BMC Medical Research Methodology

RESEARCH Open Access

Count data models for outpatient health services utilisation

Nurul Salwana Abu Bakar[1*], Jabrullah Ab Hamid[2], Mohd Shaiful Jefri Mohd Nor Sham[3], Mohd Nor Sham[3] and Anis Syakira Jailani[4]

Abstract

Background Count data from the national survey captures healthcare utilisation within a specific reference period, resulting in excess zeros and skewed positive tails. Often, it is modelled using count data models. This study aims to identify the best-fitting model for outpatient healthcare utilisation using data from the Malaysian National Health and Morbidity Survey 2019 (NHMS 2019) and utilisation factors among adults in Malaysia.

Methods The frequency of outpatient visits is the dependent variable, and instrumental variable selection is based on Andersen's model. Six different models were used: ordinary least squares (OLS), Poisson regression, negative binomial regression (NB), inflated models: zero-inflated Poisson, marginalized-zero-inflated negative binomial (MZINB), and hurdle model. Identification of the best-fitting model was based on model selection criteria, goodness-of-fit and statistical test of the factors associated with outpatient visits.

Results The frequency of zero was 90%. Of the sample, 8.35% of adults utilized healthcare services only once, and 1.04% utilized them twice. The mean-variance value varied between 0.14 and 0.39. Across six models, the zero-inflated model (ZIM) possesses the smallest log-likelihood, Akaike information criterion, Bayesian information criterion, and a positive Vuong corrected value. Fourteen instrumental variables, five predisposing factors, six enablers, and three need factors were identified. Data overdispersion is characterized by excess zeros, a large mean to variance value, and skewed positive tails. We assumed frequency and true zeros throughout the study reference period. ZIM is the best-fitting model based on the model selection criteria, smallest Root Mean Square Error (RMSE) and higher R2. Both Vuong corrected and uncorrected values with different Stata commands yielded positive values with small differences.

Conclusion State as a place of residence, ethnicity, household income quintile, and health needs were significantly associated with healthcare utilisation. Our findings suggest using ZIM over traditional OLS. This study encourages the use of this count data model as it has a better fit, is easy to interpret, and has appropriate assumptions based on the survey methodology.

Keywords Healthcare utilisation, Outpatient, Count model, Zero-inflated model, Health behavioral model

*Correspondence:
Nurul Salwana Abu Bakar
salwana.ab@moh.gov.my
[1]Centre for Health Policy Research, Institute for Health Systems Research, National Institutes of Health, Ministry of Health, Shah Alam, Malaysia
[2]Centre for Health Equity Research, Institute for Health Systems Research, National Institutes of Health, Ministry of Health, Shah Alam, Malaysia
[3]Centre for Health Economics Research, Institute for Health Systems Research, National Institutes of Health, Ministry of Health, Shah Alam, Malaysia
[4]Centre for Health Outcome Research, Institute for Health Systems Research, National Institutes of Health, Ministry of Health, Shah Alam, Malaysia

Background

Count data arising from measuring health care utilisation is a common outcome in health research, especially from national survey data. Count refers to the number of times an event occurs [1] and usually exhibits a skewed distribution. Often times, the presence of excess zeros and long positive tails leads to data dispersion. With this distinct characteristic, the assumption of normality is violated and a need for probability distribution approaches to handle the dispersions [2].

Count-valued outcomes are typically modeled using discrete distributions, such as the Poisson or negative binomial distributions [3]. In such cases, the data could either be unmodified, zero-inflated or zero-truncated relative to the standard model (linear or logistic regression model), where flexible mixture distributions are often needed to accommodate the unique features of the data. Previous studies have compared the overall performance of count regression models (Poisson, negative binomial, and their zero-inflated and hurdle variants) in modeling an outcome variable with extra zeros [3],[4]. Marginalized zero-inflated model also commonly used in exploring health care data[5, 6]. Other types of model such as Waring regression allows to distinguish between unobserved heterogeneity [7], hyper-Poisson regression utilizes the mean of the regressors [8] as well as Generalized Poisson model caters for under-dispersion property [9] were explored but the characteristics of these model not of interest in this current study.

Several estimation approaches have been developed to address zero-modified count data such as the application of the zero-inflated model in the discipline of arts [10, 11] and two-part or hurdle models in healthcare [12, 13]. While these models vary in terms of their distributional assumptions and parametric forms, they incorporate an underlying two-part model: a logistic part for excess zeros and a count part catering for zero and non-zero observations [5, 6].

