RESEARCH ARTICLE

Editorial Process: Submission:04/14/2023 Acceptance:08/14/2023

Using Multilevel Negative Binomial Modeling to Detect Active Smoking in Colorectal Cancer Screening

Nittaya Phuangrach[1], Pongdech Sarakarn[2,3]*

Abstract

Background: Multilevel analysis, in several forms, has been extensively utilized over the past few decades. While utilizing for colorectal cancer (CRC) screening may be unclear, especially at community level. The study aimed to explain the use of multilevel negative binomial analysis, developed as a practical guide through data obtained from a study of CRC screening in Thailand. **Method:** We analyzed the data of 2,475 fecal immunochemical test (FIT) cases in treatment arms from a population-based randomized controlled trial for CRC screening in the Khon Kaen province of Thailand. We summarized the statistical methodology, highlighting the advantages and disadvantages of data analysis using a multilevel negative binomial method compared with a standard negative binomial approach based on the data obtained in the randomized controlled trial for CRC screening; where active smoking and fecal hemoglobin (f-Hb) concentration were considered as the main exposure and outcome, respectively. **Results:** Our findings showed differences of significant value and magnitude in the effects of both methods. Active smoking was statistical significantly with an f-Hb concentration IRR_{adj} = 1.47 (95%, CI: 1.01-2.14) through the use of the standard negative binomial method, whereas the multilevel negative binomial approach produced a non-statistical significance of IRR_{adj} = 1.30 (95%, CI: 0.89-1.90). **Conclusion:** Utilizing a standard statistical approach in CRC screening, the data analyzed were equal to zero. Hierarchical data, based on contextual factors and using a multilevel modeling approach, must be addressed. The f-Hb concentration, occurred over-dispersion, which implies that further studies utilize over-dispersion for improved appropriate statistical analysis.

Keywords: Multilevel- negative binomial- colorectal cancer screening

Asian Pac J Cancer Prev, **24 (8)**, 2823-2827

Introduction

Multilevel analysis has been extensively utilized over the past few decades in diverse fields; from health and social sciences to econometrics (StataCorp, 2013); including multidisciplinary design and analytic approaches to advance prospective research on the multilevel determinants of child health (Johnson et al., 2017), dental caries in 12-year-old schoolchildren: a multilevel analysis of individual and school environment factors in Goiânia (Oliveira et al., 2015), regression models for mixed over-dispersed poisson and continuous clustered data: modeling BMI and the number of cigarettes smoked per day (Atem et al., 2012), and the application of generalized linear latent and mixed models in epidemiology: repeated measures, measurement error, and multilevel modeling (Anders and Sophia, 2003). Multilevel is a mixed model analysis that investigates the relationship between individuals and social groups that are conceptualized as a hierarchical system of individuals nested within groups, with individuals and groups defined at separate levels of this hierarchical system, where variables may be defined at each level (Hox, 2010). Health science research, specifically in cancer screening, presented a zero outcome with a variance greater than the mean, resulting in over-dispersion. A Poisson estimator handles over- or under-dispersion by moving away from the complete distributional specification to one within the first two moments. Alternatively, one can specify a distribution that permits more flexible modeling of the variance than the Poisson, which is the standard parametric model to account for over-dispersion; the negative binomial (Cameron and Trivedi, 1998). Although the standard negative binomial model is used to model over-dispersed count data, for which the variance is greater than that of a Poisson model. A multilevel negative binomial is an extension of the standard negative binomial model that incorporates normally distributed random effects at different hierarchical levels with mixed-effect negative binomial regression containing both fixed and random

*[1]Ph.D. Candidate in Epidemiology and Biostatistics, Faculty of Public Health, Khon Kean University, Khon Kaen, Thailand. [2]ASEAN Cancer Epidemiology and Prevention Research Group, Khon Kaen University, Khon Kaen, Thailand. [3]Department of Epidemiology and Biostatistics, Faculty of Public Health, Khon Kaen University, Khon Kaen, Thailand. *For Correspondence: spongd@kku.ac.th*

effects (StataCorp, 2013). We applied multilevel modeling with other standard statistics dependent on each type of outcome (Rabe-Hesketh and Skrondal, 2012). In CRC screening with the fecal immunochemical test (FIT) method, we found that f-Hb concentration occurred with over-dispersion. This study's goal was to explain how to use the multilevel negative binomial analysis and demonstrate a practical guide using data from a CRC screening study in Thailand.

