STATISTICS IN MEDICINE, VOL. 13, 343–355 (1994)

APPLICATION OF SAMPLE SURVEY METHODS FOR MODELLING RATIOS TO INCIDENCE DENSITIES

LISA M. LAVANGE

Centre for Medical Statistics, Research Triangle Institute, P.O. Box 12194, Research Triangle Park, NC 27709, U.S.A.

LYNETTE L. KEYES

Frank Porter Graham Child Development Center, University of North Carolina, Chapel Hill, NC 27599, U.S.A.

GARY G. KOCH

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, U.S.A.

AND

PETER A. MARGOLIS

Department of Pediatrics, University of North Carolina, Chapel Hill, NC 27599, U.S.A.

SUMMARY

We describe ratio estimation methods for multivariately analysing incidence densities from prospective epidemiologic studies. Commonly used in survey data analysis, these ratio methods require minimal distributional assumptions and take into account the random variability in the at-risk periods. We illustrate their application with data from a study of lower respiratory illness (LRI) in children during the first year of life. One question of interest is whether children with passive exposure to tobacco smoke have a higher rate of LRI, on average, than those with no exposure and in a setting where age of child and season are taken into account. A second question is whether the relationship persists after adjusting for background variables such as family's socioeconomic status, crowding in the home, race, and type of feeding. The basic strategy consists of a two-step process in which we first estimate subgroup-specific incidence densities and their covariance matrix via a first-order Taylor series approximation. These estimates are used to test for differences in marginal rates of LRI between children exposed to tobacco smoke and those not exposed. We then fit a log-linear model to the estimated ratios in order to test for significant covariate effects. The ability to produce direct estimates of adjusted incidence density ratios for risk factors of interest is an important advantage of this approach. For comparison purposes and to address the limitations of the ratio method with respect to the number of covariates that can be controlled simultaneously, we consider survey logistic regression methods for the example data as well as logistic and Poisson regression models fitted via generalized estimating equation methods. Although the analysis strategy is illustrated with illness data from an epidemiologic study, the context of application is broader and includes, for example, data on adverse events from a clinical trial.

1. INTRODUCTION

Prospective epidemiologic studies often involve the measurement of new occurrences of a disorder or an unfavourable event for the purpose of comparing incidence across risk groups. Incidence densities are calculated as the ratio of the number of events to the time at risk for subjects in a particular group and are typically the basis for subgroup comparisons. Our purpose

0277–6715/94/040343–13$11.50

Received June 1992
Revised June 1993

here is to consider the use of sample survey methods of ratio estimation in estimating incidence densities and modelling their variation. We apply these methods to child illness data and compare the results to alternative modelling strategies.

Since incidence densities are ratios of the sums of two random variables, ratio estimation methods provide a direct approach to analysing disease incidence. These methods offer a convenient way of constructing confidence intervals and hypothesis tests about the estimated densities without assumptions about the specific nature of the underlying distributions of the variables. All that is required for valid inference are adequate sample sizes such that the sums (or averages) comprising the numerators and denominators of the ratios are approximately normal for the risk groups of interest. Ratio estimation methods have historically been used in the context of sample surveys,[1] where it is often of interest to estimate rates or proportions (for example, numbers of physician visits per person, proportion of health care expenditures covered by insurance, proportion with a regular dentist, etc.) for population subgroups defined by a cross-classification of survey variables, and where large sample sizes are available to support the analysis. The methods are easy to understand and readily available via commercial software packages, and provide flexibility for a variety of analytical settings. We suggest that ratio methods may be useful for epidemiological studies in which quantities such as incidence of an event during a follow-up period are the outcomes of interest.

Our investigation stems from involvement in a follow-up study of lower respiratory illness (LRI) in children during the first year of life. Data from this study are characterized by three features that affect the analysis strategy: (1) repeat occurrences for a given subject, (2) at-risk periods that are subject to random variability, and (3) covariates that change over the course of follow-up. We apply ratio estimation procedures in estimating the incidence of LRI in an attempt to account for each of these features. The primary goal of the analysis was to determine if differences exist in the incidence of LRI due to passive exposure to tobacco smoke, and if so, whether the differences could be explained by other covariates. Ratio estimation methods provide a direct means for examining the hypothesized differences in incidence and assessing confounding. The ability to produce adjusted incidence density ratios that one can compare with their unadjusted counterparts is an important feature of the proposed approach.

