.75	0.860	0.681	0.844	0.836	0.829	0.846	0.836	0.827
		Triangular	0.50	0.869	0.768	0.869	0.867	0.865	0.868	0.868	0.863
		Triangular	0.75	0.862	0.773	0.862	0.860	0.846	0.861	0.859	0.848
		Slope	1.00	0.871	0.780	0.868	0.868	0.869	0.874	0.874	0.874
		Slope	1.00	0.871	0.780	0.868	0.868	0.869	0.874	0.874	0.874

This table shows the simulated rejection rates for the LRT for the inclusion of the quartile categories for sample sizes of $n_0 = n_1 = 200$ at the 10 per cent significance level. The columns denote the eight methods of analysis described in the text. The rows denote the means of the cases (μ), the shapes of the trends (τ , step or triangular/slope), the magnitudes of the trends (η), and the locations of the jump for step trends and the peak for triangular/slope trends (ζ). Note that the same simulated $M = 10\,000$ data sets were used to calculate the rejection rates for all entries in the same row.

using the trend-contaminated data $\{x_i^{\text{trend}}\}$ gives subnominal rejection rates. The rejection rates decrease more for step trends than for triangular or slope trends, and more for trends with larger magnitudes ($\eta = 2$ versus $\eta = 1$). Conversely, using the lowess adjusted data $\{x_i^{\text{lowc}}\}$ with $f = 1/3$ tends to give supranominal rejection rates, especially for small sample sizes, so this method should not be considered further. Furthermore, using the lowess adjusted data $\{x_i^{\text{lowc}}\}$ with $f = 1/2$ and $2/3$ and $\{x_i^{\text{lowb}}\}$ with $f = 1/3, 1/2$ and $2/3$ tends to give near nominal

Table V. Simulated rejection rates for the LRT for the inclusion of the quartile categories for various underlying distributions, trends and methods of analysis for sample sizes of $n_0 = 200$ and $n_1 = 100$ at the 10 per cent significance level.

μ	Trend parameters			x_i^{true}	x_i^{trend}	Method of analysis					
	η	τ	ζ			$\{x_i^{\text{lowc}}\}$ with f			$\{x_i^{\text{lowb}}\}$ with f		
						1/3	1/2	2/3	1/3	1/2	2/3
0.00	1	Step	0.50	0.102	0.094	0.109	0.104	0.104	0.106	0.102	0.102
		Step	0.75	0.106	0.092	0.113	0.107	0.106	0.110	0.105	0.104
		Triangular	0.50	0.104	0.098	0.112	0.108	0.106	0.108	0.103	0.107
		Triangular	0.75	0.106	0.105	0.110	0.114	0.108	0.109	0.106	0.103
0.00	2	Slope	1.00	0.101	0.096	0.105	0.103	0.101	0.102	0.105	0.102
		Step	0.50	0.105	0.063	0.108	0.102	0.096	0.104	0.100	0.098
		Step	0.75	0.109	0.069	0.107	0.101	0.099	0.101	0.101	0.097
		Triangular	0.50	0.104	0.076	0.109	0.103	0.100	0.107	0.104	0.099
0.00	2	Triangular	0.75	0.102	0.079	0.102	0.102	0.097	0.102	0.099	0.096
		Slope	1.00	0.100	0.080	0.104	0.103	0.101	0.102	0.100	0.100
		Step	0.50	0.449	0.358	0.453	0.445	0.439	0.458	0.448	0.442
		Step	0.75	0.443	0.390	0.438	0.431	0.431	0.443	0.438	0.433
0.25	1	Triangular	0.50	0.456	0.422	0.458	0.456	0.453	0.464	0.458	0.459
		Triangular	0.75	0.454	0.420	0.455	0.452	0.448	0.458	0.457	0.451
		Slope	1.00	0.449	0.412	0.454	0.452	0.453	0.455	0.453	0.455
		Step	0.50	0.454	0.195	0.434	0.416	0.402	0.438	0.421	0.403
0.25	2	Step	0.75	0.446	0.277	0.426	0.413	0.403	0.430	0.415	0.412
		Triangular	0.50	0.447	0.339	0.458	0.454	0.453	0.457	0.460	0.452
		Triangular	0.75	0.458	0.344	0.455	0.449	0.436	0.460	0.456	0.436
		Slope	1.00	0.454	0.344	0.454	0.449	0.450	0.455	0.455	0.455
0.35	1	Step	0.50	0.710	0.622	0.709	0.706	0.702	0.712	0.710	0.704
		Step	0.75	0.707	0.651	0.702	0.698	0.695	0.706	0.699	0.698
		Triangular	0.50	0.710	0.676	0.706	0.706	0.708	0.711	0.710	0.709
		Triangular	0.75	0.709	0.675	0.709	0.704	0.706	0.716	0.714	0.710
0.35	2	Slope	1.00	0.710	0.675	0.707	0.710	0.708	0.714	0.714	0.710
		Step	0.50	0.706	0.378	0.688	0.675	0.658	0.692	0.678	0.664
		Step	0.75	0.706	0.494	0.685	0.674	0.661	0.689	0.679	0.670
		Triangular	0.50	0.703	0.579	0.704	0.699	0.696	0.704	0.702	0.699
0.35	2	Triangular	0.75	0.711	0.593	0.720	0.707	0.692	0.713	0.712	0.694
		Slope	1.00	0.707	0.586	0.711	0.704	0.703	0.710	0.708	0.707