Materials and Methods

Data

This study utilized secondary CRC screening data at the Nam Phong district, Khon Kaen province Thailand through FIT with 2,475 cases in treatment arms from a population-based randomized controlled trial for CRC screening (Sarakarn et al., 2017). The secondary data was made available by the corresponding author, as head of the Asian Cancer Epidemiology and Prevention group (ACEP). Ethical consideration was approved by the Institutional Review Board of Health Science Research, Khon Kaen University (HE641437), Khon Kaen, Thailand.

Variables

Main outcome variable

CRC is the development of cancer from the colon or rectum, in which more than 80% of CRC cases arise from adenomatous polyps and is the third major type of cancer worldwide affecting both sexes (GLOBOCAN, 2022). Screening for this cancer is effective not only for early detection but also for prevention (Cunningham et al., 2010). Screening via a colonoscopy is limited, as is more expensive than FIT. CRC screening, therefore, was conducted through FIT in a population-based randomized controlled trial. The procedures employed to collect FIT analyzed the quantitative human hemoglobin content of each stool specimen collected and was measured in a laboratory using an OC-Sensor (Sarakarn et al., 2017). Some studies cut-off a hemoglobin concentration ≥100 ng/mL is considered as a positive FIT which is cases group and the negative FIT group with the hemoglobin concentration less than 100 ng/mL is considered as controls and using multiple logistic regression analyzed the data (Pramual et al., 2018) which this method leading to information is loss of the real data. Then in this study the FIT method identified f-Hb concentration as count data that start from minimum = 0 ng/mL to maximum = 6,055 ng/mL and was the main outcome of this study.

Main effect variable

Smoking is a major problem in public health. Studies have shown an increased incidence of CRC in smokers (István et al., 2016), and have determined a significant association of CRC in people who have smoked for more than 30 years, and more than 20 cigarettes per day (Cleary et al., 2010; Hannan et al., 2009). The association between smoking and CRC found that men exposed to cigarette smoke are more likely to develop cancer than women. However, the association between smoking and CRC remains unclear, as some studies have shown a positive relationship between smoking and the incidence of CRC, whereas others have not (Tsoi et al., 2009). Our study employs secondary data that defines active smoking as a self-imposed task, lighting the cigarette or tobacco product and inhaling the fumes, drawing them deep into one's lungs, and leaving deadly, toxic residues that will eventually lead to health issues. The secondary data herein measured active smoking via a questionnaire that classified two categories (0 and 1). The (0) category contained non-smokers, whereas the (1) category consisted of current smokers, as well as those who have smoked in the past but no longer smoke. We then applied statistical methods appropriate for the cluster data gathered for this project, where active smoking was the main effect.

Cluster variable

We considered data clusters within the modeling factors related to CRC screening and found that most researchers had not considered data structures as a hierarchical system in CRC screening. The Nam Phong district in Khon Kaen province covers 12 sub-districts that include hierarchical variables; such as 18 primary care units (PCU), with different public health officers, villages, populations, community stores, and hospital's distant. The PCUs, often distant from the villages and communities, were responsible to make contact with the participants, collect data, and provide follow up during the entirety of the study (Sarakarn et al., 2017).

Statistical analysis

Descriptive results on individual-level and primary care unit-level, data were summarized with univariate statistical means and the standard deviation for continuous variables, and n and % for categorical variables.