A more traditional approach to estimating incidence densities is based upon the assumption of a uniform or Poisson distribution for the event occurrences within a follow-up interval. Methods are available for the construction of confidence intervals about the density estimates.[2, 3] Under the assumption of a Poisson process, one can fit log-linear or approximate logistic regression models to compare average incidence rates across groups while controlling for other risk factors or confounding variables.[4] If the phenomenon under consideration is one that can occur repeatedly for a given subject, then methods allowing for within-person correlations can be applied.[5, 6] Under the assumption that an individual's illness episodes follow a Poisson process with parameters unique to that individual, recent methodology for subject-specific regression models provide another framework for the analysis.[7]

As previously stated, the application of ratio methods avoids the necessity of specifying the exact nature of the underlying variable distributions and, at the same time, provides a convenient method for producing adjusted density ratios. Moreover, they are applicable to situations with possibly complicated sampling procedures for subjects (for example, unequal selection probabilities, clustering at two or more stages, and/or multi-way stratification, as in national health surveys). However, the more traditional modelling strategies are more flexible in terms of the number and type of covariates that can be accommodated. With a large number of covariates, the asymptotic normality assumption for subgroup-specific estimates required by the ratio approach may not be appropriate. Consequently, we illustrate the use of survey logistic regression methods

as a straightforward way of producing adjusted odds ratios while accounting for additional potential confounding variables. Like the ratio method, minimal assumptions are required. As a point of comparison, we also present for the child illness data two applications of the generalized estimating equation (GEE) methodology,[5, 6] repeated logistic and Poisson regression.

The motivation for this research, including a description of the study data, is presented in Section 2. The proposed approach for analysing incidence densities is presented in Section 3, followed by a brief description of survey logistic regression models and the GEE applications. Results of the data analysis are given in Section 4. We conclude with a discussion of the advantages and disadvantages of the ratio methods in Section 5.

2. MOTIVATION

The motivation for this discussion arose from an investigation of the relationship between passive exposure to tobacco smoke and LRI in children. The literature indicates substantial evidence that passive smoking is related to respiratory illness.[8–10] Our objective was to compare the incidence of LRI among children with and without passive smoke exposure and to determine if any observed differences could be explained by other potentially confounding variables. Data from a community-based cohort study of respiratory illness during the first year of life in central North Carolina were used to examine the hypothesis. This large study was designed to assess the effect of an intervention aimed at reducing infants' exposure to tobacco smoke and, thereby, their rate of LRI in the first year of life.[11, 12] Data from the control arm of this experiment formed the basis for analyses reported here.

The study population consisted of infants born in two counties in central North Carolina between 1986 and 1988. Data about potential risk factors for LRI were collected during home interviews at 3 weeks, 7 months, and 12 months of age. Symptoms of acute LRI were collected via telephone interview every 2 weeks. An LRI event was defined as the parents' report of the presence of cough, wheezing, and rattling in the chest. In the control group, 294 respondents had telephones in their homes and were therefore eligible for the telephone interviews, and completed the first and third home interviews.

In an analysis of the relationship between socioeconomic status (SES) and chronic respiratory symptoms measured at 12 months, we identified a set of risk factors of respiratory illness that were of interest here.[13] These included particularly the time-dependent factors of child's age and respiratory season (of the LRI event occurrence), defined as spring/summer versus fall/winter. Relevant background factors included race, breast feeding, and crowding in the home. For analyses reported here, passive smoke exposure refers to exposure measured at the first home interview, classified as some versus none. The infant's SES was classified into two levels according to the highest level of education achieved by the head of household; low/mid SES corresponded to less than or equal to a high school education and high SES to more than a high school education. Crowding was classified as fewer than 0·5 persons per room versus greater than or equal to 0·5 persons per room.

The analysis of LRI events was complicated owing to the data collection strategy. As mentioned above, each family was called every two weeks for a year, and asked about the presence of LRI symptoms since the last call. If the symptoms indicated that a new episode had occurred, it was assumed that the child became ill midway through the reference period and was at risk for only the first week of the reference period. If the child was still ill from a previous episode, then the child was considered not to be at risk for a new episode for the entire reference period. Incidence densities were then computed as the ratio of the number of new episodes divided by the time at risk, with these definitions in force.

It is not uncommon for a child to experience several episodes of LRI during the first year of life.[14] A child prone to repeated occurrences will affect the incidence density both by adding a larger number of events to the numerator and by adding a smaller than average number of weeks at risk to the denominator, since the child still suffering from an episode is not considered at risk for a new episode. Incidence densities estimated from these data are, therefore, ratios of random variables. Sample survey methodology for ratios[1] provided a way of computing variances and testing hypotheses about the densities consistent with realistic assumptions about the data (that is, adequate sample size to support approximate normality for the ratios).

For comparison purposes, we also considered model-based approaches to the data analysis. The first consisted of managing the two-week reports of LRIs as sequences of correlated binary outcomes with missing data and applying survey logistic regression methods to estimate the expected odds ratio with respect to passive smoke exposure, controlling for other potential confounding variables. Logistic models were also fitted by applying the GEE methodology to the two-week report data and specifying the correlational structures. A third strategy was to use GEE methods for correlated Poisson outcomes and directly estimate the risk by exposure group, controlling for different lengths of at risk periods as a fixed covariate in the model. The results of the three modelling strategies were compared with the ratio methods for the example data.

