Community Dent Oral Epidemiol 2004; 32: 183–9

COMMUNITY DENTISTRY AND ORAL EPIDEMIOLOGY

The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status

J. D. Lewsey[1] and W. M. Thomson[2]

Departments of [1]Preventive and Social Medicine, and [2]Oral Sciences, School of Dentistry, The University of Otago, Dunedin, New Zealand

Lewsey JD, Thomson WM. The utility of the zero-inflated poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. Community Dent Oral Epidemiol 2004; 32: 183–9. © Blackwell Munksgaard, 2004

Abstract – *Objectives:* To examine the utility of the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) modelling approaches for modelling four sets of dental caries data from the same cohort study [with particular attention to the influence of childhood socioeconomic status (SES)]: cross-sectional data on the deciduous dentition at age 5 years; cross-sectional data on the permanent dentition at age 18 and 26 years; and longitudinal data on caries increment between ages 18 and 26 years. *Methods:* Data on dental caries occurrence at ages 5, 18 and 26 years were obtained from the Dunedin Multidisciplinary Health and Development Study (DMHDS). ZIP and ZINB models were fitted to the cross-sectional ($n = 745$) and longitudinal ($n = 809$) data sets using Stata (Intercooled Stata 7.0). The dependent variables for the three cross-sectional analyses were the DMFS indices at age 5, 18, and 26 years, and net DFS increment (NETDFS) was the dependent variable for the longitudinal analysis. *Results:* The empty ZIP model was a poor fit for all four data sets, whereas the empty ZINB model showed good fit; consequently both the cross-sectional and longitudinal analyses were conducted using ZINB modelling. Being in the high-SES group during childhood was associated with a greater probability of being caries-free by age 18 years, over and above that which would be expected from the negative binomial process. Low childhood SES also had the largest coefficient in the modelling of the negative binomial process, but at age 5 years, where the adjusted mean dmfs score in the low-SES group was 6.8 (compared with 4.7 and 2.9 in the medium- and high-SES groups, respectively). The substantial SES differences which existed at age 5 years (in the deciduous dentition) had reduced somewhat by age 18 years, and had widened again by age 26 years. In the longitudinal analysis, 'baseline' caries experience (age 18-year DMFS) was a predictor both of being an extra zero and of caries severity. *Conclusion:* This investigation of the utility of the zero-inflated approach for modelling both cross-sectional and longitudinal caries data has shown that ZIP/ZINB models can provide new insight into disease patterns. It is anticipated that they will become increasingly useful in epidemiological studies that use the DMF index as the outcome measure.

Key words: dental caries; modelling; socioeconomic status

W. M. Thomson, Department of Oral Sciences, School of Dentistry, The University of Otago, PO Box 647, Dunedin, New Zealand
Tel: +64 3 479 7116
e-mail: mthomson@gandalf.otago.ac.nz

Submitted 17 June 2003;
accepted 29 October 2003

Epidemiological studies of dental caries occurrence invariably use the DMF index (1), a simple count of the number of decayed, missing or filled teeth (or surfaces) which represents the cumulative severity of dental caries experience in an individual or a population. The index has well-documented shortcomings, and there have been recent calls for it to be replaced with newer outcome measures which better reflect (a) the presence and distribution of noncavitated lesions, and (b) the fact that caries presentation is a continuum rather than a 'present–absent' dichotomy (2). To date, however, there have been few signs of a wholesale shift from the DMF index, most probably because of its continuing usefulness and the need to be able to make historical comparisons. It would be valid to assume that the DMF index will be in use for some time yet.

Despite over six decades of use, the DMF index is associated with contentious statistical issues. Today, the typical DMF distribution is highly positively skewed and has a high proportion of zero scores (sometimes resulting in a bimodal distribution). By courtesy of the central limit theorem, between-group comparisons of mean DMF scores using the *t*-test are usually valid where there is sufficient statistical power. However, because most dental epidemiological studies are observational, regression methods are often utilized to control confounding (and thus use more than one independent variable). The assumptions of multiple linear regression (MLR) will be violated because of skewed residuals (the difference between the observed and predicted values) and the 'spike' of zero scores, even if transformation (3) of DMF scores is performed. In recent years, there has been a move away from MLR to consideration of the family of generalized linear models (GLMs), of which MLR is itself a member (4). In essence, the right-hand side of the equation is the same for all GLMs, but the dependent variable (the left-hand side of the equation) can be assumed to follow probability distributions other than the normal distribution (such as the binomial or Poisson distributions).

