

CHINESE MEDICAL SCIENCES JOURNAL

ORIGINAL ARTICLE

Study of Zero-Inflated Regression Models in a Large-Scale Population Survey of Sub-Health Status and Its Influencing Factors[△]

Tao Xu¹, Guangjin Zhu², Shaomei Han^{1*}

¹Department of Epidemiology and Statistics, ²Department of physiopathology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College, Beijing 100005, China

Key words: zero-inflated; negative binomial regression; sub-health; population survey

Objective Sub-health status has progressively gained more attention from both medical professionals and the publics. Treating the number of sub-health symptoms as count data rather than dichotomous data helps to completely and accurately analyze findings in sub-healthy population. This study aims to compare the goodness of fit for count outcome models to identify the optimum model for sub-health study.

Methods The sample of the study derived from a large-scale population survey on physiological and psychological constants from 2007 to 2011 in 4 provinces and 2 autonomous regions in China. We constructed four count outcome models using SAS: Poisson model, negative binomial (NB) model, zero-inflated Poisson (ZIP) model and zero-inflated negative binomial (ZINB) model. The number of sub-health symptoms was used as the main outcome measure. The alpha dispersion parameter and *O* test were used to identify over-dispersed data, and Vuong test was used to evaluate the excessive zero count. The goodness of fit of regression models were determined by predictive probability curves and statistics of likelihood ratio test.

Results Of all 78 307 respondents, 38.53% reported no sub-health symptoms. The mean number of sub-health symptoms was 2.98, and the standard deviation was 3.72. The statistic *O* in over-dispersion test was 720.995 ($P < 0.001$); the estimated alpha was 0.618 (95% *CI*: 0.600-0.636) comparing ZINB model and ZIP model; Vuong test statistic *Z* was 45.487. These results indicated over-dispersion of the data and excessive zero counts in this sub-health study. ZINB model had the largest log likelihood (-167519), the smallest Akaike's Information Criterion coefficient (335112) and the smallest Bayesian information criterion coefficient (335455), indicating its best goodness of fit. The predictive probabilities for most counts in ZINB model fitted the observed counts best. The logit section of ZINB model analysis showed that age, sex, occupation, smoking, alcohol drinking, ethnicity and obesity were determinants for presence of sub-health symptoms; the binomial negative

Received: March 30, 2017.

*Corresponding author Tel: 86-10-69156408, E-mail: hansm1@vip.sina.com

△Fund supported by the Basic Performance Key Project, the Ministry of Science and Technology of the People's Republic of China (No. 2006FY110300).

section of ZINB model analysis showed that sex, occupation, smoking, alcohol drinking, ethnicity, marital status and obesity had significant effect on the severity of sub-health.

Conclusions All tests for goodness of fit and the predictive probability curve produced the same finding that ZINB model was the optimum model for exploring the influencing factors of sub-health symptoms.

Chin Med Sci J 2017; 32(4):218-225. DOI:10.24920/J1001-9294.2017.054

RECENTLY sub-health becomes an important public health issue and has attracted more and more attention from both medical professionals and the public. However, people usually think that sub-health is a borderline state between health and disease, which might be caused by strong social stress. Sub-health is a suboptimal health status, and an intermediate health state between health and disease, which is characterized by a decline in vitality, physiological function and capacity for adaptation, and it refers to medically undiagnosed or functional somatic syndromes.^{1,2} It has been reported that 60%-70% of Chinese people suffering from sub-health.³ It is very important to properly assess sub-health and explore the its influencing factors in order to prevent diseases and promote the health status of the general population.

Nowadays, sub-health status is usually assessed with a variety of rating scales. Delphi self-rating sub-health scale is very applicable to sub-health assessment for community population. It assesses 18 symptomatic items. If a subject has been suffering from one or more symptoms for more than one month in the past year, he or she will be considered as of sub-health status.⁴ Previous studies showed that Delphi self-rating sub-health scale had good reliability and repeatability.⁴⁻⁶