Using negative binomial and multilevel negative binomial for analyzing data, a crude analysis depicted the association of each independent variable and f-Hb concentration in the multilevel model. Variables selection was determined by the p-value of the association 0.25 (Hosmer and Lemeshow, 2000) in clinical or epidemiological CRC screening. All values of $p > 0.25$ were retained in the final model, which created the initial multivariable of the multilevel model. The multicollinearity of the model was confirmed through the variance inflation factors (VIF) > 10 or a tolerance close to 0 and had a significantly high association (Kleinbaum et al., 1998). The high correlation variables were dropped from the multilevel model. The likelihood of over-dispersion was tested (Long, 1997). We then selected the best model by considering the p-value of the association with wald statistics, employing backward elimination to delete the model variables based on the goodness of fit by chi-square. Data were further analyzed via the STATA program version 14. The effects of the standard negative binomial regression model and multilevel negative binomial regression model presented an adjusted incidence rate ratio (IRR_{adj}) and 95% CI ($p < 0.05$). Lastly, we considered the ecological fallacy that often leads to erroneous conclusions between a standard negative binomial regression model and the multilevel negative binomial regression model based on unbiased evidence (Hox, 2010).

Results

The research subjects consisted of 2,475 presented mean of fecal hemoglobin concentration = 86.83 (sd = 448.86) median = 7 (minimum = 0, maximum = 6,055) and box plot showed effect of the outliers leading to fecal hemoglobin concentration occurred over-dispersion (Figure 1), subjects 2,475 consisted of females (66.83%), aged 45-59 years old (55.19%). Some were not educated (1.05%), not working (10.42%), married (82.02%), and had a family income of less than 5,000 baht per month (51.72%). Among their personal characteristics were smoking (25.13%) and drinking; specifically, beer (46.99%), wort (20.24%), rice whisky (24.57%), brandy (16.53%), and other alcohol (13.25%), as well as soft drinks (70.95%), tea (29.25%), and coffee (54.59%).

The participants had a BMI < 26 (64.61%), as well as current household health problems; including a father or mother with cancer (10.91% and 7.39%, respectively), or a brother (8.36%) or sister (4.97%) with cancer. The subjects (2,475) also presented diabetes mellitus (13.45%).

When considering factors in cluster variable, we found that public health officer per PCU < 10 person = 90.98%, community per PCU ≥ 8 community = 59.31%, population per PCUs ≥ 4,001 person = 82.66%, community store per PCUs ≥ 31 stores = 61.82% and hospital distance from PCUs ≥ 21 km = 52.77%.

The negative binomial model demonstrated that smoking affected the f-Hb concentration [IRR_{adj} = 1.47 (95% CI: 1.01-2.14)]; however, when compared with the multilevel negative binomial model, smoking was deemed insignificant with the f-Hb concentration [IRR_{adj} = 1.30 (95% CI: 0.89-1.90)] (Table 1). Conversely, Tsoi performed a meta-analysis on the basis of prospectively published data to examine the association between smoking and incidences of CRC. The gender difference, dose responses, and CRC were assessed. We found that current smokers

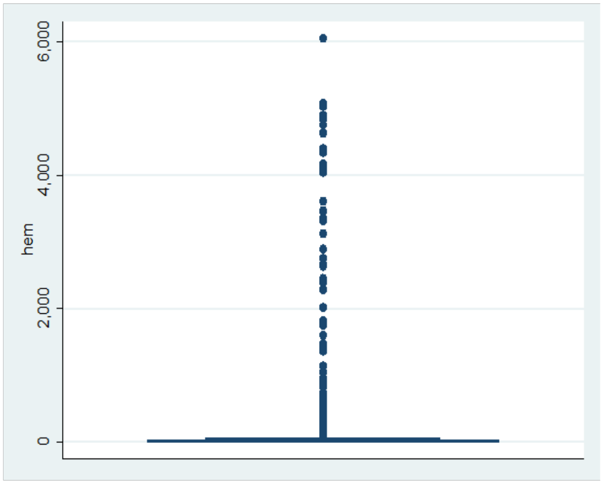

Figure 1. Characteristic of Fecal Hemoglobin Concentration.