Unfortunately, even Poisson and negative binomial regression violate the assumption of normally distributed standardized residuals when there is a high proportion of zero scores and/or the DMF distribution is bimodal. Zero-inflated modelling has recently been suggested as an approach which gives a better fit to these types of count data. Two possible alternatives are the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) models (5). Böhning et al. (6) used ZIP modelling to examine changes in DMFT scores among children in a preventive dentistry trial in Brazil, and reported model predictions close to the observed DMFT distribution.

A feature of zero-inflated modelling is that the effects of covariates can be examined simultaneously in the extra zero and Poisson (or negative binomial) components of the model. For the DMF index, for example, the covariates which affect the probability of being caries-free (i.e. DMF $= 0$) over and above that expected from a usual Poisson (negative binomial) process are modelled simultaneously with the covariates that affect the average severity of caries experience. A more detailed description of zero-inflated modelling can be found in the Appendix.

The aim of this investigation was to examine the utility of the ZIP and ZINB approaches for modelling four sets of dental caries data from the same cohort study [with particular attention to the influence of childhood socioeconomic status (SES)]: cross-sectional data on the deciduous dentition at age 5 years; cross-sectional data on the permanent dentition at age 18 and age 26 years; and longitudinal data on caries increment between ages 18 and 26 years.

Methods

The sample

Data were obtained from the assessments at ages 5, 18 and 26 years in the Dunedin Multidisciplinary Health and Development Study (DMHDS), a longitudinal study of children born in Dunedin during 1972–73 (7). Perinatal data were obtained and the sample for the longitudinal study was defined at age 3 years. This initially comprised 1037 children assessed within a month of their third birthdays and again at ages 5, 7, 9, 11, 13, 15, 18, 21 years and, most recently, at 26 years, when 980 (96%) of the surviving 1019 study members were assessed. Barriers to study members' participation were minimized by the unit assuming the costs of participation (such as travel, lost wages, child care). The various assessments (e.g. oral health, mental health, physical health) are presented as standardized modules in counterbalanced order and each is conducted by a different examiner who is kept blind to all study data.

Ethical approval for the current study was obtained from the Otago Ethics Committee, and

informed consent was obtained from all participants.

Measures

An estimate of social class was obtained for each participant by using data collected on parental SES. Standard New Zealand occupation-based indices (8) were used; these employ a six-interval classification (where, e.g. a doctor scores '1' and a labourer scores '6'). The variable we used is the average of the highest SES level of either parent, assessed repeatedly at the study member's birth and at ages 3, 5, 7, 9, 11, 13 and 15 years (in order to capture cumulative exposure to SES while, at the same time, allowing for change of SES during childhood). The resulting scores were used to assign each individual to one of three childhood SES groups using predetermined thresholds: scores of 1 and 2 were allocated to the 'high-SES' group; those scoring 3 or 4 were allocated to the 'medium-SES' group; and the remainder (scores 5 or 6) were categorized as 'low SES'.

Residential fluoride exposure to age 5 years was computed as the percentage of the child's life spent in a fluoridated area. Similarly, residential fluoride exposure to ages 18 and 26 years was computed as the percentage of the years to each of those ages which were spent in a fluoridated area.

Use of dental services at age 26 years was determined by asking study members whether they usually visited the dentist for a check-up or because of a problem. Those who reported the latter were designated 'episodic users' of dentistry.

Dental examinations at each age were conducted using calibrated dental examiners. Diagnosis was based on clinical examinations, and no radiographs were taken. All questionable lesions were recorded as sound. Repeat examinations were not possible because of the logistical constraints imposed by the tightly scheduled assessment undergone by study members. Caries increment between ages 18 and 26 years was computed by comparing the status of each tooth surface at age 26 years with that at age 18 years. The 8-year net DFS increment was computed for each individual by subtracting the number of reversals from the caries increment.