In rating scales, subjects are usually categorized into two or several groups based on whether they had positive symptoms. The prevalence rate and logistic regression are usually used to study the incidence intensity and the relationship between sub-health and its risk factors.⁶⁻¹⁰ However, these traditional methods may lose information, because every subject has different number of sub-health symptoms so that categorization will result in inability to assess the severity of sub-health status. Actually, the number of sub-health symptoms is a kind of count data, in which observations take only non-negative integer values {0, 1, 2, 3 ...}. During statistical treatment, if they are considered as continuous outcomes or transferred to dichotomous data, the numbers of sub-health symptoms are often extremely concentrated and don't follow the normal distribution. Consequently, arithmetic means and standard deviations are not applicable, and linear regression is not appropriate because of skewed distribution

and over-dispersion. Although linear regression and logistic regression models are often used to treat count data, the results are likely to be inefficient, inconsistent or biased.^{11,12}

In view of these limitations, Poisson regression or negative binomial (NB) regression is commonly used to model count outcomes based on their specific distributions. However, a lot of subjects have no sub-health symptom and there is a large proportion of zero count, which has adverse effect on the goodness of fit of Poisson regression or NB models. Neglect of excess zeroes will bias the estimation of parameters.¹³ Zero-inflated (ZI) regression models consider the raw dataset as a mixture of an all-zero subset and another subset following Poisson or NB distribution.¹³⁻²⁰ ZI models were firstly introduced by Lambert to explain excess zero counts.²¹ Cheung YB mentioned that ZI regression models can be interpreted as reckoning a two-step disease regression.¹³ At the beginning subjects are not at the risk, so they have zero counts. The influence of covariates may move them into the at-risk population and the outcomes follow a Poisson or NB regression distribution. A covariate may or may not have the same direction of impact in the two steps.

This study was designed with a large-scale population survey to compare the goodness of fit of four count outcome models: Poisson model, NB model, Zero-inflated Poisson (ZIP) model and Zero-inflated negative binomial (ZINB) model, aiming at identifying the optimum model for the study of sub-health status.

MATERIALS AND METHODS

Sample and participants

A large-scale population survey about physiological and psychological constants was conducted in Chinese people during 2007-2011. It was supported by the basic performance key project of the Ministry of Science and Technology of the People's Republic of China. This survey was conducted in four provinces: Sichuan, Heilongjiang, Hunan, Yunnan; and two autonomous regions: Inner Mongolia, and Ningxia Hui. Two-stage cluster sampling method was used to select eligible subjects in every province. Firstly, two or three cities were sampled based

on their economical status and ethnic inhabitations of minorities. Then several communities or villages in each city were randomly selected, where all eligible people were referred to participate the survey. Eligible subjects were those who didn't run a high fever in the past 15 days, and weren't suffering from dysplasia or serious chronic diseases involving main organs such as heart, liver, lung, brain, kidney, etc. This study has been approved by the institutional review board. Signed informed consent forms were acquired from all subjects who were willing to take part in the survey.

Sub-health assessment

A sub-health rating scale framed by researchers from Medical College of Jinan University was used to assess the sub-health status of community population.⁴ This scale includes 18 symptom items that are grouped into 6 dimensions: physical symptoms, psychological symptoms, vigor, social adaptability, immunity, and going to hospital. The 18 sub-health symptoms consist of fatigue, headache or dizziness, tinnitus, numbness or stiffness in the shoulders or legs, a sense of pharyngeal foreign bodies, upset, loneliness, inattention, anxiety, dreamful, forgetful, decreased vitality, not interested in surroundings, moody, feeling tired at work, incompatible with coworkers, susceptible to flu or other diseases, feeling suffering from undiagnosed diseases. The number of sub-health symptoms, which was the count of "positive" responses of these 18 questions (range 0-18), was the main outcome measurement of this study.

Strict quality control standards were carried out. All surveyors were trained according to the training manual and were tested before the survey. A preliminary survey was conducted in order to verify competence of the surveyors.

ZI models construction

SAS (version 9.2) was used in the regression model construction. Two-tailed $P \leq 0.05$ was considered as statistically significant. Four count outcomes models: Poisson model, NB model, ZIP model and ZINB model were constructed. Poisson regression and NB regression are commonly used to model count outcomes based on Poisson distribution or NB distribution. But with excess zero counts, ZI models may be better. ZI regression models can be interpreted as reckoning a two-step disease regression.