3. METHODS

The basic approach consists of first estimating incidence densities for each cell of a cross-classification of the explanatory variables, that is passive smoke exposure and other confounders. The corresponding estimated covariance matrix of the cell-specific ratio estimates is computed via a first-order Taylor series approximation of the deviations of the estimates from their expected values. This approximation for large samples is well known[15] and has seen considerable application in the field of survey data analysis.[16, 1] We then fitted linear models to the logarithms of the ratio estimates by applying weighted least squares methods[17, 18] in order to assess the simultaneous effects of the explanatory variables.

Let $\lambda = \mu_y/\mu_x$ denote the ratio of two quantities in the target population about which inference is desired. For our application, μ_y corresponds to the population average number of illness episodes per person and μ_x corresponds to the population average time at risk per person. We are interested in estimating λ with a random sample of n units selected under some specified design. The estimator for λ has the form $R = y/x$ where $x = \Sigma w_i x_i$ and $y = \Sigma w_i y_i$, with w_i being the reciprocal of the probability of selection for the ith unit in the sample and x_i and y_i being the time at risk and number of illness episodes for the ith unit, respectively. For a simple random sample where each unit is selected with equal probability ($\pi = 1/w$), $\bar{x} = x/wn$ and $\bar{y} = y/wn$ are sample-based estimators of the population mean parameters μ_x and μ_y respectively, and $R = (y/x) = (\bar{y}/\bar{x})$. A variance estimator for R can be computed by noting that R is a non-linear function of two linear sample statistics and can be expanded via a first-order Taylor series about $\mu_x = E(x)$ and $\mu_y = E(y)$ as follows:

$$F_L(x, y) = F(\mu_x, \mu_y) + \partial F_x\,(\mu_x, \mu_y)\,(x - \mu_x) + \partial F_y\,(\mu_x, \mu_y)\,(y - \mu_y). \tag{1}$$

Noting that the variable portion in the expansion is given by

$$z = (\partial F_x)x + (\partial F_y)y, \tag{2}$$

the variance of $F(x, y)$ can be approximated by $\text{var}(z)$.[16] Although primarily used in survey settings where the variance is computed according to a complex sampling scheme, this approximation is applicable under the assumption of simple random sampling from a population, where

expectation throughout refers to expectation under repeated sampling. Consequently, the resulting variance estimators do not involve any formal distributional assumptions about either x or y.

For ratio estimators, we have

$$\partial F_x = -y/x^2, \qquad \partial F_y = 1/x. \tag{3}$$

For the ith sampling unit, define

$$z_i = w(y_i - Rx_i)/x. \tag{4}$$

Then under simple random sampling, we can approximate $\text{var}(R)$ by

$$\begin{aligned} \text{var}(z) &= n\sum_{i=1}^{n} (z_i - \bar{z})^2/(n-1) \\ &= \sum_{i=1}^{n} (y_i - Rx_i)^2/n(n-1)\bar{x}^2 \end{aligned} \tag{5}$$

since $w/x = 1/(n\bar{x})$. Other studies with more complex sampling designs can be accommodated by applying the appropriate variance formula for the estimate of a population total under the specific sampling design in computing $\text{var}(z)$.

For the case of repeated measurements on a subject, let i denote subject and $j = 1, \ldots, m_i$ index the measurements. Then assuming that n subjects were randomly selected with replacement from the target population, $\text{var}(z)$ can be approximated by the between-subject mean square error, substituting

$$z_i = \sum_{j=1}^{m_i} z_{ij}$$

into (5).[1] Here, z_{ij} is the linearized value, obtained by applying (4) to the jth observation on the ith subject, and y_{ij} and x_{ij} are the number of events and time at risk, respectively. Recall that x_{ij} is a function of previous illness episodes and will vary from observation to observation.

Note that no assumptions are required concerning the structure of the correlations among repeated observations on a subject in computing this variance estimator. The only assumption made thus far is that subjects were selected with replacement according to a probability sampling scheme. The selection of measurements per person need not be specified. Under repeated sampling from the same population, this estimator is consistent for $\text{var}(R)$, relative to the sample size n for subjects in the study being sufficiently large.