Data analysis

ZIP and ZINB models were fitted to the cross-sectional ($n = 745$) and longitudinal ($n = 809$) data sets using Stata (Intercooled Stata 7.0; Stata Corporation, College Station TX, USA). The dependent variables for the three cross-sectional analyses were the DMFS indices at age 5 (dmfs5), 18 (DMFS18), and 26 years (DMFS26), and net DFS increment (NETDFS) was the dependent variable for the longitudinal analysis.

For SES, two dummy variables were created, SESLOW (low SES) and SESMED (medium SES); 'high SES' was the reference category. Dummy variables were created (for each of ages 5, 18 and 26 years) to dichotomize the person's residential fluoride exposure into less than half their life exposed (coded 1) and half or more their life exposed (coded 0). The dummy variables are called FEXP5 < 50%, FEXP18 < 50% and FEXP26 < 50% for ages 5, 18, and 26 years, respectively. The dummy variable FEMALE was created to represent gender.

The 'usual' versus 'episodic' use of dental services was included in the modelling process for the longitudinal analysis only (represented by the dummy variable EPISODIC). For this analysis, it was also necessary to include DMFS18 as an independent variable to control for 'baseline' caries experience. To assess the influence of residential fluoride exposure, that up to age 18 years (FEXP18 < 50%) was included to control for 'baseline' exposure, and fluoride exposure between 18 and 26 years was included as a binary variable (FEXP18_26) which represented 1 or more years of residence in a fluoridated area between ages 18 and 26 years.

A Wald test (9) was used to assess whether the model coefficients were statistically different from zero ($\alpha = 0.05$).

Results

Cross-sectional results

The first step of model formulation is to find a suitable probability distribution for the observed dependent variable data (4). To illustrate how the probability distributions which underpin the ZIP and ZINB approaches fit the observed data, the predicted proportions from 'empty' models (only the intercept fitted) were compared with the observed proportions from the DMF indices (Fig. 1a–d).

For all three data sets, the ZIP model was a poor fit. It predicted the proportion of zeros well, but, for DMFS >0, the distribution of the model predictions was very different to the distribution of the observed data. The ZINB model, however, produced a good fit for the entire range of DMFS