ZIP model refers to a raw dataset as a mixture including an all-zero subset and a subset following Poisson distribution.^{21,22} ZIP model supposes that:

$$P(y_i) = \begin{cases} \pi_i + (1 - \pi_i)e^{-\mu} & y_i = 0 \quad \text{Logit section} \\ (1 - \pi_i) \frac{e^{-\mu} \mu^{y_i}}{y_i!} & y_i \geq 1 \quad \text{Poisson section} \end{cases}$$

At the same time, ZINB model refers to a raw dataset as a mixture including an all-zero subset and a subset following NB distribution.²¹⁻²² The probability density function of ZINB model is that:

$$P(y_i) = \begin{cases} \pi_i + (1 - \pi_i) \left(\frac{1}{1 + \alpha\mu} \right)^{\alpha^{-1}} & y_i = 0 \quad \text{Logit section} \\ (1 - \pi_i) \frac{\Gamma(y_i + \alpha^{-1})(\alpha\mu)^{y_i}}{y_i! \Gamma(\alpha^{-1})(1 + \alpha\mu)^{(y_i + \alpha^{-1})}} & y_i \geq 1 \quad \text{NB section} \end{cases}$$

π_i is the probability of excessive zero counts. In the logit section, explanations for regression coefficients are similar to those in logistic regression. And then in the Poisson or NB section, the explanations are the same as those in the traditional Poisson or NB regression models.

Determination of the optimum model

The alpha dispersion parameter and O test were used to identify over-dispersion of count data.²³ When the alpha dispersion parameter equals to 0, NB model and ZINB model will reduce to Poisson model and ZIP model. So the alpha dispersion parameter can be used to compare the nested models, Poisson vs. NB, and ZIP vs. ZINB. Vuong test was conducted to compare non-nested models, NB vs. ZINB, in order to judge whether there were excessive zero count.^{22, 24} The goodness of fit of regression models were determined by the predictive probability curve, likelihood ratio test statistics: log-likelihood, Akaike's Information Criterion (AIC) and Bayesian information criterion (BIC). Predictive probability distributions of the four count outcome models were examined to see how these models fitted the observed proportions of 19 counts.

In the total population, a sample was randomly selected based on the proportional stratified random sampling method with a sampling proportion of 10 percent, in order to further assess the goodness of fit of the four count outcome models.

Definition of potential determinants

Potential determinants of sub-health symptoms included in the models were age, sex, hypertension, occupation, tobacco smoking, alcohol drinking, nationalities, marital status, and obesity. Age was a continuous variable. Sex was a dichotomized variable: male and female. Blood pressure was measured in the morning after the subject rested for five minutes in the seating position with her or

his back being supported, feet on the floor and the right arm being supported with the cubit fossa at the level of the heart. The appropriate cuff was chosen according to the arm circumference. Subjects were defined as being hypertensive if they had an average systolic blood pressure (SBP) equal to or greater than 140mmHg, and/or average diastolic blood pressure (DBP) equal to or greater than 90mmHg, and/or if they were diagnosed as being hypertensive in the past, and/or reported currently on antihypertensive medications. Marital status included married, single, divorced or widowed. Occupation referred to physical labor or mental labor. With a broad range of ethnicity, subjects were grouped as Han, Yi, Miao, Mongolia, Tibetan, Hui, Tujia, Korean, Manchu, Yao, Dai, Qiang and others. Smoking was categorized as non-smokers, current smokers or former smokers. Alcohol drinking was categorized as regular drinkers or non-drinkers. Obesity was defined as a body mass index (BMI) equal to or greater than 28 kg/m².

RESULTS

Demographic characteristics

82,336 subjects signed the inform consent of this survey, among whom 78,307 subjects completed all survey scales. The completion rate was 95.1%. The mean age of all respondents was 32.4±19.7 (ranged from 10 to 80) years. The demographic characteristics of the sample were presented in Table 1. The percentage of female respondents was 56.85%. About two-thirds (66.86%) were mental labors, such as teachers, doctors, professionals, students, governmental and institute employees. Widowed or divorced respondents accounted for 3.40% of all respondents. About three-fourths (74.18%) were Han nationalities. Percentages of regular smokers and drinkers were 14.91% and 15.08% respectively. Percentages of hypertensive subjects and obese ones were 19.69% and 8.06% respectively.