This table shows the simulated rejection rates for the LRT for the inclusion of the quartile categories for sample sizes of $n_0 = 200$ and $n_1 = 100$ at the 10 per cent significance level. The columns denote the eight methods of analysis described in the text. The rows denote the means of the cases (μ), the shapes of the trends (τ , step or triangular/slope), the magnitudes of the trends (η), and the locations of the jump for step trends and the peak for triangular/slope trends (ζ). Note that the same simulated $M = 10\,000$ data sets were used to calculate the rejection rates for all entries in the same row.

rejection rates, and note that the rejection rates tend to decrease slightly as the data-fraction f increases.

Next, when the mean of the cases is greater than that of the controls ($\mu = 0.25, 0.35$), using the trend-contaminated data $\{x_i^{\text{trend}}\}$ gives substantially lower power compared to using the uncontaminated data $\{x_i^{\text{true}}\}$, as expected, since the presence of the trend tends to obscure differences between the cases and controls. In particular, power decreases more for step trends (especially with jumps at $\zeta = 0.5$), for trends with larger magnitudes, and for smaller sample sizes. Conversely, using the lowess adjusted data $\{x_i^{\text{lowc}}\}$ and $\{x_i^{\text{lowb}}\}$ gives power levels

Table VI. Simulated raw biases in the logistic regression estimates of β_4 for various underlying distributions, trends, and methods of analysis for sample sizes of $n_0 = n_1 = 100$.