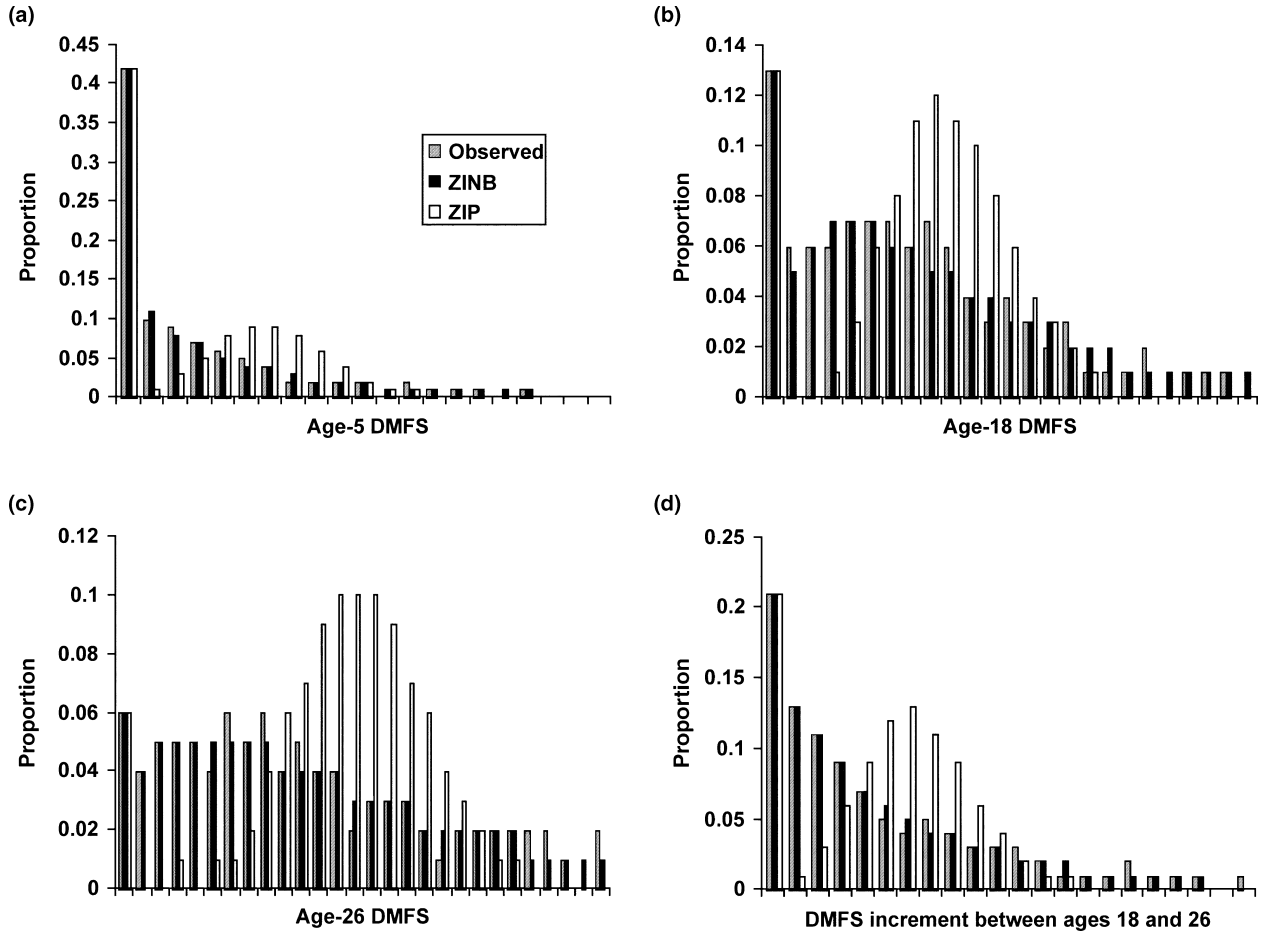

Fig. 1. Predicted proportions from intercept-only ZIP and ZINB models compared with the observed proportions from (a) dmfs at age 5 years, (b) DMFS at age 18 years, (c) DMFS at age 26 years and (d) net DFS increment.

values. For this reason, the ZIP model will no longer be considered for the cross-sectional data.

The outcome of the ZINB model with the covariates fitted for the age 5-, 18-, and 26-year data sets is shown in Table 1. The first and third columns relate to the modelling of the extra zeros (in the logit scale) and the negative binomial process (in the natural log scale), respectively. Interpreting the coefficients for the intercepts is made easier if, e.g. the age 5-year intercept value is viewed as representing the estimate for an individual who is male, of high SES, and has spent more than 50% of his life in a fluoridated area. The second column in the table presents the probability (calculated from the coefficient in the first column) of an individual with that particular characteristic being an extra zero. The fourth column presents the adjusted mean DMFS (calculated from the coefficients in the third column).

The largest coefficient (ignoring the intercepts) in the modelling of the extra zeros process is associated with the SESLOW variable (age 18 years). The probability of being an extra zero (i.e. 'caries-free') in the 'low-SES' group is anti-logit $(-1.288 - 1.304) = 0.07$, and the probability of being an extra zero in the 'high-SES' group is anti-logit$(-1.288) = 0.22$. The difference in probability between the two groups (0.15) is statistically significant, indicating that being in the high-SES group during childhood was associated with a greater probability of being caries-free by age 18 years, over and above that which would be expected from the negative binomial process. The SESLOW variable also has the largest coefficient in the modelling of the negative binomial process, but only at age 5 years. The adjusted mean dmfs score in the low-SES group is anti-$\log_e$ $(1.061 + 0.861) = 6.8$; the mean dmfs score in the high-SES group is anti-$\log_e(1.061) = 2.9$; the mean dmfs score in the medium-SES group is anti-$\log_e(1.061 + 0.484) = 4.7$. The mean values of the medium- and low-SES groups are statistically

Table 1. The coefficients and their 95% confidence intervals (in parentheses) from ZINB models for the age 5, 18, and 26 years cross-sectional caries data sets