Sub-health symptoms

The Cronbach alpha coefficient was 0.814, and Guttman split-half coefficient was 0.765, which showed good reliability of the sub-health rating scale. Of all 78,307 respondents, 38.53% didn't report any sub-health symptoms, 11.18% and 9.59% reported one and two sub-health symptoms respectively. The proportion of respondents decreased with the increase of sub-health symptoms.

The goodness of fit of count outcome models

The mean number of sub-health symptoms was 2.98±3.72.

Table 1. Demographic characteristics of respondents about this study in year 2007-2011 (*n*=78307)

Characteristics	<i>n</i>	%
Sex		
Male	33 793	43.15
Female	44 514	56.85
Occupation		
Mental	52 357	66.86
Physical	25 950	33.14
Ethnicity		
Han	58 087	74.18
Yi	3 914	5.00
Miao	1 146	1.46
Mongolia	2 998	3.83
Tibetan	1 787	2.28
Korean	2 612	3.34
Hui	4 639	5.92
Tujia	1 657	2.12
Others	1 467	1.87
Marital status		
Married	40 207	51.35
Single	35 434	45.25
Widowed or divorced	2 666	3.40
Hypertension	15 416	19.69
Obesity	6 312	8.06
Tobacco smoking	11 677	14.91
Alcohol drinking	11 811	15.08

The sample variance was 13.84, which was significantly bigger than the mean. The over-dispersion test statistic *O* was 720.995, and the *P* value was less than 0.001. The alpha parameter was estimated as 1.671 for comparing the NB model with Poisson model (95% *CI*: 1.646-1.695). Furthermore, the alpha parameter was 0.618 (95% *CI*: 0.600-0.636) for comparing ZINB model with ZIP model. Alpha dispersion parameters and *O* test findings showed that the number of sub-health symptoms was over-dispersed. Both Poisson distribution and ZIP distribution were worse than corresponding NB distribution and ZINB distribution for count outcomes of the number of sub-health symptoms.

Vuong test was used to compare NB model with ZINB model. The test statistic *Z* was 45.487 and the *P* value was

less than 0.001, which indicated that there were too many zero counts to be treated with traditional NB distribution. ZINB model was the best model to fit excessive zero counts. Table 2 showed that the statistics of goodness of fit of four count outcomes models. The log likelihood of ZINB model was larger than that of NB model by 3580. ZINB model was better than NB model based on likelihood ratio test considering that the difference of log likelihood obeys chi-square distribution. The Poisson regression model fitted worst. ZINB model showed the best goodness of fit, with the largest log likelihood and the smallest AIC statistic and BIC statistic.

Figure 1 showed the predictive probability distribution curves of four models and the observed proportions of 19 observations. It was clear that the Poisson regression model fitted worst, in which the predictive probabilities of almost all counts were significantly different from the observed proportions. ZIP model was a little better than Poisson model. ZIP model predicted zero count accurately, but the predictive probabilities of other counts were terribly inconsistent with the observed proportions. The

Table 2. Statistics of goodness of fit of the four regression models

Models	Log likelihood	AIC statistic	BIC statistic
Poisson	-243 314	486 665	486 831
ZIP	-186 623	373 318	373 651
NB	-171 099	342 236	342 412
ZINB	-167 519	335 112	335 455

AIC, Akaike's information criterion; BIC, Bayesian information criterion; NB, negative binomial regression; ZIP, zero-inflated Poisson regression; ZINB, zero-inflated negative binomial regression.

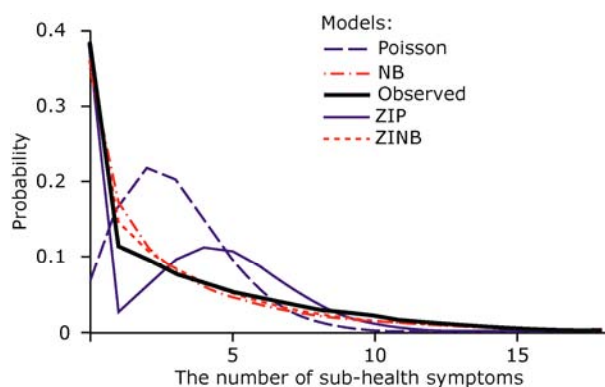


Figure 1. Predictive probability distribution curves of four models and the observed proportions of 19 observations. The predictive probabilities for most counts in ZINB model fitted the observed counts best.

predictive probabilities for zero count in the NB model and ZINB model were very close to the observed proportions. The predictive probabilities for most counts in ZINB model fitted the observed counts very well, except that the predictive probability of count 1 was a little bigger than the observed count. The NB model fitted the count outcomes a little worse than ZINB model.

