Considerations in Poisson and Negative Binomial Model Selection for Identification of Risk Factors for Caries Development in Potentially Heterogeneous High-Risk Populations

Pagán-Rivera K[1,2], Dawson DV[1,2], Weber-Gasparoni K[2], Warren JJ[2], Kramer KWO[3], Marshall TA[2], Johnsen DC[2]

[1]Department of Biostatistics, College of Public Health, University of Iowa, 145 N. Riverside Drive 100 CPHB, Iowa City, IA 52242

[2]College of Dentistry, University of Iowa, 801 Newton Rd, Iowa City, IA 52246

[3]Melbourne, FL

Abstract

Data were combined from four studies conducted in four eastern Iowa pediatric populations at high risk for dental caries (N=882). All studies scored caries at the surface level using the d1d2mfs caries score criteria. Multiple risk factors were evaluated using similar or identical protocols, including demographics, and children's dietary and oral hygiene habits. A set of candidate variables were chosen based upon bivariate analyses (Spearman and Kruskal-Wallis tests). Poisson and Negative Binomial models were fitted with those candidate variables. Zero-inflated Poisson and Negative Binomial models were also considered. Goodness-of-fit tests were computed for a subset of candidate models and information criteria were used to choose the best model. Challenges to model selection were posed by the possible heterogeneity of the different populations represented, and strategies to address this consideration were addressed. The Negative Binomial models were always preferred over the Poison models and Zero-inflated models were required with these data.

Key Words: Model Selection, Negative Binomial, Poison, Information Criteria

1. Introduction

Identification of caries patterns and risk factors for dental caries in children represents an important step in the development and implementation of preventive measurements (O'Sullivan and Tinanoff 1996, Chankanka et al. 2011). Moreover, the identification of oral health-related disparities among populations could improve those preventive measurements (Levin et al. 2009). In order to identify them, generalized linear models have been used. It is known that the Poisson model is used when count data are available. However, when the variance is different from the mean, a negative binomial model is preferred. In particular, Lewsey et al. (Lewsey et al. 2000) suggested the use of Negative Binomial and Poisson models when investigating factors associated with the number of decayed, missing or filled surfaces. The fact that some data might include individuals with non-cavitated teeth, brings the concern about having a bimodal distribution. It is in this scenario when zero-inflated models could be used, when the data contains a considerable large amount of zeros (Lambert 1992). That is, there is a probability, π, that the outcome

will be zero and a probability, $(1-\pi)$, that the outcome will follow a Poisson or a Negative Binomial distribution. When analyzing caries data some researchers have suggested approaches that included these kinds of distributions (Lewsey and Thomson 2004).

Combined data would allow the use of a bigger sample size which would increase the statistical power. However, the possibly heterogeneity of the datasets could lead to challenges when performing model selection. Therefore, the aim of this study is to identify factors that affect caries development in Iowa children from multiple high risk populations while overcoming the challenges of model selection.

2. Methods

2.1 Data

Demographic information, socioeconomic status, children's dietary information and oral hygiene habits were collected from four eastern Iowa pediatric populations at high risk for dental caries (N=882). Study-specific sample sizes are as follow: Davenport [n=415 (Weber-Gasparoni, Reeve, et al. 2013, Weber-Gasparoni, Warren, et al. 2013)], Muscatine [n=157 (Saba et al. 2014)], Carver [n=195 (Warren et al. 2009)], and Cedar Rapids [n=115 (Saba et al. 2014)]. Similar, and in many instances, identical protocols were used to collect caries risk factor information in the four studies. In particular, studies utilized a common protocol for the dental caries examinations, which recorded the number of d1d2mf surfaces. All studies used the same d1d2mfs caries scoring criteria, which included noncavitated (d1) and cavitated (d2) lesions, missing and filled surfaces (Warren et al. 2009).