Healthcare utilisation refers to being in contact with a certified medical or health facility, involving the process of seeking care to prevent or treat health problems [14]. The utilisation of healthcare services is the result of a complex decision-making process with multiple determinants, often occurring from the interaction between contextual and individual factors. These include individual characteristics, access to health services, and organization of the healthcare system [15]. Although utilisation of healthcare services is primarily decided by the choices of patients, the factors leading to such decisions are not merely individual preferences, but more complex choices involving the institutional, socioeconomic, and cultural backgrounds of the individual.

The Andersen Healthcare Utilisation Model (Andersen's Behavioral Model of Health Services Use) is one of the most popular models of healthcare utilisation [16, 17]. Andersen's model focuses on the social and economic factors that determine the use of health care. This model explains that healthcare utilisation depends on various factors ranging from the propensity of individuals to use services, the ability of individuals to access services, and individual's health condition, each of which is represented by predisposing, enabling, and need factors. Predisposing characteristics are demographic variables that make some individuals more likely to use healthcare services than others. Enabling factors measure individuals' ability to access health care from an economic standpoint. Need variables include risk factors for diseases, individual health states, and experiences of diseases that lead to seeking medical assistance. Need factors are the strongest predictors of healthcare utilisation, followed by enabling and predisposing factors [18, 19]. This study uses variables based on the Andersen Behavioral Model to identify associated factors in outpatient health care utilisation in Malaysia.

This study aims to identify the best-fitting count model for outpatient health care utilisation using data from the Malaysian National Health and Morbidity Survey 2019. We estimated different models and used several model selection criteria to identify the best-fitting criteria. This study also identifies the factors of health care utilisation among adults in Malaysia, which are vital for healthcare planners and managers.

Methods

Data source

This study utilized data from the National Health and Morbidity Survey (NHMS) 2019, a cross-sectional household survey in Malaysia conducted every four years to gather community-based data for health care utilisation and needs. The NHMS uses a two-stage stratified cluster sampling method conducted through face-to-face interviews. Details of the survey method are described in the official report [20]. The survey captures details on sociodemographic, health status, health problems, household income, and utilisation patterns, including frequency, service provider, and payment sources. Adults aged 18 years and older were included in this study. This study captured outpatient visits in the last 14 days for both public and private facilities. From a total of 11,674 sampled populations, only 8.1% reported utilized outpatient care at least once. The short reference period led to two types of zero: true zero reflects non-users because they did not get sick during the reference period, while frequency zero reflects individuals who fell sick during the reference period but did not seek care.

Theoretical approaches and studied variables

This study utilized Andersen's health behavioral model [16] to determine the predisposing (demographic), enabling (personal/family), and predictive (perceived/evaluated) factors of seeking outpatient care with the availability of the best presented variables collected from the NHMS The independent variables were selected based on the Andersen model. The dependent variable was the frequency of outpatient visits, and the selection of instrumental variables was done accordingly. Total household income was log-transformed using Ln(X), and imputation was performed on missing values based on working status and education level of the same group stratification. Statistical analysis was performed using STATA 14 (Stata Corporation, College Station, TX). The 'svyset' command were used and weight estimation for the complex survey design based on the probability of sampling, the non-response and post-stratification adjustment by ethnicity, age, and gender [21].

Comparison of regression models

To explore the data, six regression models were used in this study. Initially, data was explored using ordinary least squares (OLS) as a common regression for healthcare utilisation analysis. The count data models considered in this study were Poisson regression, negative binomial (NB) regression, zero-inflated models (ZIM) such as zero-inflated Poisson (ZIP) and marginalized zero-inflated negative binomial (MZINB), and the hurdle or two-part model (probit and truncated at zero negative binomial).