The result of random sample with one-tenth of respondents was consistent with that of the total population: ZINB model showed the best goodness of fit with the largest log likelihood and the smallest AIC, BIC (Table 3). Based on the alpha dispersion parameter, over-dispersion *O* test, Vuong test, statistics of the goodness of fit, and the predictive probabilities of counts, ZINB model was an optimum model fitting the number of sub-health symptoms.

Results of ZINB model

Regression coefficients of the potential determinants of ZINB model were shown in Table 4. Binary section on the left side was the model for zero count dataset. Age, sex, occupation, smoker, alcohol drinker, ethnicity and obesity were determinants for whether encountering sub-health symptoms or not. Elderly, female, regular smoker, alcohol drinker, Yi-nationality, Hui-nationality and other-nationality were susceptible to suffer one or more sub-health symptoms ($P < 0.05$). In addition, mental labors ($P < 0.001$), Korean ($P < 0.001$), Miao-nationality ($P < 0.001$), Tujia-nationality ($P < 0.001$) and obese people ($P = 0.020$) were more insensitive to sub-health symptom. The NB section on the right side showed that sex, occupation, smoking, alcohol drinking, ethnicity, marital status and obesity had significant effect on the severity of sub-health status. Female, mental laborer, regular smoker, alcohol drinker, Yi-nationality, Korean, Mongolian, Hui-nationality, widowed or divorced, single and obese people reported more sub-health symptoms. However mental laborer, Miao-nationality and Tujia nationality respondents had fewer sub-health symptoms.

Table 3. Statistics of the goodness of fit of regression models in the randomly selected sample ($n = 7830$)

Models	Log likelihood	AIC	BIC
Poisson	-24 136	48 307	48 433
ZIP	-18 550	37 172	37 423
NB	-17 002	34 042	34 174
ZINB	-16 639	33 353	33 610

AIC, Akaike's information criterion; BIC, Bayesian information criterion; NB, negative binomial regression; ZIP, zero-inflated Poisson regression; ZINB, zero-inflated negative binomial regression.

Table 4. ZINB regression coefficients of the potential determinants of ZINB model for the number of sub-health symptoms

Potential determinants	Binary section					Negative binomial section				
	β	Z	P	95% CI for β		β	Z	P	95% CI for β	
Age	-0.029	-17.740	<0.001	-0.032	-0.025	0.000	0.040	0.972	-0.001	0.001
Female	-0.410	-17.550	<0.001	-0.455	-0.364	0.194	18.230	<0.001	0.173	0.215
Hypertension	0.039	1.070	0.284	-0.032	0.110	-0.008	-0.690	0.492	-0.031	0.015
Mental laborer	0.213	6.310	<0.001	0.147	0.280	-0.090	-7.490	<0.001	-0.114	-0.067
Smoking	-0.206	-5.010	<0.001	-0.287	-0.126	0.084	5.770	<0.001	0.055	0.112
Alcohol drinking	-0.383	-9.670	<0.001	-0.461	-0.305	0.066	4.800	<0.001	0.039	0.093
Yi nationality	-1.357	-17.830	<0.001	-1.506	-1.208	0.125	7.110	<0.001	0.091	0.160
Tibetan	0.027	0.420	0.677	-0.100	0.155	-0.004	-0.130	0.897	-0.062	0.055
Korean	0.599	11.570	<0.001	0.498	0.701	0.138	5.430	<0.001	0.088	0.187
Mongolia	-0.031	-0.590	0.557	-0.134	0.072	0.049	2.200	0.028	0.005	0.094
Miao nationality	0.449	5.560	<0.001	0.291	0.608	-0.144	-3.530	0.000	-0.224	-0.064
Hui nationality	-0.425	-9.260	<0.001	-0.515	-0.335	0.144	8.100	<0.001	0.109	0.178
Tujia nationality	0.293	4.060	<0.001	0.152	0.434	-0.084	-2.580	0.010	-0.148	-0.020
Other nationality	-0.178	-2.390	0.017	-0.324	-0.032	0.048	1.500	0.133	-0.014	0.110
Widowed or divorced	-0.099	-1.140	0.256	-0.269	0.072	0.110	5.040	<0.001	0.067	0.153
Single	-0.072	-1.550	0.122	-0.163	0.019	0.040	2.450	0.014	0.008	0.073
Obesity	0.107	2.330	0.020	0.017	0.197	0.050	3.160	0.002	0.019	0.080
Alpha	0.618	66.620	<0.001	0.600	0.636					