	Logit	Probability of being an extra zero	Negative binomial	Adjusted DMFS
Age 5 years (dmfs)				
Intercept	−0.762 (−1.587, 0.063)	0.32	1.061 (0.675, 1.447)	2.9
SESLOW	−0.508 (−1.394, 0.378)	0.22	0.861 (0.447, 1.275)	6.8
SESMED	−0.320 (−1.061, 0.421)	0.25	0.484 (0.117, 0.851)	4.7
FEXP5 < 50%	0.024 (−0.490, 0.538)	0.32	−0.003 (−0.234, 0.228)	2.9
FEMALE	−0.013 (−0.530, 0.504)	0.32	−0.085 (−0.318, 0.145)	2.7
Age 18 years (DMFS)				
Intercept	−1.288 (−1.925, −0.651)	0.22	1.929 (1.743, 2.115)	6.9
SESLOW	−1.304 (−2.284, −0.324)	0.07	0.226 (0.008, 0.444)	8.6
SESMED	−1.014 (−1.659, −0.369)	0.09	0.144 (−0.040, 0.328)	8.0
FEXP18 < 50%	−0.245 (−0.835, 0.345)	0.18	0.089 (−0.033, 0.211)	7.5
FEMALE	0.384 (−0.196, 0.964)	0.29	−0.085 (−0.318, 0.148)	6.3
Age 26 years (DMFS)				
Intercept	−3.153 (−5.201, −1.105)	0.04	2.268 (2.080, 2.456)	9.7
SESLOW	−0.983 (−3.343, 1.377)	0.02	0.397 (0.178, 0.618)	14.4
SESMED	−0.532 (−2.110, 1.046)	0.02	0.205 (0.019, 0.391)	11.9
FEXP26 < 50%	−1.070 (−2.934, 0.794)	0.01	0.132 (0.005, 0.259)	11.0
FEMALE	0.809 (−0.839, 2.457)	0.09	0.021 (−0.106, 0.148)	9.9

significantly different from that of the high-SES group, and show a clear biological gradient. Thus, 5-year-old children from low-SES groups had, on average, nearly four more surfaces affected than their high-SES counterparts, and medium-SES children fell between those two groups.

The ZINB-adjusted estimates for caries severity and for the probability of being an extra zero by SES are presented in Fig. 2a,b (which uses the data from Table 1). The caries severity pattern indicates that the substantial SES differences which existed at age 5 years (in the deciduous dentition) had reduced somewhat by age 18 years, and had widened again by age 26 years. By contrast, the SES patterns in the probability of being an extra zero were substantial at age 5 years, greater at age 18 years (particularly with respect to the high SES group), and almost nonexistent by age 26 years.

Longitudinal results

The predicted proportions of 'empty' ZIP and ZINB models alongside the observed proportions