Let $g = 1, \ldots, G$ index the levels of a risk factor or combination of risk factors for which incidence density ratios are to be estimated. Then $R(g)$, $z_i(g)$, and $\text{var}[z(g)]$ can be defined by summing over subjects or observations with membership in the gth subgroup in the above formulas. In our application, two risk factors are defined at the reporting period level, respiratory season and age of child. Since a subject can contribute observations to more than one level of these risk factors, any test of their effects must take into account the resulting correlations among the cell-specific ratios. Then the covariance between ratios corresponding to levels g and g' is given by

$$\text{cov}(z(g), z(g')) = n\sum_{i=1}^{n} (z_i(g) - \bar{z}(g))\,(z_i(g') - \bar{z}(g'))/(n-1). \tag{6}$$

Let $\mathbf{R}$ denote the $(G \times 1)$ vector of estimated incidence densities and $\mathbf{V}(\mathbf{R})$ denote the Taylor series approximation for the covariance matrix of $\mathbf{R}$, that is, v_{gg} is computed from (5) and

$v_{gg'}$ from (6), $g, g' = 1, \ldots, G$. Then $\mathbf{V(R)}$ is a consistent estimator of the population covariance matrix. Let $\mathbf{C}$ denote a $(c \times G)$ contrast matrix. Then the hypothesis H_0: $\mathbf{C}\lambda = 0$ can be tested via the Wald test statistic

$$\mathbf{W} = (\mathbf{CR})' (\mathbf{CV(R)C}')^{-1} \mathbf{CR}, \tag{7}$$

which is approximately distributed as a chi-square statistic with c degrees of freedom. When the number of sample units or subjects is moderate rather than large relative to c, $\mathbf{W}$ has been shown to be too liberal,[19, 20, 21] and a transformed F statistic is recommended. Two candidates are the Hotelling's T^2-type statistic,[22] which has been shown to perform well in small samples,[23] and the Satterthwaite adjusted χ^2 statistic introduced by Rao and Scott[24] for survey data analysis.

The SUDAAN software package[25] was employed for analyses reported here. This package provides the capability for estimating ratios and testing contrasts among them. All three test statistics mentioned above are computed for tests of hypotheses. In addition, the vector of estimated ratios, $\mathbf{R}$, and corresponding covariance matrix, $\mathbf{V(R)}$, can be output for secondary data analysis. In particular, categorical data analysis methods[17, 18] can be applied to fit linear models to $\mathbf{R}$.

For this application, $\mathbf{R}$ corresponds to the vector of cell-specific ratios from a cross-classification of the risk factors of interest. We first estimated $\mathbf{R}$ and $\mathbf{V(R)}$ via PROC RATIO in SUDAAN. We then transformed the estimates to their natural logarithms to provide additional covariance matrix stability (that is, to eliminate the tendency for the variances to be proportional to the squares of the ratios) as well as to enable the use of multiplicative (log-linear) models. We then input the transformed vector $\mathbf{log_e(R)}$ and associated covariance matrix, $\mathbf{D_R^{-1}V(R)D_R^{-1}}$, where $\mathbf{D_R}$ is a diagonal matrix with $\mathbf{R}$ on the diagonal, to SAS PROC CATMOD. Linear models were fitted to the ratios via weighted least squares in order to assess the main effects and interactions among the various risk factors.

In order to apply these methods, we assumed that the averages comprising the numerators and denominators were approximately distributed as multivariate normal random variables. Subgroup-specific sample sizes should be large enough to support this assumption. In the presence of many risk factors, the ratio approach may not be feasible owing to small sample sizes resulting from the complete cross-classification of the factors. For such cases, model-based methods may be more appropriate. We considered logistic and Poisson regression models as tools for approximating the adjusted risk of LRI. For the first, we modelled the probability of contracting an LRI in a two-week reference period as a function of the selected risk factors. Logistic models were fitted to the repeated binary outcomes on a subject. The predicted probabilities for each exposure group were then inflated to approximate rates per person year at risk. If, for a given two-week period, a child was still ill and therefore not at risk for a new LRI, we set the outcome to missing. Because no knowledge was to be gained concerning a new LRI, we assumed that these missing data were non-informative.[26]

Two methods were used to fit the logistic models. The first corresponds to applying survey logistic regression methods to the repeated observations on a subject. With this method, a weighted analogue of the usual likelihood equations is solved for the model parameters, and the variances are then adjusted for the sample design.[27] To apply this method to the example data, we assumed that each subject corresponded to a cluster, and variance formulas appropriate for single-stage cluster sampling with replacement were applied. Survey logistic regression methods are valid in the presence of within-subject correlations but may not be efficient, as the correlational structure is not used to refine the parameter estimates. Second, we fitted two GEE logistic regression models[5, 6] assuming an independent working correlation matrix and equal correlations, respectively. Although developed from different underlying principles, the GEE model with

independence yields similar results to the survey logistic model. The major difference is that the GEE method uses the number of clusters as a divisor in computing variances, while the survey logistic regression method uses the number of clusters minus one.[28, 29]

For the Poisson model, we first accumulated LRI events and time at risk for each person by each cell in the complete cross-classification of the risk factors. Recall that each subject contributed to more than one cell owing to the time-varying covariates (age and season). We then fitted the following Poisson regression model to the log of the LRI events:

$$E[\log(y_{ij})] = \log(x_{ij}) + \sum_{k=1}^{q} a_{ijk}\beta_k, \tag{8}$$

where y_{ij} denotes the LRI count, x_{ij} the time at risk, and $\mathbf{a}_{ij} = (a_{ij1}, \ldots, a_{ijq})'$ the set of predictors for the ith subject and the jth observation (within the ith subject). Note that $\mathbf{a}_{ij}$ would account for whatever explanatory variables were relevant to the group g for which y_{ij} and x_{ij} were representative. GEE methods were used to fit the models assuming equal correlations among repeated counts for a subject and non-informatively missing data for periods in which the subject was ill and not eligible for a new LRI.