μ	Trend parameters			x_i^{true}	x_i^{trend}	Method of analysis					
	η	τ	ζ			$\{x_i^{\text{lowc}}\}$ with f			$\{x_i^{\text{lowb}}\}$ with f		
						1/3	1/2	2/3	1/3	1/2	2/3
0.00	1	Step	0.50	-0.004	-0.001	-0.003	-0.002	-0.003	-0.002	-0.001	-0.002
		Step	0.75	0.002	0.003	-0.001	0.002	0.002	0.001	-0.000	0.000
	Triangular	0.50	-0.002	-0.002	-0.002	-0.001	0.000	-0.002	-0.002	-0.002	-0.001
		0.75	0.005	0.007	0.004	0.005	0.004	0.005	0.005	0.005	0.006
		1.00	0.002	0.001	0.002	0.002	0.002	0.001	0.001	0.001	0.000
0.00	2	Step	0.50	0.003	0.001	0.000	0.000	0.000	-0.001	0.001	0.001
		Step	0.75	0.010	0.017	0.007	0.009	0.006	0.008	0.007	0.006
	Triangular	0.50	0.002	0.003	0.002	0.003	0.003	0.001	0.001	0.001	0.002
		0.75	-0.000	0.002	0.001	-0.001	0.001	0.000	-0.001	-0.001	-0.001
		1.00	-0.001	-0.000	-0.002	-0.002	-0.002	-0.001	-0.002	-0.002	-0.001
0.25	1	Step	0.50	0.008	-0.058	-0.026	-0.020	-0.015	0.011	0.005	0.002
		Step	0.75	0.016	-0.033	-0.019	-0.013	-0.012	0.018	0.010	0.006
	Triangular	0.50	0.016	-0.010	-0.018	-0.007	-0.001	0.022	0.020	0.018	0.018
		0.75	0.009	-0.018	-0.023	-0.012	-0.008	0.020	0.016	0.016	0.010
		1.00	0.018	-0.005	-0.011	0.000	0.004	0.030	0.025	0.024	0.024
0.25	2	Step	0.50	0.014	-0.179	-0.038	-0.037	-0.042	0.001	-0.013	-0.028
		Step	0.75	0.009	-0.122	-0.042	-0.041	-0.042	-0.004	-0.019	-0.025
	Triangular	0.50	0.018	-0.074	-0.010	-0.001	0.004	0.028	0.024	0.018	0.018
		0.75	0.013	-0.078	-0.017	-0.010	-0.013	0.022	0.017	0.001	0.001
		1.00	0.011	-0.077	-0.022	-0.011	-0.006	0.021	0.016	0.015	0.015
0.35	1	Step	0.50	0.021	-0.077	-0.024	-0.016	-0.013	0.026	0.019	0.010
		Step	0.75	0.022	-0.044	-0.027	-0.019	-0.015	0.026	0.018	0.013
	Triangular	0.50	0.020	-0.018	-0.024	-0.009	-0.002	0.032	0.027	0.023	0.023
		0.75	0.024	-0.016	-0.019	-0.003	0.001	0.037	0.031	0.024	0.024
		1.00	0.025	-0.012	-0.023	-0.008	0.001	0.037	0.032	0.029	0.029
0.35	2	Step	0.50	0.011	-0.255	-0.059	-0.059	-0.068	-0.006	-0.027	-0.048
		Step	0.75	0.020	-0.159	-0.048	-0.048	-0.049	0.002	-0.015	-0.028
	Triangular	0.50	0.021	-0.102	-0.022	-0.007	-0.001	0.035	0.029	0.023	0.023
		0.75	0.017	-0.110	-0.030	-0.017	-0.022	0.028	0.019	0.001	0.001
		1.00	0.019	-0.107	-0.024	-0.008	-0.001	0.031	0.028	0.027	0.027

This table shows the simulated raw biases in the logistic regression estimates of β_4 , namely $\text{BIAS}(\beta_4) = E(\hat{\beta}_4) - \beta_4$, for sample sizes of $n_0 = n_1 = 100$. Note that $\beta_4 = 0.0, 0.636$ and 0.892 for $\mu = 0.0, 0.25$ and 0.35 , respectively. The columns denote the eight methods of analysis described in the text. The rows denote the means of the cases (μ), the shapes of the trends (τ , step or triangular/slope), the magnitudes of the trends (η), and the locations of the jump for step trends and the peak for triangular/slope trends (ζ). Note that the same simulated $M = 10\,000$ data sets were used to calculate the biases for all entries in the same row.

comparable to those for using the uncontaminated data. Note also that the power decreases as the data-fraction increases ($f = 1/3, 1/2, 2/3$), particularly for step trends and for trends with larger magnitudes. Additional simulations (not shown) show similar patterns in the rejection rates for the LRT for the inclusion of the quartile categories at the 5 per cent significance level, $G^2 > \chi_{3,0.95}^2 = 7.81$, for the same sample sizes.