OLS is a common and basic form of regression with a distinct assumption of normality. The most common evidence published using this National Health Morbidity dataset uses OLS for the analysis of continuous outcome variables [21]. However, with the healthcare utilisation concept, Poisson regression, a basic count model, is used. Poisson assumes equi-dispersion of mean and variance [1], while the NB model is equipped for a parameter to account for overdispersion. It is deemed to be a better model for an overdispersion variance to mean. Both, ZIM and hurdle model allow zero and positive counts, but cater to different decision-making processes [22]. In this dataset, the users of the outpatient healthcare service were based on a few assumptions specified by each model. For ZIM, we assumed that all patients had access to outpatient services and affordability was not an issue to obtain care in the Malaysian healthcare setting [23]. Thus, the occurrence of zero in this dataset was assumed to be a true zero, because a person is a non-user as he or she did not get sick within the study duration (sampling zeros). However, frequency zero represents a person who is sick but chooses not to use outpatient healthcare services (structural zeros). For the hurdle model, we assumed that in this dataset, the first visit was on account of the patient, while the subsequent visit was determined by a joint decision of the patient and their healthcare provider [24]. This principal-agent model allows two different processes. While for marginalized-ZINB, allows for differentiation of latent class of zero for 'not at risk' individual and 'at risk' individual for outpatient utilisation [5].

Model selection was based on a few steps. The initial step was to check for data distribution using Stata command 'summarize', 'detail'. The occurrence of overdispersion may suggest using NB, ZIM, or hurdle models [24]. Akaike information criteria (AIC) and Bayesian information criterion (BIC) used to compare between models [25][22], where lower AIC and BIC values is preferred [26]. An additional step was also taken by conducting the Vuong test, where a positive value indicates that zero-inflation is appropriate for the said model rather than using a single-equation count model [27] (i.e., Poisson vs. ZIP, NB vs. ZINB). Corrected Vuong test accounting for AIC and BIC value was conducted using an updated "*zipcv*"and zinbcv, [27] and "mzinb" Stata command [28] Root Mean Square Error (RMSE) calculated together and presented with R2 for goodness-of-fit measures [29]. A comparison of the observed and predicted values was also compared [29].

Table 1 Frequency distribution of outpatient visits (number of observations = 11,674)

Total number of outpatient visit (n = 11,674)	Frequency	Percent (%)
0	10,467	89.66
1	1,002	8.58
2	121	1.04
More than 3	84	0.72
y̲(mean)	0.14	
s^2 y (variance)	0.39	

Results

The frequency distribution of outpatient visits showed that 90% of the population were non-users, while 8.58% utilized healthcare services only once, followed by a smaller percentage of other counts as in Table 1. The maximum distribution of outpatient visits was 25 visits (0.01%) over 14 days. Overdispersion of the mean and variance was observed. Figure A provided as supplementary file showed the skewness of 14.9569 and kurtosis = 363.3711, a large value indicating a positive skew distribution and a high-peak of data distribution [30] .

Table 2 shows the variables in each category of the Andersen model. The initial model had 14 independent variables. For predisposing factors: states (13 States and 3 Federal territory) [31], age, ethnic group (5 major ethnic group) [32], sex and education. The six enabler factors

Table 2 Summary statistics of the variables used in the demand equation

Variable	Frequency(%)	Mean(± SD)
Predisposing factors		
State		
Johor	1052 (9.01)	
Kedah	669 (5.73)	
Kelantan	709 (6.07)	
Melaka	636 (5.45)	
Negeri Sembilan	653 (5.59)	
Pahang	745 (6.38)	
Penang	688 (5.89)	
Perak	578 (4.95)	
Perlis	667 (5.71)	
Selangor	1324 (11.34)	
Terengganu	730 (6.25)	
Sabah	855 (7.32)	
Sarawak	710 (6.08)	
Federal Territory of Kuala Lumpur	563 (4.82)	
Federal Territory of Labuan	643 (5.51)	
Federal Territory of Putrajaya	452 (3.87)	
Age (years)		44.83 (± 16.55)
Ethnic group		
Malay	7,613 (65.21)	
Chinese	1,483 (12.7)	
Indian	753 (6.45)	
Bumiputera Sabah	651 (5.58)	
Bumiputera Sarawak	488 (4.18)	
Others ethnic	686 (5.88)	
Sex		
Male	5, 517 (47.26)	
Female	6,157 (52.74)	
Education level		
No formal education	679 (5.82)	
Primary education	2,540 (21.76)	
Secondary education	5,593 (47.91)	
Tertiary education	2,862 (24.52)	
Marital status		
Not married*	3,744 (32.07)	
Married	7,930 (67.93)	
Enabling factors		
Working Status		
No	4,857 (41.61)	
Yes	6,817 (58.39)	
Percentage (50%) of working adults in Household		
No	8,130 (69.64)	
Yes	3,544 (30.36)	
Government Coverage		
No	8,760 (75.04)	
Yes	2,914 (24.96)	
Employer Coverage		
No	9,506 (81.43)	
Yes	2,168 (18.57)	
Household income quintile		
Poorest quintile	2,500 (21.42)	
Second quintile	2,291 (19.62)	