The SUDAAN LOGISTIC procedure was used for the survey logistic models. Software obtained from the authors[5, 6] was used for the GEE models.

4. RESULTS

We present results from our investigation of potential risk factors of LRI in order to illustrate the various methods of analysis. A more comprehensive investigation of the relationship between passive smoke exposure and LRI is currently under way by the project investigators. The overall incidence of LRI for the study population was 0·95 (SE = 0·09). The observed incidence densities for subgroups defined by selected risk factors are presented in Table I. Also included are the numbers of subjects, subjects with an LRI and person-years at risk. Variance estimates were computed via the Taylor series approximation methods described in Section 3. Lower and upper 95 per cent approximate confidence limits and p-values for a test of the hypothesis of any difference among the levels of a risk factor are presented. These results were computed using the RATIO procedure in SUDAAN. The LRI rate is considerably greater among children with passive exposure to smoke (1·24 versus 0·65 per person year, $p = 0{\cdot}001$). In addition, children are at significantly higher risk during the fall/winter respiratory season, at 4–6 or > 6 months of age (compared with 0–3 months), in families of lower SES and in families with more crowding ($\geqslant 0{\cdot}5$ versus $< 0{\cdot}5$ persons per room).

The second stage of analysis involved modelling the incidence densities. One question of interest was whether the child characteristics, age and respiratory season, explained the apparent passive smoke exposure effect. These variables were established as significant risk factors of LRI in the univariate analysis (Table I) and were time dependent, thereby providing an illustration of the ability of the proposed methodology to handle correlations among repeated events for a given individual. The cell-specific sample sizes in the passive smoking by age by season cross-classification were adequate for supporting the large-sample approximations required by this method. All cells contained at least 30 subjects and 80 per cent of the cells contained at least 5 subjects with LRI episodes. The remaining cells contained at least 3 subjects with LRI episodes. Results from a study of the small-sample properties of chi-square goodness of fit tests in log-linear models[30] suggest the use of these or similar sample size requirements. The results obtained via the ratio methods were confirmed by those obtained via other analytic methods, further indicating

Table I. Incidence densities for risk factors of LRI

Factor	Level	Children	LRIs	Person-years at risk	ID	var*	95% confidence interval	*p*-value†
Passive smoking	Exposed	158	130	104·75	1·24	0·024	[0·94, 1·54]	0·001
	Unexposed	136	67	103·65	0·65	0·009	[0·46, 0·83]	
Season	Spring/summer	276	81	109·46	0·74	0·009	[0·55, 0·93]	0·001
	Fall/winter	286	116	98·95	1·17	0·017	[0·92, 1·43]	
Age	0–3 months	293	35	62·42	0·56	0·010	[0·37, 0·75]	0·001
	4–6 months	267	65	48·48	1·34	0·035	[0·98, 1·71]	
	> 6 months	261	97	97·50	0·99	0·016	[0·75, 1·24]	
SES	⩽ HS	138	116	87·77	1·32	0·026	[1·01, 1·64]	< 0·001
	> HS	156	81	120·64	0·67	0·010	[0·47, 0·87]	
Crowding (persons per room)	< 0·5	161	79	119·28	0·66	0·011	[0·46, 0·86]	< 0·001
	⩾ 0·5	133	118	89·13	1·32	0·025	[1·01, 1·63]	
Feeding	Bottle	154	112	102·45	1·09	0·018	[0·83, 1·36]	0·11
	Breast	139	85	105·91	0·80	0·015	[0·56, 1·05]	
Race	White	218	145	161·82	0·90	0·011	[0·69, 1·10]	0·32
	Black	76	52	46·58	1·12	0·039	[0·73, 1·51]	
Overall		294	197	208·4	0·95	0·008	[0·77, 1·13]	

* Taylor series variance estimate
† Probability level for a test of differences among the levels of the risk factor based on the Wald chi-square statistic