Table VI shows the simulated raw biases in the estimation of β_4 for the same methods of analysis and trends as in the previous tables with sample sizes of (a) $n_0 = n_1 = 100$. These

biases should be interpreted in comparison with the true values $\beta_4 = 0.0, 0.636$ and 0.892 for $\mu = 0.0, 0.25$ and 0.35 , respectively. When the cases and controls have the same means ($\mu = 0$), we observe that all eight methods described above give near zero biases. When the mean of the cases is greater than that of the controls ($\mu = 0.25, 0.35$), using the uncontaminated data $\{x_i^{\text{true}}\}$ gives negligible positive biases. By comparison, using the trend-contaminated data $\{x_i^{\text{trend}}\}$ gives substantial negative biases, which reflects the fact that the estimates of $\{\beta_j\}$ are attenuated towards zero. These biases are greater in absolute value for step trends (especially with jumps at $\zeta = 0.5$) and for trends with larger magnitudes. Conversely, the lowess adjusted data $\{x_i^{\text{lowc}}\}$ and $\{x_i^{\text{lowb}}\}$ gives relatively small biases of either sign. Additional simulations (not shown) for sample sizes of (b) $n_0 = n_1 = 200$ and (c) $n_0 = 200$ and $n_1 = 100$ show that the absolute biases in the estimation of β_4 using the uncontaminated data were proportionately smaller than those for (a) $n_0 = n_1 = 100$, whereas the absolute biases using the trend-contaminated data and the lowess adjusted data were comparable for all sample sizes.

These simulations show that it may be valuable to examine carefully the nature of systematic trends over time in continuous predictor data, especially when the cases and controls may have a common trend. We recommend that one estimate the common trend by lowess methods using both the cases and controls combined whenever the numbers of cases and controls are (nearly) balanced on each day of analysis. We also generally recommend that one choose a smaller data-fraction for step trends and a larger data-fraction for triangular trends, although one should experiment with several values to determine the impact of this parameter on the process of trend estimation and removal.

5. DISCUSSION

The analysis of the case-control study presented in this paper involved the estimation and removal of a systematic trend over time from the predictor variable for sphinganine (Sa), a biomarker of fungal toxin exposure. The ln-Sa measurements for the cases and controls both had trends of approximately the same triangular shape and location. The underlying common trend was assumed to depend on the day of HPLC analysis and not on the true Sa levels themselves. We employed lowess methods to estimate and then remove this common trend on the logarithmic scale, using both the cases and controls combined with a data-fraction of $f = 2/3$. We then categorized the lowess adjusted data by the empirical quartiles of the adjusted controls. In turn, we constructed a logistic regression model to relate the risk of oesophageal cancer to the quartile categories of the lowess adjusted ln-Sa measurements, along with covariates for age, gender, smoking and drinking. We concluded that sphinganine is not a useful predictor of the risk of oesophageal cancer. It is also reassuring that another sphingolipid examined in the main study, sphingosine, was measured without trend and was not a useful predictor of the risk of oesophageal cancer either [4].

In addition, we designed a simulation study to validate the use of lowess methods to estimate and then remove systematic trends over time from continuous predictor data when the trends for the cases and controls have a common shape. We simulated data for hypothetical case-control studies of various sample sizes, with various differences between the means of the cases and controls, and with various shapes of contaminating trends over time. We next used lowess methods with various data-fractions to estimate the trend using the controls only and using both the cases and controls combined. These simulations show that

performing logistic regression with the trend-contaminated data tends to give attenuated parameter estimates and lower significance and power levels than using the uncontaminated data. Conversely, performing logistic regression with the lowess adjusted data computed using both the cases and controls combined with an appropriate data-fraction tends to give nearly unbiased parameter estimates, near nominal significance levels, and power comparable to that for the uncontaminated data.

We have discussed the use of lowess methods to estimate and remove common systematic trends over time from predictor variables in the context of case-control studies. More generally, these methods could be applied in either prospective or retrospective studies to adjust either predictor or response continuous variables contaminated by trends, before these variables are transformed and used in regression models with linear or non-linear link functions. Moreover, although the above simulations only examined the use of quartile category predictors without other covariates, the ideas developed here easily extend to situations where the model includes covariates (like the motivating example), provided that these covariates are independent of the trend. Furthermore, one may also wish to explore the use of more advanced versions of lowess to estimate systematic trends semi-parametrically using other covariates in addition to time. These potential applications collectively make lowess methods powerful tools for removing systematic trends from contaminated laboratory measurements, especially when standards or quality control samples are unusable or unavailable for this purpose. Finally, if in addition to a systematic trend over time, there exist other odd features in the data, such as gaps or outliers, one must carefully consider the impact these features have on the process of trend estimation and removal.