Table 2 (continued)

Variable	Frequency(%)	Mean(± SD)
Third quintile	2,335 (20. 0)	
Fourth quintile	2,301 (19.71)	
Richest quintile	2,247 (19.25)	
Total household income (ln)		7.52(± 1.72)
Health need factors		
Had any self-reported health problem		
No	8,130 (69.64)	
Yes	3,544 (30.36)	
Perceived health status		
Excellent & good	8, 751 (74.96)	
Fair	2,639 (22.61)	
Poor & Very poor	284 (2.43)	
Number of diagnosed NCD**		
0	8,363 (71.64)	
1	1,517 (12.99)	
2	1,036 (8.87)	
3	758 (6.49)	

*Not married includes single/widower/divorcee

**NCD: Non-communicable disease, any combination of diabetes, hypertension and hypercholesterolemia

Table 3 Comparisons across all models using LL, AIC and BIC.

Test statistic	Model					
	OLS	Poisson	Negative Binomial (NB)	ZIP	MZINB*	Hurdle (Probit & NB)
LLa	-11,743	-5,186	-4,608	-1,698	-1,680	-5,624
AICa	23,541	10,425	9,272	3,455	3,420	11,355
BICa	23,740	10,624	9,478	3,668	3,641	11,753
RMSEa	0.5184	0.5169	0.5179	0.3547	0.3548	0.5178
R2a	0.0368	0.0785	0.0553	0.6979	0.6974	0.0537
Vuong testb						
Uncorrected	-	-	-	1259.9c	1984.8c	-
AIC	-	-	-	1259.9c	1984.8c	-
BIC	-	-	-	1259.9c	1984.8c	-

Notes : Abbreviation: LL=log likelihood; AIC=Akaike's information criterion; BIC=Bayesian information criterion; RMSE=root mean square error, R2=r-square

[a] Lower LL, AIC, and BIC were preferred. Lower RMSE and higher R2 values indicate lesser prediction errors

[b] Positive Vuong statistics value indicates zero-inflated model is more appropriate than conventional

[c] Statistical significance at $p<0.001$

* indicates preferred model

comprise of working status, percentage (50%) of working adults in household, government coverage, employer medical coverage, household income quintile and total household monthly income (ln). Need factors were self-reported health problems, perceived health status, and number of non-communicable diseases (NCDs) diagnosed by health care workers. For final model, manual variable selection was performed and reiteration of regrouping some variables in case if the model can be improved also conducted. The instrumental variables were used consistently across all models.

To select the best model, we used model selection criteria based on the AIC and the BIC (Table 3). This table also indicates ZIM both ZIP and MZINB show the smallest log likelihood (LL), AIC, and BIC, indicating that the zero-inflated models were preferred. The positive Vuong test confirms that the ZIM is superior to its respective single-equation count model. In this case, the Vuong test with the AIC and BIC correction for both ZIM (ZIP and MZINB) yielded positive values ($p<0.001$), which corresponds to a statistically significant selection of the zero-inflated model. Measures of goodness-of-fit using RMSE showed ZIM: ZIP (0.3547) and MZINB (0.3548) had a smallest value and R2 of 0.69. Smaller RMSE, reflect the smaller bias between predicted and observed values for each count on the count models considered as depicted in Fig. 1. It shows the distribution of the observed count and predicted count in the models respectively. The red diagonal line served as the best-fitted line for the counts. Additional Figure B: Scatterplot of predicted values vs.