Table 1. Risk factors at individual Levels

Independent variables	Negative Binomial			Multilevel Negative Binomial		
	IRR_{adj}	95%CI	P	IRR_{adj}	95%CI	P
Sex						
Male	0.86	0.61-1.62	0.4	0.86	0.61-1.21	0.4
Age						
≥60	1.66	1.32-2.09	<0.0001	1.57	1.24-1.98	<0.0001
Occupation						
Not working	0.86	0.67- 1.10	0.24	0.85	0.66-1.09	0.22
Marital status						
Divorced	1.22	0.93-1.61	0.14	1.26	0.95-1.67	0.1
Family income						
≤ 5000 baht/month	1.12	0.90-1.40	0.28	1.1	0.88-1.37	0.38
Smoking						
Yes	1.47	1.01-2.14	0.04	1.3	0.89-1.90	0.16
Drinking beer						
Yes	0.69	0.53-0.90	0.006	0.7	0.54-0.92	0.01
Drinking rice whisky						
Yes	1.52	1.09-2.12	0.014	1.63	1.16-2.28	0.004
Brother with cancer						
Yes	1.62	1.10-2.40	0.014	1.62	1.09-2.40	0.015
Sister with cancer						
Yes	1.5	0.94-2.38	0.084	1.42	0.89-2.26	0.131
With DM						
Yes	0.73	0.54-0.98	0.039	0.76	0.56-1.03	0.087

Table 2. Risk factors in the primary Care Unit Level.

Independent variables	Negative Binomial			Multilevel Negative Binomial		
	$IRR._{adj}$	95% CI	P	$IRR._{adj}$	95% CI	P
Public health officers						
per PCU						
< 10	1.19	0.95-1.48	0.123	1.2	0.77-1.88	0.416
Community						
per PCU						
≥ 8	1.04	0.71-1.52	0.81	1.03	0.70-1.52	0.845
Population						
per PCU						
≥4001	1.12	0.87-1.43	0.353	1	0.78-1.28	0.976
Community stores						
per PCU						
≥ 31	1.45	1.02-2.09	0.04	1.45	0.91-2.31	0.112
Hospital distance						
from a PCU						
≥ 21 km	0.98	0.73-1.31	0.894	1.16	0.78-1.73	0.449

showed a modestly higher risk of colorectal cancer [RR_{adj} = 1.20 (95% CI: 1.10-1.30) than nonsmokers (Tsoi et al., 2009).

An assessment of the physiological characteristics revealed that related f-Hb concentrations within the results of diabetes mellitus in CRC screening were unclear, in which the multilevel negative binomial model showed IRR_{adj} = 0.76 (95% CI: 0.56-1.03) (Table 1). The results differed from the effects of diabetes mellitus on CRC prognosis for overall survival using the random effect model HR_{adj} = 1.18 (95% CI: 1.12-1.24), I^2 = 64.8%, P for heterogeneity < 0.001. However, the sensitivity analysis of the overall pooled HR on the effects of DM on CRC prognosis produced the lowest HR_{adj} of overall survival at 1.18 (95% CI: 1.12-1.24), and the highest HR_{adj} of 1.38 (95% CI: 1.31-1.46) (Zhu et al., 2017).

The cluster data of eighteen primary care units (PCUs) identified different subjects who had received health services, selected based on the population and area in each PCU. In this study, each cluster differed by the number of public health officers, communities or villages, populations, community stores, and the hospital's distance from a PCU. The negative binomial model found that community stores presented a risk factor [IRR_{adj} = 1.45 (95% CI: 1.02-2.09], however, the multilevel negative binomial model found that community stores were not significant [IRR_{adj} = 1.45 (95% CI: 0.91-2.31)] (Table 2). The Center for Disease Control and Prevention analyzed the extent of point-of-purchase tobacco advertising and marketing found in various types of stores. Generalized estimating equations were used to analyze the data with a logit link function. Overall, 2,999 stores observed were tobacco retailers and were eligible for inclusion in the study compared with smaller grocery stores [OR_{adj} = 9.5 (95% CI: 4.7-19.2)], which were more likely to have high-intensity exterior tobacco advertising (Centers for Disease Control and Prevention C, 1999).

