Table of content:

Description of data

Participants were selected using a multistage stratified cluster sampling method. First, each province was divided into 2 strata, urban and rural areas, and then two districts or counties were randomly selected from each stratum. Second, three middle schools were randomly chosen from every district or county from the list of middle schools provided by the council. Third, at each middle school, a random sample of 320 children aged 12 to 15 years old was recruited using the list of enrollees, with 80 children enrolled for each age group. A target sample of 3,840 participants was initially established per province for a total of 119,040 children nationally. Finally, a total of 118,601 school-age participants (27,821 12-year-olds, 30,961 13-year-olds, 30,691 14-year-olds and 29,128 15-year-olds) completed the clinical examination and questionnaire survey.

All permanent teeth of the participants were examined, and dental caries were diagnosed according to the WHO criteria. Unified training sessions were provided for representative survey examiners from all provinces before the national survey began. For the reliability assessment, duplicate examinations were conducted during the main survey. Five percent of the participants were re-examined to calculate inter-examiner reliability. The Kappa score was 0.94 for the examination of dental caries in 12- to 15-year-olds.

Estimation of $\boldsymbol{\beta}^{(k)}$

Consider model (1) as a latent class regression model. To each objection $y_i, i = 1, \dots, n$, there exists an unobserved latent subgroup indicator vector $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})^{\mathrm{T}}$ with $z_{ik} = 1$ if y_i comes from subgroup k and 0 otherwise. It is obvious that to each $\mathbf{z}_i$, only one component is 1 and all others are 0. In addition, we know that $\mathbf{z}_i, i = 1, \dots, n$ independently belong to multinomial distribution with probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^{\mathrm{T}}$. As the conditional probability function of y_i given $\mathbf{z}_i$ is $\prod_{k=1}^{K} \mathrm{Poi}\{y_i|\lambda_k(\mathbf{x}_i, \boldsymbol{\beta}^{(k)})\}^{z_{ik}}$, the complete likelihood function takes the form of

$$L(y, \mathbf{z};\ \boldsymbol{\pi}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k \mathrm{Poi}\{y_i|\lambda_k(\mathbf{x}_i; \boldsymbol{\beta}^{(k)})\}]^{z_{ik}},$$

and the complete log-likelihood function is

$$l(y, \mathbf{z};\ \boldsymbol{\pi}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log{[\pi_k \mathrm{Poi}\{y_i|\lambda_k(\mathbf{x}_i; \boldsymbol{\beta}^{(k)})\}]},$$

where $\boldsymbol{\beta} = (\{\boldsymbol{\beta}^{(k)}\}^{\mathrm{T}}, k = 1, \dots, K)^{\mathrm{T}}$.

As $\mathbf{z}_i, i = 1, \dots, n$ are unobserved, we use the EM algorithm to estimate $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\pi}^{\mathrm{T}})^{\mathrm{T}}$. In the E-step, we replace z_{ik} by its posterior expectation.

$$\frac{\pi_k \mathrm{Poi}\{y_i|\lambda_k(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^{(k)})\}}{\sum_{k=1}^{K}\pi_k \mathrm{Poi}\{y_i|\lambda_k(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^{(k)})\}}.$$

Then, in the M-step, optimizing the complete log likelihood function with $\mathbf{z}_i$ fixed at the values computed at E-step, we could get the estimator of $\boldsymbol{\beta}^{\mathrm{T}}$ and $\boldsymbol{\pi}^{\mathrm{T}}$.

Finally, we can classify all individuals into different subgroups by comparing the estimator of posterior probabilities with form:

$$\hat{P}\left(k\middle|y_i,\mathbf{x}_i;\ \widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\beta}}\right)=\frac{\hat{\pi}_k \mathrm{Poi}\left\{y_i\middle|\lambda_k\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^{(k)}\right)\right\}}{\sum_{k=1}^{K}\hat{\pi}_k \mathrm{Poi}\left\{y_i\middle|\lambda_k\left(\mathbf{x}_i;\widehat{\boldsymbol{\beta}}^{(k)}\right)\right\}}.$$