In terms of the broader methodological picture, careful consideration should be given to statistical methods that employ empirical quantile categories rather than categories defined by preset cutpoints. These empirically-based methods may occur in the context of regression, as in the above example, or association, such as contingency tables with categories defined by the empirical quantiles of the marginal data [9, 10]. These methods tend to be robust to many types of measurement error, especially when greater confidence exists in the relative ranks of the measurements rather than in their absolute values. However, these methods may require special variance estimation techniques and may entail a significant loss of power, and thus the advantages and disadvantages of their use in any particular application should be prudently evaluated.

The first author, Dr C. B. Borkowf, can provide sample computer code written in the S-Plus 2000 programming language to perform the lowess calculations described in this paper.

ACKNOWLEDGEMENTS

The authors wish to thank Sanford M. Dawsey, Philip R. Taylor and Lisa M. McShane for helpful discussions and comments on this manuscript.

REFERENCES

1. Breslow N, Day NE. *Statistical Methods in Cancer Research. Volume I – The Analysis of Case-Control Studies*. IARC Scientific Publications: Lyon, 1980.
2. Blot WJ, Li J-Y, Taylor PR, Guo W, Dawsey S, Wang G-Q, Yang CS, Zheng S-F, Gail M, Li G-Y, Yu Y, Liu B-Q, Tangrea J, Sun Y-H, Liu F, Fraumeni JF Jr., Zhang Y-H, Li B. Nutrition intervention trials in Linxian, China: Supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population. *Journal of the National Cancer Institute* 1993; **85**(18): 1483–1492.

3. Li B, Taylor PR, Li J-Y, Dawsey SM, Wang W, Tangrea JA, Liu B-Q, Ershow AG, Zheng S-F, Fraumeni JF Jr, Yang Q, Yu Y, Sun Y, Li G, Zhang D, Greenwald P, Lian G-T, Yang CS, Blot WJ. Linxian nutrition intervention trials: Design, methods, participant characteristics, and compliance. *Annals of Epidemiology* 1993; **3**(6):577–585.
4. Abnet CC, Borkowf CB, Qiao Y-L, Albert PS, Wang E, Merrill AH Jr, Mark SD, Dong Z-W, Taylor PR, Dawsey SM. Sphingolipids as biomarkers of fumonisin exposure and risk of esophageal squamous cell carcinoma. *Cancer Causes and Control* 2001; **12**(9):821–828.
5. Merrill AH Jr, Wang E, Vales TR, Smith ER, Schroeder JJ, Menaldino DS, Alexander C, Crane HM, Xia J, Liotta DC, Meredith FI, Riley RT. Fumonisin toxicity and sphingolipid biosynthesis. *Advances in Experimental Medicine and Biology* 1996; **392**:297–306.
6. Turner PC, Nikiema P, Wild CP. Fumonisin contamination of food: Progress in development of biomarkers to better assess human health risks. *Mutation Research* 1999; **443**(1–2):81–93.
7. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 1979; **74**(368):829–836.
8. MathSoft Inc., Data Analysis Products Division. *S-Plus 2000 Programmer's Guide*. MathSoft Inc.: Seattle, Washington, 1999.
9. Borkowf CB, Gail MH, Carroll RJ, Gill RD. Analyzing bivariate continuous data grouped into categories defined by empirical quantiles of marginal distributions. *Biometrics* 1997; **53**(3):1054–1069.
10. Borkowf CB, Gail MH. On measures of agreement calculated from contingency tables with categories defined by the empirical quantiles of the marginal distributions. *American Journal of Epidemiology* 1997; **146**(6):520–526.