2.2 Statistical Analyses

Possible associations between d1d2mfs and putative risk factors were initially assessed via bivariate analyses (Spearman rank correlations and Kruskal-Wallis tests); adjustment for multiple comparisons was made using the Bonferroni method in conjunction with an overall Type I error level of 0.05. In order to identify which risk factors were associated with dental caries, Poisson and Negative Binomial models were fitted. The model selection technique used was Akaike Information Criterion (AIC) (Akaike 1973). Both Poisson and Negative Binomial zero-inflated models were considered. Goodness-of-fit tests were performed for a subset of candidate models and information criteria were used to choose the best model. Moreover, site-specific models were fitted and the results were compared to the final model with the combined data. A model selection based on the significance of the variables was also performed and compared to the model selected by using the AIC. Models fitted included those with and without site-specific adjustments. Two-way interactions were also included in the models.

3. Results

3.1 Descriptors

The sample for this study consisted of a total of 882 children between 12 and 74 months of age from four studies, as described in Tables 1 and 2. In the combined sample, 49.6% were females, 38.7% minority, 50.8% had carious lesions, 34.6 % did not brush, 37.4% did not use fluoridated toothpaste, and 75.1% used city (fluoridated) water. Also, 80.6 % of the mothers had at least a high school or GED and 42.7 % were married at the time of the data collection.

Table 1: Descriptive Statistics

Variable	*N*	*Mean*	*Std. Dev*	*Min*	*Lower Quartile*	*Median*	*Upper Quartile*	*Max*
Children's age in months	881	30.77	13.70	12.17	20.43	28	38.2	74
Number of maxillary incisor teeth with visible plaque	875	1.61	1.55	0	0	1	3	5
Total number of all erupted teeth	882	16.32	4.40	2	16	17	20	20
d1d2mfs	882	3.77	7.77	0	0	1	4	72

Table 2: Frequencies

Variable	*Frequency*	*Percent*
Child's Sex		
Male	428	50.41
Female	421	49.59
Child's Race		
African American	81	9.27
Asian	3	0.34
Caucasian	536	61.33
Latino/a	144	16.48
Other	110	12.59
Annual Household Income		
$0-5,000	166	19.37
$5,001-10,000	110	12.84
$10,001-15,000	78	9.10
$15,001-20,000	182	21.24
$20,001-25,000	109	12.72
$25,001-30,000	86	10.04
$30,001+	126	14.70
Mother's Highest Level of Education		
less than HS	169	19.45
HS/GED	521	59.95
2-year	95	10.93
4-year	57	6.56
graduate	27	3.11
Brushing Habits		
no	301	34.60
yes, occasionally	135	15.52
yes, daily	434	49.89
Use of Fluoridate Toothpaste		
no	304	37.44
yes, occasionally	118	14.53
yes, daily	390	48.03
Water Source		
well	45	5.18
city	653	75.14
bottled	102	11.74
filtered/other/>1	69	7.94

3.2 Bivariate Analyses

When exploring the relationship between the d1d2mfs and the explanatory variables, we found that the d1d2mfs caries score criteria was positively correlated with the child's age in months ($r=0.60$; $p<0.0001$), the total number of all erupted teeth ($r=0.61$; $p<0.0001$) and the number of maxillary incisors with visible plaque ($r=0.25$; $p<0.0001$). Caries score was

negatively correlated with the mother's highest level of education (r=-0.18; p<0.0001) (Table 3). Kruskal-Wallis tests showed that there was a difference in the d1d2mfs caries score criteria among ethnic groups (p<0.0001), annual household income (p=0.0117), and brushing and use of fluoridate toothpaste habits (both p<0.0001). After adjusting for multiple comparisons, the associations between caries score and child's age in months, total number of all erupted teeth, mother's highest level of education, number of maxillary incisor teeth with visible plaque, ethnic groups, and brushing and use of fluoridate toothpaste habits remained significant.