Appendix Table 1. Variables involved in the present study

Categories	Subcategories	Options
Social demographic factors	X_1: indicator variable 1 to region*	1 = "east area", 0 = "otherwise".
	X_2: indicator variable 2 to region*	1 = "west area", 0 = "otherwise".
	X_3: age,	ranges from 11.92 to 16.08.
	X_4: gender	1 = "male", 0 = "female".
	X_5: type of census register	1 = "urban", 0 = "rural", 7 observations are missing.
	X_6: only child	1 = "yes", 0 = "no", 9 observations are missing.
	X_7: paternal educational level	1 = "never accepted education", 2 = "elementary school", 3 = "middle school", 4 = "high school", 5 = "secondary school", 6 = "college", 7 ="undergraduate", 8 = "graduate or higher", 13445 observations are missing and the missing rate is 11.34%.
	X_8: maternal educational level	The score is of the same with X_7, 13,851 observations are missing and missing rate is 11.76%.
Oral hygiene behavior	X_9: Frequency of toothbrushing	1 refers to "daily or often", 0 represents "seldom or never", 7 observations are missing.
Sugar consumption habits	X_{10}: score of sugar consumption	The score of sweet foods, drinks and milk/coffee/tea taken frequency, ranges from 1 to 16, higher score indicates more frequent, 18 observations are missing.
Pit-and-fissure sealant history	X_{11}	1 = "yes", 0 = "no".
Oral health knowledge	X_{12}: the score of oral knowledge	ranges from 0 to 8, higher score means better knowledge equipment.
	X_{13}: self-assessment to tooth	1 = "very good", 2 = "good", 3 = "fair", 4 = "poor" and 5 = "bad", 9 observations are missing.
	X_{14}: ever visit dentist or not	1 = "yes", 0 = "no", 11 observations are missing.

* According to the National Health Statistical Yearbook 2015, the 31 provinces were divided into three regions, which were east, middle, and west.

Appendix Table 2. Results of model fitting with different subgroup numbers.

Iteration	Converged	k	k_0	log-likelihood	BIC
32	True	2	2	-856569.4	1713607
112	True	3	3	-838691.8	1678092
342	True	3	4	-836939.2	1674587
341	True	3	5	-835813.9	1672336
503	True	3	6	-836938.4	1674585

Note: k_0 represents pre-selected number of subgroups, k is the final number of subgroups during the computation, "Iteration" indicates the times of iteration.

Appendix Table 3. Distributions of observed and fitted DMFT values by Poisson mixture regression model / Poisson mixture regression model with ZIP modification

Value	0	1	2	3	4	5	>=6
Subgroup 1							
Observed DMFT	39087	6260	4063	1818	798	312	238
Poi-mixture	40099	7183	3554	1169	340	144	87
ZIP Poi-mixture	39563	2938	7815	1901	308	51	0
Subgroup 2							
Observed DMFT	19809	10296	6987	3121	1236	343	177
Poi-mixture	20098	12602	7514	1307	321	96	31
ZIP Poi-mixture	19804	4816	15275	1885	176	13	0
Subgroup 3							
Observed DMFT	0	0	91	839	1721	1203	2587
Poi-mixture	0	4	235	1303	2004	1508	1387
ZIP Poi-mixture	-	-	-	-	-	-	-

Appendix Table 4. The *P* value for the effects of the same variable in different three subgroups

Variables	*P* value for subgroup1&2&3
Intercept	< 0.001
X_1	< 0.001
X_2	< 0.001
X_3	< 0.001
X_4	< 0.001
X_5	< 0.001
X_6	< 0.001
X_7	< 0.001
X_8	**0.305**
X_9	< 0.001
X_{10}	< 0.001
X_{11}	< 0.001
X_{12}	**0.314**
X_1X_{10}	< 0.001