Discussion

Multilevel problems have led to methods of analysis that aggregate or disaggregate all variables to a single level of interest statistics. However, assessing variables from several levels at a single common level is insufficient, resulting in two common issues. The first issue was that too much information was lost, reducing the power of statistical analysis. When data is disaggregated, however, employing a higher number of disaggregated cases for the sample size results in significance tests that reject the null hypothesis (Hox, 2010); such as the effects of active smoking and community stores. The second problem, due to the ecological fallacy known as the 'Robinson effect', as well as 'Simpson's paradox', refers to the conclusions drawn through grouped data of a single heterogeneous population, that are then collapsed and analyzed (Hox, 2010) such as the diabetes mellitus analyzed with the negative binomial model indicated a protective factor [IRR_{adj} = 0.73 (95% CI: 0.54-0.98)], however, when compared with the multilevel negative binomial model, diabetes mellitus was not significant [IRR_{adj} = 0.76 (95% CI: 0.56-1.03)] that the results showed different direction from significant to non-significant when control with the cluster variable we call this situation is 'Simpson's paradox'. However, when considered the magnitudes of effect which is incident rate ratio of the diabetes mellitus, we found that IRR_{adj} change around 5% from protective factor convergent to risk factor we call this situation is 'Robinson effect' which the result of diabetes mellitus of the negative binomial modeling occurred ecological fallacy that leads to erroneous conclusions.

The results of the negative binomial and multilevel negative binomial models may seem to be similar, however, when compared with an adjusted incidence rate ratio and 95% CI, we found the multilevel negative binomial to have more precision than that of the negative binomial. Moreover, many of the advantages of multilevel models over traditional methods are at the expense of greater model complexity. If the model is accurate, multilevel estimates are then less biased and more efficient than those obtained using other methods (Diez-Roux, 2000). Our study revealed limitations, as the multilevel models were less parsimonious and required larger data sets. Sample size and power calculations for multilevel hypothesis testing are particularly complex. Effectiveness depends both on the number of groups and the number of individuals per group (Diez-Roux, 2000). Realistically, mixed models will have more effect than traditional methods. However, the use of secondary data presents data limitations, particularly in cluster variables, that will be inappropriate and provide an incorrect assessment of the differences between multilevel modeling and traditional regression. We measured the active smoking characteristics in the Nam Phong district, Khon Kaen province, Thailand, which represents part of a community or PCU level that will define differences in the contextual data.

Therefore, when the data analyzed is equal to zero and the outcomes produce count data that occurred through over-dispersion, the standard statistical approach, such as Poisson regression may be inappropriate, especially in CRC screening. Hierarchical data based on contextual factors through a multilevel modeling approach should be further evaluated.

Author Contribution Statement

NP was the principal investigator involved in the conceptualization, methodology, formal analysis, visualization, and preparation of the original draft. PS acted as a supervisor involved in the conceptualization, methodology, statistical analyses, and editing of the manuscript.

Acknowledgements

The authors would like to thank the Asian Cancer Epidemiology and Prevention group (ACEP) for their supported.

Funding statement

The authors received no specific funding for this work.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

Anders S, Sophia R-H (2003). Some applications of generalized linear latent and mixed models in epidemiology: Repeated measures, measurement error, and multilevel modeling. *Norsk Epidemiologi*, **13**, 265-78.

Atem F, Ngwa J, Adeniji A (2012). Regression Models for Mixed Over-Dispersed Poisson and Continuous Clustered Data: Modeling BMI and Number of Cigarettes Smoked Per Day. *J Modern Appl Statist Methods*, **11**, 19.