Table 3: Spearman Rank Correlation between explanatory variables and caries score (d1d2mfs)

Explanatory Variable	*N*	*Correlation Coefficient*	*p-value*
Child's age in months	447	0.60	<0.0001
Total number of all erupted teeth	448	0.61	<0.0001
Mother's higher level of education	440	-0.18	<0.0001
Number of maxillary incisor teeth with visible plaque	447	0.25	<0.0001

3.3 Model Selection

The six risk factors identified in bivariate analyses formed the basis of the set of candidate variables for initial modeling of d1d2mfs; other candidate variables also considered included: child's sex, child's race, annual household income, mother's marital status, and water source. Additionally, first-order interactions between these variables were evaluated. The primary criterion used in model selection was the AIC. This model selection technique always preferred the Negative Binomial models over Poisson models. Moreover, zero-inflated negative binomial models were appropriate.

Based on the AIC, the best model is the one that includes: child's age in months, total number of all erupted teeth, annual household income, mother's highest level of education, use of fluoridate toothpaste, number of maxillary incisor teeth with visible plaque, child's sex and race, site/study, mother's marital status and an interaction term between site/study and child's age in months (Table 4). Goodness-of-fit tests and model diagnostics raised no concerns regarding model assumptions or fit. It is important to note that even though the child's sex was not statistically significant, it was kept in the model since it was a confounder of interest, there was a difference of more than 70 in the AIC when it was removed, and its removal was associated with substantial changes in the other parameters.

Table 4: Results of the final model

Variable	*β estimate*	*p-value*
Child's age in months	-0.1127	0.0216
Total number of all erupted teeth	0.1714	0.0001
Number of maxillary incisor teeth with visible plaque	0.1049	0.0034
Child's sex (Reference: female)	0.1054	0.3267
Site/Study (Reference: Cedar Rapids)		overall p-value = 0.0005
Carver	-3.0645	0.0217
Davenport	-4.6356	0.0005
Muscatine	-3.0061	0.0679
Interaction term between site and child's age in months		0.0031
Annual household income (Reference: $0-5,000)		overall p-value = 0.0025
$5,001-10,000	-0.5714	0.0024
$10,001-15,000	-0.4584	0.0343
$15,001-20,000	-0.3393	0.0395
$20,001-25,000	-0.0839	0.6801
$25,001-30,000	0.0895	0.6730
$30,001+	-0.5381	0.0061
Mother's highest level of education (Reference: less than HS)		overall p-value = 0.0350
HS/GED	-0.3178	0.0162
2-year	-0.4568	0.0274
4-year	-0.5285	0.0407
Graduate	-0.8738	0.0424
Use of fluoridated toothpaste (Reference: no)		overall p-value = 0.0038
yes, occasionally	0.0781	0.6863
yes, daily	0.4725	0.0034
Children's race (Reference: Caucasian)		overall p-value = 0.0166
African American	0.1751	0.3574
Asian	-0.0893	0.8974
Latino/a	0.4445	0.0012
Other	-0.0568	0.7501
Mother's Marital Status (Reference: Single, never married)		overall p-value = 0.0485
Married	-0.0214	0.8784
Separated	-0.1737	0.4640
Divorced	0.2116	0.3001
Live with significant other	0.4805	0.0121
Widowed	-1.2706	0.2320
Zero-Model		
Child's age in months	-0.0746	0.0044
Total number of all erupted teeth	-0.3938	0.0001
Use of fluoridated toothpaste (Reference: no)		overall p-value = 0.0120
yes, occasionally	-0.7390	0.2793
yes, daily	0.9038	0.0899
Site/Study (Reference: Cedar Rapids)		overall p-value = 0.0001
Carver	-0.2939	0.6648