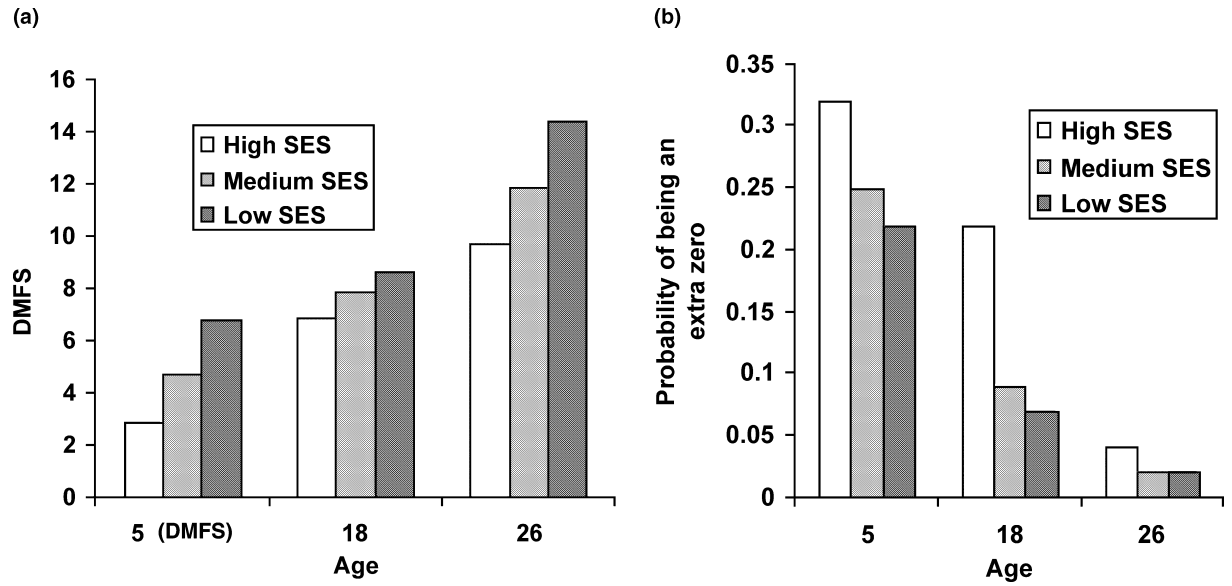

Fig. 2. (a) SES differences in caries severity (mean DMFS) and (b) probability of being an 'extra zero' at ages 5, 18 and 26 years.

Table 2. The coefficients and their 95% confidence intervals (in parentheses) from ZINB models for DFS increment between ages 18 and 26 years

	Logit	Probability of being an extra zero	Negative binomial	Adjusted DFS increment
Intercept	−0.734 (−2.146, 0.678)	0.32	1.341 (1.043, 1.639)	3.8
SESLOW	0.639 (−0.781, 2.059)	0.48	0.119 (−0.136, 0.374)	4.3
SESMED	0.790 (−0.389, 1.969)	0.51	−0.030 (−0.243, 0.182)	3.7
FEXP18 < 50%	−0.947 (−1.835, −0.059)	0.16	0.038 (−0.118, 0.194)	4.0
FEXP18_26	−0.761 (−1.649, 0.127)	0.18	−0.039 (−0.229, 0.152)	3.7
FEMALE	0.374 (−0.380, 1.127)	0.41	−0.142 (−0.292, 0.008)	3.3
EPISODIC	0.403 (−0.393, 1.198)	0.42	0.071 (−0.083, 0.224)	4.1
DMFS18	−0.348 (−0.525, −0.172)	0.25	0.044 (0.033, 0.055)	4.0

of the NETDFS index are shown in Fig. 1. As with the cross-sectional findings, the ZIP model provided a poor fit while the ZINB model provided a good fit. For this reason, the ZIP model will no longer be considered for the longitudinal data. The outcome of the ZINB model with covariates is presented in Table 2. 'Baseline' caries experience (age 18-year DMFS) was a predictor both of being an extra zero and of caries severity. This result should be considered with caution because there are methodological difficulties with modeling change when baseline score is included in the model as a predictor (10). However, when the relationship was assessed by plotting NETDFS against the mean of DMFS18 and DMFS26 (11), an association remained apparent. The only other significant predictor (negative) of being an extra zero was having spent less than half of one's life in a fluoridated area up to age 18 years.

Discussion

This investigation examined the utility of the ZIP and ZINB approaches for modelling three types of dental caries data from the same cohort study. The ZINB modelling approach was found to have the best fit, not only with cross-sectional deciduous and permanent dentition caries data, but also with longitudinal data on caries increment. This indicates that even after accounting for the 'spike' of extra zeros in the data sets, the remainder of the DMF distribution was too over-dispersed to be considered a Poisson distribution. It can be seen in Fig 1a–d that the ZIP models predict bimodal distributions, but the observed DMF distributions are clearly unimodal. It is unclear how typical the DMF distributions under study are, but they are similar to a deciduous caries distribution from Manchester, UK (12). However, for a Brazilian deciduous data set (6) the distribution was bimodal, and a ZIP model fitted well.

The models reveal some interesting differences in the way in which SES was associated with caries severity and prevalence in the cohort. The apparent reduction in SES differences in caries severity between ages 5 and 18 years and their widening again by age 26 years may be due to two possibilities. First, it may be that the universal access to publicly funded dental care for New Zealand children from age 5–17 years was responsible for the reduction in SES inequalities. Utilization prior to age 5 years is rather less uniform, and is strongly influenced by SES (13), as is the use of dental services after age 18 years, when virtually all public support ceases (14). Secondly, it may be that the differences merely reflect the fact that, by age 18 years, a substantial proportion of the permanent dentition has not been 'at risk' for many years. The actual situation is likely to be a combination of these. The longitudinal data indicate that the highest increment was observed among low-SES individuals (although the SES differences were not statistically significant), suggesting that the health effects of a low-SES childhood persist well into adulthood.