Table II. Model results for log transformed incidence densities

Effect	Beta	SE	*p*-value	IDR	95% CI
Intercept	− 1·19	0·23	< 0·001	—	
Passive smoking	0·71	0·19	< 0·001	2·04	(1·40, 2·97)
Season	0·47	0·14	0·001	1·60	(1·21, 2·11)
Age: 4–6 months	0·83	0·21	< 0·001	2·30	(1·53, 3·47)
> 6 months	0·56	0·21	0·006	1·76	(1·18, 2·63)
Goodness of fit chi-square (7 d.f.) = 7·48			$p = 0·38$		
Predicted incidence density (95% CI):					
Exposed	1·25	(0·99, 1·59)			
Unexposed	0·62	(0·46, 0·83)			

the adequacy of the cell-specific sample sizes. Owing to the collinearity present among the passive smoking, crowding, and SES measures, adding either crowding or SES to the cross-classification resulted in cells with too few subjects by these minimal criteria. Consequently, we examined further potential confounding effects via the various modelling strategies.

Estimated ratios and their covariance matrix were computed via SUDAAN PROC RATIO for the cross-classification of passive smoking, age, and respiratory season. These estimates were then input to SAS PROC CATMOD for the modelling. Goodness of fit tests were performed to determine a parsimonious set of predictors. No interactions among the predictors were found to be significant. The final model results for the ratio analysis are presented in Table II. All

Table III. Model results for repeated logistic regression

Effect	Beta	SE	p-value	OR	95% CI
Intercept	−4·85	0·24	< 0·001	—	
Passive smoking	0·45	0·22	0·04	1·58	(1·03, 2·40)
Season	0·44	0·14	0·002	1·55	(1·18, 2·05)
Age: 4–6 months	0·79	0·22	< 0·001	2·20	(1·43, 3·38)
> 6 months	0·62	0·20	0·002	1·86	(1·25, 2·77)
Crowding	0·48	0·24	0·04	1·62	(1·02, 2·57)
SES	0·33	0·26	0·21	1·39	(0·84, 2·29)
Predicted no. of LRI per year (95% CI):					
Exposed	0·92	(0·70, 1·21)			
Unexposed	0·59	(0·45, 0·78)			

predictors remained significant, including passive smoking ($p < 0·001$). The model adjusted incidence density ratio was similar to the unadjusted ratio (2·04 versus 1·92 for some versus no exposure), indicating no apparent confounding of exposure effect by age and season. Approximate 95 per cent confidence intervals for the two ratios were (1·40, 2·97) and (1·32, 2·79) respectively.

In order to assess the effects of additional covariates, we applied the modelling strategies described in Section 3. Effects for age, season, crowding, and SES were included in the models as potential confounders based on the evidence in Table I. Table III gives results for the model fit to the dichotomous outcome, indicating an LRI event in a two-week period, using survey logistic regression methods. All of the covariates were significant except SES ($p = 0·21$). Apparently, crowding and SES were measuring similar characteristics, and of the two, crowding appeared to be a better predictor. The predicted probabilities of an LRI in the two-week reference period were computed by exposure group, assuming uniform distributions for the other covariates in the model, and used to estimate incidence per person year at risk. The predicted densities were 0·92 and 0·59 illnesses per person year for exposed and unexposed subjects, respectively. The associated odds ratio, 1·58 with 95 per cent confidence interval of (1·03, 2·40), indicates evidence of some confounding when compared with the unadjusted or crude IDR (1·92 for illnesses per person year at risk). In order to verify that this difference was is fact due to confounding and not to a poor approximation of an incidence density ratio with an odds ratio, we fitted a repeated logistic model that included only passive smoking, age, and season. The adjusted odds ratio for exposure from this model was 1·97 with 95 per cent confidence interval of (1·35, 2·89). This estimate is almost identical to the adjusted incidence density ratio computed from the ratio analysis that included the same set of explanatory variables (Table II), indicating that the odds ratio provides a reasonable approximation for the incidence density ratio in this example.

The results for the GEE logistic regression model with independent working correlation matrix were identical to those presented in Table III for the survey logistic model. The results for the GEE logistic model assuming equal correlations across time are presented in Table IV. Modelling the correlation matrix resulted in slightly smaller estimated standard errors for this example. The conclusions are the same, with all covariates except SES remaining significant. The adjusted odds ratio for passive smoke exposure was 1·56, with 95 per cent confidence interval of (1·02, 2·38).