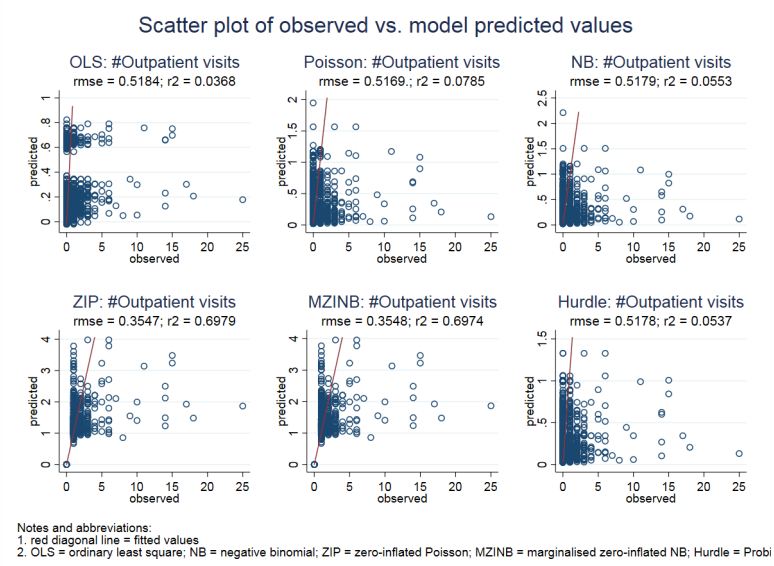

Fig. 1 Scatter plot of observed vs. predicted values

residuals and Figure C: QQ-plot of the residuals provided as Supplementary files.

In Fig. 1,

Thus, we are confident in choosing ZIM of either ZIP or MZINB as the appropriate model based on the lowest AIC and BIC values with a positive Vuong test ($p<0.001$), Table 4 lists the coefficients of one ZIM best-fit model, MZINB. Through regression modeling, the final model included four instrumental variables. States, ethnicity, household income quintile, and perceived health status were significantly associated with the total number of outpatient visits. This study found that the major ethnic groups: Malay, Chinese and India, and those with perceived health status of "poor and very poor," had significantly higher number of total outpatient visits. Adult population in certain states in Peninsular Malaysia; Johor, Melaka, Pahang and Terengganu showed a significant outpatient utilisation parallel with significant outpatient utilisation among Sabah and Sarawak population.

Discussion

Outpatient utilisation data from the NHMS 2019 survey have large zeros, non-negative integers, and continuous data with discrete events. In this dataset, zeros accounted for 89.96% of the total. Approximately 8.58% of the respondents made one outpatient visit, while 1.69% of the population made follow-up outpatient visits. Our data showed a large excess of zeros with a long positive skewed tail, with a maximum of nine outpatient visits over 14 days. This is consistent with other survey data (24, 25) in Asia. Neighboring countries in Indonesia have large zeros amounting to 85% outpatient visits in public facilities and 92% in private facilities [33] in a four-week reference period. Similarly, the Jordan National Health Survey captured 80% of zero [34] with a fourteen-day recall period for outpatient care. In contrast, a study in Norway with a 12 month reference period recorded 78.5 −86.2% of outpatient utilisation [35]. These examples show that a shorter reference period results in a larger zero. A large zero with a short reference period is inevitable because of the survey design. This study included a 14-day reference period. This study makes assumptions about the two types of zeros— frequency zeros due to no outpatient visits throughout the reference period and a true zero that might be due to no illness or presence of illness but not seeking outpatient care.

The predictors of one outpatient visit are usually determined using OLS. However, count variables, especially those involving healthcare data, rarely meet the distributional assumptions of ordinary least square regressions [24] of normality and constant variance. The OLS results depicted the highest LL, AIC, and BIC in this study. This can result in inaccurate estimates of standard errors, p-values, and confidence intervals. However, a recently published local study [21] of oral healthcare has taken measures to limit the data analysis for one visit and make the assumption that zero occurrences are true zero.

In our study, Poisson's stand as a basic count model has a strict condition of equal variance and mean [36] and an outcome variable with a Poisson distribution. Owing to its distinct characteristics, Poisson is unsuitable for this data distribution. Our outcome variable has an excess of zero in front, with a long positive tail. While NB caters for overdispersion of mean and variance [37] [38], comparing the value of LL between NB and Poisson shows a smaller LL value in NB than in Poisson. These