The SES patterns in the probability of being an extra zero were different, however, being substantial at age 5 years, greater at age 18 years (particularly with respect to the high SES group), and almost nonexistent by age 26 years. The latter is probably a reflection of the very high prevalence of caries experience by age 26 years (with only 6% of the cohort having a DMFS of 0), but the considerably higher probability of being an extra zero at age 18 years which was associated with high SES suggests that high SES confers a protective effect which operates over and above that which is observable with caries severity. There is the intriguing possibility that this is a consequence of

SES-associated differences in intervention threshold, with dentists (or dental therapists, who treat children up until they begin secondary schooling) being more reluctant to 'drill the first tooth' in the mouths of children who are of higher SES.

This investigation of the utility of the zero-inflated approach for modelling both cross-sectional and longitudinal caries data has shown that ZIP/ZINB models can provide new insight into disease patterns. It is anticipated that they will become increasingly used in epidemiological studies that use the DMF index as the outcome measure.

Acknowledgements

We thank the Dunedin Study members and their parents, Unit research staff, the New Zealand Dental Research Foundation, Air New Zealand, study founder Dr Phil A. Silva, and former dental principal investigator Dr R. Harvey Brown. The Dunedin Multidisciplinary Health and Development Research Unit is supported by the Health Research Council of New Zealand.

References

1. Klein H, Palmer CE, Knutson JW. Studies on dental caries. Public Health Rep 1938;53:751–65.
2. Spencer AJ. Skewed distributions – new outcome measures. Community Dent Oral Epidemiol 1997;25:52–9.
3. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc 1964;26:211–52.
4. Dobson AJ. An introduction to generalized linear models. 2nd ed. Boca Raton: Chapman & Hall; 2002.
5. Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med 2002;21:1461–9.
6. Böhning D, Dietz E, Schlattmann P, Mendonca L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. J R Stat Soc 1999;162:195–209.
7. Silva PA, Stanton W. From child to adult: the Dunedin Multidisciplinary Child Development Study. Auckland: Oxford University Press; 1996.
8. Elley WB, Irving JC. A socio-economic index for New Zealand based on levels of education and income from the 1966 census. New Zealand J Educ Stud 1972;7:153–67.
9. Everitt BS. The Cambridge dictionary of statistics. 2nd ed. Cambridge: Cambridge University Press; 2002.
10. Tu Y-K, Gilthorpe MS, Griffiths GS. Is reduction of pocket probing depth correlated with the baseline value or is it ''mathematical coupling''? J Dental Res 2002;81:722–6.
11. Oldham PD. A note on the analysis of repeated measurements of the same subjects. J Chronic Dis 1962;15:969–77.
12. Lewsey JD, Gilthorpe MS, Bulman JS, Bedi R. Is modelling dental caries a 'normal' thing to do? Community Dental Health 2000;17:212–7.
13. Beautrais AL, Fergusson DM, Shannon FT. Use of preschool dental services in a New Zealand birth cohort. Community Dent Oral Epidemiol 1982; 10:249–52.
14. Thomson WM. Use of dental services by 26-year-old New Zealanders. New Zealand Dental J 2001;97:44–8.

Appendix

The zero-inflated Poisson model can be expressed as:

$$\Pr(y|x) = \begin{cases} \pi + (1-\pi)\frac{e^{-\mu}\mu^y}{y!} & \text{for } y = 0 \\ (1-\pi)\frac{e^{-\mu}\mu^y}{y!} & \text{for } y > 0 \end{cases}$$

where y denotes the dependent variable, π is the probability of being an extra zero and $\mu = x\beta$, with x representing the independent variables and β the coefficients associated with x. Individuals with $y = 0$ can be thought of as consisting of two groups: the first is not part of the Poisson process, and the second is part of a Poisson distribution with mean μ but only taking zero values. Individuals with $y > 0$ are all considered part of the Poisson process. Note that the Poisson distribution is scaled by the probability of not being an extra zero. Model estimation is carried out by maximum likelihood. The zero-inflated negative binomial model is formulated as above by replacing the Poisson distribution ($e^{-\mu}\mu^y/y!$) with the negative binomial distribution. The expression for the negative binomial distribution is complex, and not shown here. The interested reader can find it in 5, along with further insights into zero-inflated modelling.