Cameron C, Trivedi P (1998). Regression analysis of count data. Cambridge University Press, Cambridge. pp 70-7.

Centers for Disease Control and Prevention C (1999). Point-of-purchase tobacco environments and variation by store type--United States. *MMWR Morb Mortal Wkly Rep*, **51**, 184-7.

Cleary SP, Cotterchio M, Shi E, Gallinger S, Harper P (2010). Cigarette smoking, genetic variants in carcinogen-metabolizing enzymes, and colorectal cancer risk. *Am J Epidemiol*, **172**, 1000-14.

Cunningham D, Atkin W, Lenz HJ, et al (2010). Colorectal cancer. *Lancet*, **375**, 1030-47.

Diez-Roux AV (2000). Multilevel analysis in public health research. *Annu Rev Public Health*, **21**, 171-92.

GLOBOCAN. Epidemiology of colorectal cancer. Retrieved http://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf.

Hannan LM, Jacobs EJ, Thun MJ (2009). The association between cigarette smoking and risk of colorectal cancer in a large prospective cohort from the United States. *Cancer Epidemiol Biomarkers Prev*, **18**, 3362-7.

Hosmer D, Lemeshow S (2000). Applied logistic regression. Wiley, New York, pp 92-5.

Hox JJ (2010). Multilevel analysis: techniques and applications. Routledge, New York. pp 1-4.

István DM, Cristian B, Árpád T, et al (2016). The Role of Smoking in the Development of Colorectal Cancer. *Acta Medica Marisiensis*, **62**, 400-2.

Johnson SB, Little TD, Masyn K, Mehta PD, Ghazarian SR (2017). Multidisciplinary design and analytic approaches to advance prospective research on the multilevel determinants of child health. *Ann Epidemiol*, **27**, 361-70.

Kleinbaum DKL, Muller K, Nizam A (1998). Applied regression analysis and other multivariable methods. Pacific Grove: Duxbury Publishing. pp 240-52.

Long J (1997). Regression models for the categorical and limited dependent variables. SAGE, California. pp 236-7.

Oliveira LB, Moreira Rda S, Reis SC, Freire Mdo C (2015). Dental caries in 12-year- old schoolchildren: a multilevel analysis of individual and school environment factors in Goiânia. *Rev Bras Epidemiol*, **18**, 642-54.

Pramual P, Sarakarn P, Kamsa-ard S, et al (2018). Lack of Association between Red Meat Consumption and a Positive Fecal Immunochemical Colorectal Cancer Screening Test in Khon Kaen, Thailand: a Population- Based Randomized Controlled Trial. *Asian Pac J Cancer Prev*, **19**, 271-8.

Rabe-Hesketh S, Skrondal A (2012). Multilevel and Longitudinal Modeling Using Stata. Stata Press, Texas. pp 499-861.

Sarakarn P, Promthet S, Vatanasapt P, et al (2017). Preliminary Results: Colorectal Cancer Screening Using Fecal Immunochemical Test (FIT) in a Thai Population Aged 45-74 Years: A Population-Based Randomized Controlled Trial. *Asian Pac J Cancer Prev*, **18**, 2883-9.

StataCorp (2013). Stata multilevel mixed effects reference manual release 13.

Statistical Software. College Station TX, StataCorp LP. pp 1-36.

StataCorp (2013). Stata multilevel mixed effects reference manual release 13.

Statistical Software. College Station TX, StataCorp LP. pp 125-40.

Tsoi KK, Pau CY, Wu WK, et al (2009). Cigarette smoking and the risk of colorectal cancer: a meta-analysis of prospective cohort studies. *Clin Gastroenterol Hepatol*, **7**, 682-8.

Zhu B, Wu X, Wu B, et al (2017). The relationship between diabetes and colorectal cancer prognosis: A meta-analysis based on the cohort studies. *PLoS One*, **12**, e0176068.