Results from the Poisson regression analysis are presented in Table V. For this model, we accumulated illness events and time at risk for each person in each age by season period. The GEE methodology was then applied to fit a log-linear model to the counts with time at risk entered as an offset variable. Equal correlations were assumed among the multiple event counts

Table IV. Model results for GEE logistic regression with equal correlations

Effect	Beta	SE	*p*-value	OR	95% CI
Intercept	− 4·82	0·24	< 0·001	—	
Passive smoking	0·44	0·22	0·04	1·56	(1·02, 2·38)
Season	0·44	0·14	0·002	1·55	(1·18, 2·05)
Age: 4–6 months	0·80	0·22	< 0·001	2·22	(1·45, 3·39)
> 6 months	0·63	0·20	0·002	1·87	(1·26, 2·77)
Crowding	0·47	0·23	0·04	1·60	(1·01, 2·52)
SES	0·32	0·26	0·21	1·38	(0·84, 2·28)
Predicted no. of LRI per year (95% CI):					
Exposed	0·93	(0·71, 1·23)			
Unexposed	0·61	(0·46, 0·80)			

Table V. Model results for Poisson regression with equal correlations

Effect	Beta	SE	*p*-value	RR	95% CI
Intercept	− 1·60	0·25	< 0·001	—	
Passive smoking	0·45	0·22	0·04	1·57	(1·02, 2·40)
Season	0·49	0·15	< 0·001	1·63	(1·22, 2·18)
Age: 4–6 months	0·93	0·22	< 0·001	2·54	(1·65, 3·90)
> 6 months	0·65	0·22	0·003	1·92	(1·26, 2·95)
Crowding	0·53	0·23	0·02	1·70	(1·08, 2·66)
SES	0·24	0·25	0·35	1·27	(0·77, 2·08)
Predicted rate of LRI per year (95% CI):					
Exposed	1·01	(0·47, 2·17)			
Unexposed	0·64	(0·34, 1·22)			

per person. Under the Poisson assumption of the mean–variance relationship, passive smoking remained significant after adjusting for the other predictors ($p = 0{\cdot}04$). The ratio of average LRI rates for exposed to unexposed groups was 1·57 with 95 per cent confidence interval of (1·02, 2·40) under this model, again indicating some confounding due to crowding and SES when compared with the unadjusted and adjusted incidence density ratios from the ratio analysis.

5. DISCUSSION

We have described a direct method for estimating and modelling incidence densities from a prospective study based on sample survey methodology that we believe has a number of advantages in this setting. Adjusted incidence density estimates produced via the proposed approach are directly comparable with the observed ratios, with the same definitions of time at risk in force. Goodness of fit tests can be generated, allowing for comparisons of various models fitted to the ratios. The methods require minimal distributional assumptions and are applicable to a wide variety of study designs, including complex sample surveys. The effects of unequal weighting, stratification, and/or cluster sampling frequently associated with large national surveys (such as the National Health and Nutrition Examination Survey, NHANES[18]) can be accommodated by incorporating the appropriate variance formula in equation (5).

The primary limitations of the ratio method are the requirements of large sample sizes and categorical explanatory variables. Variance estimates are based on the Taylor series approximation for ratios of random variables. These estimates are consistent for their population counterparts provided that the sample sizes upon which they are based are sufficiently large. For the example data, we were limited to examining three of the five potential covariates in one model owing to small sample sizes. Nevertheless, we would argue that the marginal rates are of interest in targeting children prone to respiratory illness, regardless of the causal pathways, and therefore the ratio method is potentially a very useful tool in applications of this type. The requirement of categorical predictors is often not a problem. The complex sample survey setting is typically characterized by very large samples for which valid estimates are available through the survey design for groups based on one or more questionnaire items that define explanatory variables. In such a setting, the application of the proposed methods is a natural course to take.

For sample sizes insufficient to support ratio analyses, several options are available. The model-based approaches we considered allow for continuous covariates and will accommodate more covariates than the ratio approach in that all interactions need not be entered in the model. The lack of a global goodness of fit test for comparing competing models is a drawback. The ability to produce adjusted incidence rates via ratio methods or Poisson regression models is somewhat more appealing than producing adjusted odds ratios from the logistic models, although for the example presented here, little difference was evident. In practice, the choice of a model-based approach versus the ratio method will probably be based on the need for modelling in general, that is, to adjust for selection bias or confounding effects due to an imperfect sampling scheme or small samples. In such settings, the sample size and number of potential confounders are likely to drive the decisions on analytic methods.

Although developed from different principles, the survey logistic regression model is for the most part identical to the GEE repeated logistic model with working independence. The advantage of the former is that commercial software is readily available and provides gains in computational efficiency when compared with the latter. For the example data set of 6115 two-week reference period records for 294 clusters (subjects), the survey logistic software took approximately one-quarter to one-fifth the time to run on a 486 workstation as the GEE software, run under SAS Version 6.04, for the same model. The advantage of the GEE approach is the ability to incorporate the correlation structure and thereby gain some efficiency in the parameter estimates. For the example data, these gains were slight. Widths of approximate 95 per cent confidence intervals were comparable for all three models presented, ranging from 1·36 to 1·38. Interestingly, the confidence interval widths computed from the ratio analysis agreed with those from the repeated logistic model containing the limited set of three explanatory variables. The widths for the two approaches were 1·57 and 1·54, respectively. Gains in efficiency for the example data were realized owing to adjusting for additional covariates, and not to differences between the ratio and model-based methods, or to modelling of the correlation structure.