DISCUSSION

Count outcome data are frequently used in medical studies, and often inappropriately treated as continuous or categorical data. Over-dispersion and terribly skewed distribution reduce the utility of linear regression, therefore count outcomes should not be considered as continuous data. In the logistic model, all respondents in this study were dichotomized as two subsets: without any sub-health symptoms or with one or more sub-health symptoms. As a result, respondents with sub-health symptoms were treated with the same method regardless of the number of sub-health symptoms. So logistic regression model can only be used to explore the determinants of sub-health incidence rate but not the severity of sub-health status. Since categorizing count data would lead to loss of some useful information, logistic regression was not an appropriate model for count outcome study.^{11,12}

In general, traditional Poisson regression and NB regression were common models for count outcomes. But the strict requirement for variance equaling to the mean value causes the over-dispersed count data hardly to follow Poisson distribution. Both *O* test and the alpha dispersion parameters demonstrated over-dispersion of the sub-health data in this study. Statistics of the goodness

of fit and the predictive probability curves confirmed that neither Poisson model nor ZIP model was appropriate for the study of sub-health symptoms.

With the error item of gamma function, NB distribution takes the over-dispersion into account.²² However, the excessive zero counts had bad effects on traditional Poisson regression and NB regression models, for these regression models do not fit data with so many zero counts. ZI models were introduced just for resolving both over dispersion and excessive zero counts. ZI models provide assessment of the determinants of sub-health severity but not just the presence or absence of sub-health status, because they can model the number of sub-health symptoms in a continuum instead of dichotomous outcome. To our best knowledge, this is the first study to explore the application of ZI models to sub-health symptoms based on a survey in a large-scale population.

This study found that ZIP model fitted the data worse than ZINB model, which may be because over-dispersion of the number of sub-health symptoms restricted the utility of ZIP model. *O* test, Vuong test, statistics of the goodness of fit, likelihood ratio test and predictive probability curve indicated that ZINB model was the best model for the number of sub-health symptoms with proximately two-fifths of zero counts. This result was consistent with the results of studies that ZINB model was the best model for

count outcomes.^{19,20,25-31} However, some other studies found that ZIP model was better than ZINB model because of less severity of over-dispersion in those study data.³²⁻³³ In addition, some studies considered ZIP model as a good model for count outcome data, but these studies did not explore the ZINB model and couldn't compare the goodness of fit between ZIP model and ZINB model.³⁴⁻³⁸

The number of sub-health symptoms had a wide range from 0 to 18, which aggravated its over-dispersion, so that both Poisson model and ZIP model fitted badly. In addition, two-fifth of zero counts weakened the goodness of fit of traditional NB model.

In ZINB models, gender, occupation, smoking status, alcohol drinking, ethnicity and obesity were found to have impact on both sub-health incidence and sub-health severity. Marital status was only associated with the number of sub-health symptoms. The widowed/divorced and the singles suffered from higher counts of sub-health symptoms compared to the married. The significance of these results remind us of the value of social aspects and their application in the health care practice of preventive, predictive and personalized medicine. The elderly were more susceptible to sub-health status, but we did not find the relationship between age and sub-health severity in this study, which needs further investigation in the future.