Davenport	-2.4095	0.0004
Muscatine	-0.7959	0.1298

This model suggested that higher level of mother's education was associated with lower d1d2mfs score, and that Latino children were at higher risk of having d1d2mfs surfaces than white children. An outcome that was not expected was that the use of fluoridated toothpaste was associated with higher mean d1d2mfs. This relationship could perhaps be explained by the fact that the use of fluoridated toothpaste is also associated with the child's age and the child's number of erupted teeth which implies a higher risk of having d1d2mfs surfaces, or by recommendations to parents in children with carious lesions. The child's age in months, the number of erupted teeth, use of fluoridated toothpaste and study/site were the variables that explained the zero inflation part of this model. In particular, the log odds of being in the excess zero group would decrease as the child gets older and as the number of erupted teeth increases. The interaction term showed that in the Carver, Muscatine and Davenport population the mean d1d2mfs increased as the child's age increased; however, the age relationship was much less prominent in the Cedar Rapids population (Figure 1).

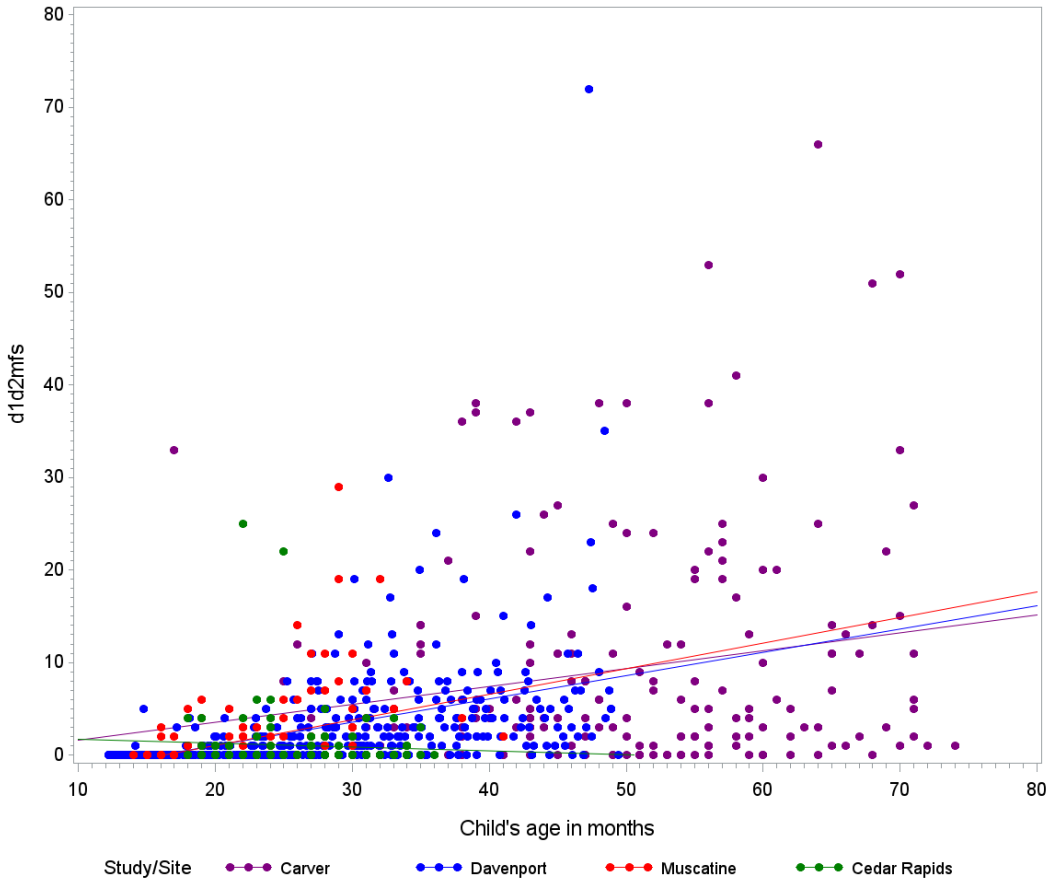

Figure 1: Relationship between d1d2mfs, child's age in month and study/site

When fitting site-specific models, a complication was faced. The Cedar Rapids site had a significantly smaller dataset which made impossible to fit a model with too many variables. However, the other three sites (Carver, Muscatine and Davenport) had a larger sample size

and more models could be fitted and compared. In fact, the best model for those sites included almost all covariates in the model derived from the combined data.