Both the ratio and the model-based approaches make assumptions about missed interview data, namely that these data are missing at random. The treatment of illness periods when a child is not at risk for a new LRI differed slightly, though the impact on the resultant estimates was the same. With the model-based approaches, these periods were considered to be non-informatively missing while with the ratio approach, no time at risk was added to the denominator of the incidence densities. All of the methods considered here are based on large-sample approximations. As described earlier, transformed F and adjusted χ^2 statistics have been proposed for survey data analysis when the number of sampling units is moderate (we employed the Wald chi-square test statistics for results presented here owing to adequate numbers of subjects). Recent work[28] suggests similar adjustments to the test statistics computed for GEE models.

We have limited our scope to a subset of possible models for analysing incidence data. Likelihood-based methods proposed by Thall for modelling Poisson event rates assuming a random subject effect are also of interest in this context.[31] The methods presented here are potentially useful in analysing side effect or adverse event data from clinical trials. With randomization of subjects, the limitations of the ratio method with respect to the number of covariates for which control is possible should not be a drawback. Other methods that have been considered in this context include Markovian models for incidence data.[32] Although repeated occurrences are accommodated, it is not obvious how their approach could be extended to capture the random variability in at-risk periods. This may be of interest for future research.

To conclude, we feel that sample survey methods may be useful in epidemiological settings such as the example described here. These methods are easy to access, straightforward to apply, and produce easily interpretable results without very stringent assumptions. Although ideal for large survey settings, we feel that these methods may offer some advantages over more traditional approaches to modelling incidence density data in other applications as well.

ACKNOWLEDGEMENTS

We wish to acknowledge Professor Ronald Helms for his helpful comments concerning the treatment of missing data in the GEE models, and Dr. Robert Greenberg who kindly made the data available to us. This research was supported in part by grant no. 28895 from the National Heart, Lung, and Blood Institutes, National Institutes of Health.

REFERENCES

1. Cochran, W. G. *Sampling Techniques*, 3rd edn., Wiley, New York, 1977.
2. Kleinbaum, D. G., Kupper, L. L. and Morgenstern, H. *Epidemiologic Research: Principles and Quantitative Methods*. Van Nostrand Reinhold, New York, 1982.
3. Sempos, C. T. *Statistical Methods in Epidemiology*, Oxford University Press, New York, 1989.
4. Koch, G. G., Atkinson, S. S. and Stokes, M. E. 'Poisson regression', in Kotz, S., Johnson, N. L. and Reade, C. B. (eds.), *Encyclopedia of Statistical Science 7*, Wiley, New York, 1986, pp. 32–41.
5. Liang, K. L. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
6. Zeger, S. L. and Liang, K. L. 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, **42**, 121–130 (1986).
7. Zeger, S. L., Liang, K. L. and Albert, P. S. 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics*, **44**, 1049–1060 (1988).
8. Fergusson, D. M., Horwood, L. J. and Shannon, F. T. 'Parental smoking and respiratory illness in infancy', *Archives of Diseases of Childhood*, **55**, 352–361 (1980).
9. Ware, J. E., Jr., Dockery, D. W., Spiro, A., Speizer, F. E. and Ferris, B. G. 'Passive smoking, gas cooking, and respiratory health of children living in six cities', *American Review of Respiratory Disease*, **129**, 366–374 (1984).
10. Wright, A. L., Holberg, C., Martinez, F. D., Taussig, L. M. and Group Health Medical Associates. 'Relationship of parental smoking to wheezing and nonwheezing lower respiratory tract illness in infancy', *Journal of Pediatrics*, **118**, 207–214 (1991).
11. Greenberg, R. A., Bauman, K. E., Strecher, V. J., Keyes, L. L., Glover, L. H., Haley, N. J., Stedman, H. C. and Loda, F. A. 'Passive smoking in the first year of life', *American Journal of Public Health*, **81**, 850–853 (1991).
12. Greenberg, R. A., Strecher V. J., Bauman, K. E., Boat, B. W., Fowler, M. G., Keyes, L. L., Denny, F. W., Chapman, R. S., Stedman, H. C., LaVange, L. M., Glover, L. H., Haley, N. J. and Loda, F. A. 'Evaluation of a home-based intervention program to reduce infant passive smoking and lower respiratory illness', to appear in *Journal of Behavioral Medicine* (1993).
13. Margolis, P. A., Greenberg, R. A., Keyes, L. L., LaVange, L. M., Chapman, R. S., Denny, F. W., Bauman, K. E. and Boat, B. W. 'Lower respiratory illness in infants and low socioeconomic status', *American Journal of Public Health*, **82**, 1119–1126 (1992).