In conclusion, the goodness of fit tests and predictive probability curves produced the same finding that ZINB model was the optimum model fitting the number of sub-health symptoms with excessive zero counts. It can be used in evaluating sub-health status and the severity.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

REFERENCES

1. Medicine CAOC. The TCM clinical guidelines of suboptimal health status. Beijing: China Press Traditional Chinese Medicine, 2006.
2. Wang W, Yan Y. Suboptimal health: a new health dimension for translational medicine. Clin Transl Med 2012; 1: 28. doi: 10.1186/2001-1326-1-28.
3. Wang DX, Zhou HL. Health, disease and subhealth. Medicine and Society 2007; 20: 5-8. Chinese.
4. Chen QS, Wang SY, Jing CX, Dong XM, Chi GB, Zhu L. Evaluation on diagnostic criterion of sub-health with Delphi method. Chinese Journal of Public Health. 2003; 19: 1467-8. Chinese. doi: 10.3321/j.issn:1001-0580.2003.12.030.
5. Xu T, Han SM, Liu JT, Zhu GJ, Mao M. Comparison of sub-health status between Tibetan people and Han people. Chin Med J 2009; 89: 2671-4. Chinese. doi: 10.3760/cma.j.issn.0376-2491.2009.38.003.
6. Xu T, Liu JT, Han SM, Zhu GJ, Mao M. Analyses for risk factors of sub-health status with logistic model and binomial model-survey in Tibetan people. J Clin Rehabilitative Tis Engineer Res 2009; 13: 6597-600. Chinese. doi: 10.3969/j.issn.1673-8225.2009.33.043.
7. Zhu L, Wang SY, Fan CX, Xiao YJ, Ou CF. Logistic regression analysis of risk factors on subhealth of young teachers in institution of higher learning. Chin J Public Health. 2003; 19: 595-6. Chinese. doi: 10.3321/j.issn:1001-0580.2003.05.048.
8. Wang Y, Chen Q. A cross-sectional study on sub-health of students in a university in Guangzhou. Chin General Practice 2005; 8: 738-40. Chinese. doi: 10.3969/j.issn.1007-9572.2005.09.019.
9. Pang H, Wang P. Research on sub-health condition and influencing factors of middle school teacher in Xinjiang. China Sport Science and Technology 2008; 44: 47-50. Chinese. doi: 10.3969/j.issn.1002-9826.2008.05.008.
10. Nie XL, Xue Q, Lai MH, Chen J, Zhao XS, Luo R. Subhealth and its influential factors among civil servants in taxation department. Chinese Journal of Public Health 2010; 26: 634-5. Chinese.
11. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometric 2000; 56: 1030-9. doi: 10.1111/j.0006-341X.2000.01030.x.
12. Agresti A. Generalized Linear Models. In: Agresti A, editors. An introduction to categorical data analysis. New York: Chichester Wiley, 1996. p74-83.
13. Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. Stat Med 2002; 21:1461-9. doi:10.1002/sim.1088.
14. Solinas G, Campus G, Maida C, Sotgiu G, Cagetti MG, Lesaffre E. et al. What statistical method should be used to evaluate risk factors associated with dmfs index? Evidence from the National Pathfinder Survey of 4-year-old Italian children. Community Dent Oral Epidemiol 2009; 37: 539-46. doi: 10.1111/j.1600-0528.2009.00500.x.
15. Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, et al. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. Biometrics 2009; 65:1003-10. doi: 10.1111/j.1541-0420.2009.01223.x.
16. Wang H, Heitjan DF. Modeling heaping in self-reported cigarette counts. Stat Med 2008; 27: 3789-804. doi: 10.1002/sim.3281.
17. Sheu ML, Hu TW, Keeler TE, Ong M, Sung H Y. The effect