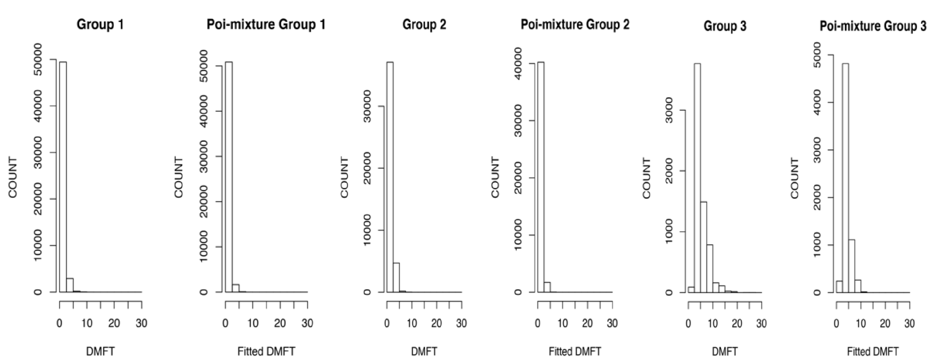
X_2X_{10}	< 0.001
X_9X_{13}	< 0.001
X_9X_{14}	< 0.001

Appendix Table 5. The P value for the effects between each two groups

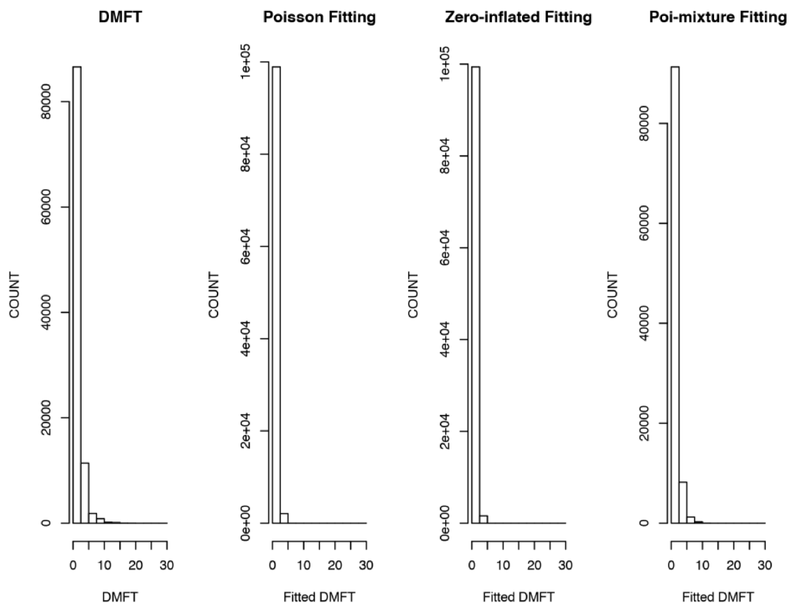
Variables	subgroup1&2	subgroup1&3	subgroup2&3
Intercept	< 0.001	< 0.001	< 0.001
X_1	< 0.001	< 0.001	< 0.001
X_2	< 0.001	**0.719**	**0.095**
X_3	< 0.001	< 0.001	0.025
X_4	< 0.001	**0.877**	< 0.001
X_5	< 0.001	< 0.001	< 0.001
X_6	< 0.001	< 0.001	< 0.001
X_7	< 0.001	< 0.001	< 0.001
X_9	< 0.001	< 0.001	< 0.001
X_{10}	< 0.001	< 0.001	< 0.001
X_{11}	**0.115**	0.028	< 0.001
X_1X_{10}	< 0.001	< 0.001	< 0.001
X_2X_{10}	< 0.001	**0.395**	0.024
X_9X_{13}	< 0.001	< 0.001	< 0.001
X_9X_{14}	< 0.001	< 0.001	< 0.001

Appendix Table 6. Distributions of observed DMFT and fitted DMFT by Poisson, zero-inflated Poisson and Poisson mixture regression model.

Value	0	1	2	3	4	5	>=6
Observed DMFT	58896	16556	11141	5778	3755	1858	3002
Poisson	9408	77152	12357	1671	315	78	5
Zero-inflated	10193	76024	13178	1423	165	3	0
Poi-mixture	60197	19789	11303	3779	2665	1748	1505



Appendix Figure 1: Histograms of observed DMFT and fitted DMFT within each subgroup.



Appendix Figure 2: Histograms of observed and fitted DMFT.