When adjustment for site differences was not part of the modeling process, the final model based on the significance of the variables was similar to the model derived via AIC. However, if the site variable was added to the best model based upon significance criteria, it was significant after adjustment for the other covariates. This suggests that some additional factors that were not part of the data collection, but are associated with site, may have a useful explanatory role in dental caries. Therefore, when combining the data, an adjustment for those site differences should be given serious consideration.

4. Discussion

The aim of this analysis was to identify risk factors for dental caries in children. The ones detected were: child's age, total number of all erupted teeth, annual household income, mother's highest level of education and marital status, use of fluoridated toothpaste, number of maxillary incisors with visible plaque, child's sex and race and the study/site. There was a significant interaction between child's age in months and the study site. There were site differences which implies that the risk factors vary from site to site. This findings have a public health impact in the sense that preventive and intervention methods should take into account the specific risk factors for that population. The interaction term showed that the child's age does not have the same impact in all sites. Appropriate adjustment for covariates will be an important preliminary step to identification and comparison of caries patterns within these high risk groups.

Acknowledgements

Supported by NIH grants R21-DE15008 and R21-DE016483, the Roy J. Carver Charitable Trust and the University of Iowa.

References

Akaike, Hirotugu. 1973. Information theory and an extension of the maximum likelihood principle. Paper read at Second International Symposium on Information Theory, at Budapest.

Chankanka, Oitip, Joseph E Cavanaugh, Steven M Levy, Teresa A Marshall, John J Warren, Barbara Broffitt, and Justine L Kolker. 2011. "Longitudinal associations between children's dental caries and risk factors." *Journal of public health dentistry* no. 71 (4):289-300.

Lambert, Diane. 1992. "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics* no. 34 (1):1-14.

Levin, Kate A, Carolyn A Davies, Gail VA Topping, Andrea V Assaf, and Nigel B Pitts. 2009. "Inequalities in dental caries of 5-year-old children in Scotland, 1993–2003." *The European Journal of Public Health* no. 19 (3):337-342.

Lewsey, James D, Mark S Gilthorpe, John S Bulman, and Raman Bedi. 2000. "Is modelling dental caries a'normal'thing to do?" *Community dental health* no. 17 (4):212-217.

Lewsey, James D, and William M Thomson. 2004. "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 Dentistry and Oral Epidemiology* no. 32 (3):183-189.

O'Sullivan, David M, and Norman Tinanoff. 1996. "The association of early dental caries patterns with caries incidence in preschool children." *Journal of public health dentistry* no. 56 (2):81-83.

Saba, Ann H, John J Warren, Karin Weber-Gasparoni, and Deborah V Dawson. 2014. "Retention of Low Income Children in Three Dental Studies Investigating Early Childhood Caries." *Journal of Health Disparities Research and Practice* no. 7 (4):77-90.

Warren, John J, Karin Weber-Gasparoni, Teresa A Marshall, David R Drake, Farideh Dehkordi-Vakil, Deborah V Dawson, and Katie M Tharp. 2009. "A longitudinal study of dental caries risk among very young low SES children." *Community dentistry and oral epidemiology* no. 37 (2):116-122.

Weber-Gasparoni, K., J. Reeve, N. Ghosheh, J. J. Warren, D. R. Drake, K. W. Kramer, and D. V. Dawson. 2013. "An effective psychoeducational intervention for early childhood caries prevention: part I." *Pediatr Dent* no. 35 (3):241-6.

Weber-Gasparoni, K., J. J. Warren, J. Reeve, D. R. Drake, K. W. Kramer, T. A. Marshall, and D. V. Dawson. 2013. "An effective psychoeducational intervention for early childhood caries prevention: part II." *Pediatr Dent* no. 35 (3):247-51.