- of a major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model. *Health Economics* 2004; 13:781-91. doi: 10.1002/hec.849.
18. Denwood MJ, Stear MJ, Matthews L, Reid SW, Toft N, Innocent GT. The distribution of the pathogenic nematode *Nematodirus battus* in lambs is zero-inflated. *Parasitology* 2008; 135: 1225-35. doi: 10.1017/S003118-2008004708.
 19. 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. doi: 10.1111/j.1600-0528.2004.00155.x.
 20. Javali SB, Pandit PV. Using zero inflated models to analyze dental caries with many zeroes. *Indian J Dent Res* 2010; 21:480-5. doi: 10.4103/0970-9290.74210.
 21. Lambert D. Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; 34: 1-14. doi: 10.2307/1269547.
 22. Joseph M, Hilbe. Problems with zero counts. In: Joseph M, Hilbe. *Negative Binomial Regression*. London: Cambridge Univ Pr 2007: 173-77.
 23. Bohning D, Dietz E, Schlattmann P. Zero-inflated count models and their applications in public health and social science. In: Rost J, Langeheine R eds. *Applications of latent trait and latent class models in the Social Sciences*. 1st ed. Germany: Munster Waxmann 1997: 333-44.
 24. Vuong QH. Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica* 1989; 57: 307-33. doi: 10.2307/1912557.
 25. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *J Biopharm Stat* 2006; 16: 463-81. doi: 10.1080/10543400600719384.
 26. Akram K, Pedersen-Bjergaard U, Carstensen B, Borch-Johnsen K, Thorsteinsson B. Frequency and risk factors of severe hypoglycaemia in insulin-treated Type 2 diabetes: a cross-sectional survey. *Diabet Med* 2006; 23:750-6. doi: 10.1111/j.1464-5491.2006.01880.x.
 27. Zaninotto P, Falaschetti E. Comparison of methods for modelling a count outcome with excess zeros: application to Activities of Daily Living (ADL-s). *J Epidemiol Community Health* 2011; 65:205-10. doi: 10.1136/jech.2008.079640.
 28. Turner AN, Miller WC, Padian NS, Kaufman JS, Behets FM, Chipato T, et al. Unprotected sex following HIV testing among women in Uganda and Zimbabwe: short- and long-term comparisons with pre-test behaviour. *Int J Epidemiol* 2009; 38:997-1007. doi: 10.1093/ije/dyp171.
 29. Carrel M, Voss P, Streatfield PK, Yunus M, Emch M. Protection from annual flooding is correlated with increased cholera prevalence in Bangladesh: a zero-inflated regression analysis. *Environ Health* 2010; 9:13. doi: 10.1186/1476-069X-9-13.
 30. Lacruz ME, Emeny RT, Haefner S, Zimmermann AK, Linkohr B, Holle R, et al. Relation between depressed mood, somatic comorbidities and health service utilisation in older adults: results from the KORA-Age study. *Age Ageing* 2012; 41: 183-90. doi: 10.1093/ageing/afr162.
 31. Meszaros ZS, Dimmock JA, Ploutz-Snyder RJ, Abdul-Malak Y, Leontiera L, Canfield K, et al. Predictors of smoking severity in patients with schizophrenia and alcohol use disorders. *Am J Addict* 2011; 20: 462-7. doi: 10.1111/j.1521-0391.2011.00150.x.
 32. Bandiera FC, Arheart KL, Caban-Martinez AJ, Fleming LE, McCollister K, Dietz NA, et al. Secondhand smoke exposure and depressive symptoms. *Psychosom Med* 2010; 72:68-72. doi: 10.1097/PSY.0b013e3181c6c8b5.
 33. Slymen DJ, Ayala GX, Arredondo EM, Elder JP. A demonstration of modeling count data with an application to physical activity. *Epidemiol Perspect Innov* 2006; 3: 3. doi: 10.1186/1742-5573-3-3.
 34. Karazsia BT, van Dulmen MH. Regression models for count data: illustrations using longitudinal predictors of childhood injury. *J Pediatr Psychol* 2008; 33:1076-84. doi: 10.1093/jpepsy/jsn055.
 35. Bergemann TL, Huang Z. A new method to account for missing data in case-parent triad studies. *Hum Hered* 2009; 68:268-77. doi: 10.1159/000228924.
 36. Ceppi M, Biasotti B, Fenech M, Bonassi S. Human population studies with the exfoliated buccal micronucleus assay: statistical and epidemiological issues. *Mutat Res* 2010; 705: 11-9. doi: 10.1016/j.mrrev.2009.11.001.
 37. Zhang HJ, Min J, Wang P, et al. A statistical method to extra zero in the data of field survey. *Acta Universities Medicinalis Nanjing* 2007; 27:634-636. doi:10.3969/j.issn.1007-4368.2007.06.035.
 38. Marioni RE, Matthews FE, Brayne C. The association between late-life cognitive test scores and retrospective informant interview data. *Int Psychogeriatr* 2011; 23: 274-9. doi: 10.1017/S1041610210001201.