Table 4 Estimated coefficients for the best fit model, MZINB

Variable	Coef.	SE	*p*-value	95% CI	
				Lower	Upper
Predisposing factors					
State					
Johor	0.326	0.135	0.016	0.061	0.591
Kedah	0.083	0.153	0.589	-0.217	0.382
Kelantan	0.193	0.160	0.228	-0.121	0.507
Melaka	0.411	0.172	0.017	0.075	0.747
Negeri Sembilan	0.264	0.141	0.061	-0.012	0.540
Pahang	0.295	0.144	0.041	0.012	0.577
Penang	*Ref*				
Perak	0.307	0.143	0.031	0.028	0.587
Perlis	0.132	0.247	0.593	-0.352	0.617
Selangor	0.171	0.123	0.165	-0.070	0.413
Terengganu	0.709	0.143	< 0.001	0.429	0.988
Sabah	0.669	0.144	< 0.001	0.387	0.951
Sarawak	0.551	0.144	< 0.001	0.269	0.834
Federal Territory of Kuala Lumpur	0.227	0.161	0.157	-0.088	0.542
Federal Territory of Labuan	0.340	0.363	0.349	-0.371	1.050
Federal Territory of Putrajaya	0.326	0.296	0.271	-0.255	0.906
Ethnic group					
Malay	0.410	0.119	0.001	0.176	0.643
Chinese	0.401	0.117	0.001	0.173	0.630
Indian	0.411	0.140	0.003	0.137	0.686
Bumiputera Sabah	-0.073	0.163	0.654	-0.393	0.247
Bumiputera Sarawak	*Ref*				
Others ethnic	0.314	0.147	0.032	0.026	0.601
Enabling factors					
Poorest quintile	*Ref*				
Second quintile	0.101	0.066	0.123	-0.027	0.230
Third quintile	0.068	0.065	0.296	-0.059	0.195
Fourth quintile	0.186	0.069	0.007	0.052	0.320
Richest quintile	0.234	0.067	< 0.001	0.103	0.366
Health need factors					
Had any self-reported health problem					
Perceived health status	*Ref*				
Excellent & good	0.032	0.045	0.487	-0.057	0.120
Fair	0.586	0.069	< 0.001	0.451	0.722
Poor & Very poor	-0.563	0.174	0.001	-0.905	-0.221
Intercept	0.326	0.135	0.016	0.061	0.591

information criteria were used to help determine appropriate models. The lower the AIC and BIC values, the better the model.

Because NB is unable to cater to overdispersion due to excessive zero, ZIM is considered as an alternative modeling strategy. In our data, ZIM showed better LL, AIC, and BIC values than the hurdle model. Across all models, ZIM was deemed suitable as evidence by the smallest values of LL, AIC, and BIC. The results of goodness-of-fit with smaller RMSE value and better R2 value also preferred ZIM. Our findings are also consistent with other findings that used ZIM [39][40]. The different underlying theories and processes of ZIM and hurdle models also serve as a basis for model selection between these two models. In our health system setting, follow-up visits are usually scheduled by healthcare professionals, especially in a public healthcare setting. Thus, it is more appropriate to use the ZIM.

The Vuong test was used to determine whether estimating a zero-inflation component is appropriate or whether a single equation count model should be used [27]. The result of Stata using Vuong is biased toward supporting the zero-inflation model. The results of both corrected and uncorrected Vuong tests show a positive value, indicating the selection of ZIP. However, with the implementation of a new *zipcv* and *zinbcv* Stata command, there are no significant differences. This study reported no large differences compared to previous studies [27]. In this study, we utilized traditional ZIP and MZINB in exploring the best-fit model for the said data. MZINB used as to cater for unobservable latent classes pertaining to the count zero[5]. Lower LL, AIC, BIC values, and positive Vuong Test, significant p-value together with smaller RMSE value and R2 yielded almost a similar value for both ZIM for our data.

In this study, states, ethnicity, household income quintile and perceived health statuswere significant ddeterminants of closely related healthcare utilisation. This trend is visible in numerous other studies that show that wealth are associated with healthcare utilisation [41]. Socioeconomic factors reflected by household income quintile play an important role in outpatient healthcare utilisation [42, 43]. A population with need factors seeks medical outpatient care. In our study, the perceived health status seems to be a significant factor for seeking outpatient care. This concurs with other studies, as perceived health status increases healthcare utilisation especially in an outpatient setting[44].

The strength of this study lies in its utilisation of national healthcare data. It reflects the entire Malaysian population regardless of citizenship. The NHMS is conducted every four years; thus, it is the best available data reflecting the accuracy and timeliness of healthcare utilisation. The model constructed in this study was adapted to meet the characteristics and population data collected by assimilating the Andersen's Behavioral Model of Health Services Use. However, this study did not explore the subsequent frequency of outpatient visits using either public or private facilities. In addition, details of the types of government coverage were not specifically explored. This could be an interesting area to explore, given that the government progressively increases initiatives to increase access to outpatient utilisations either in government or

private facilities to achieve Universal Health Coverage (UHC).

Conclusion

Our study demonstrated the statistical advantages of count data model approaches over traditional OLS. The overdispersion shows violations of the underlying assumptions of normality and constant variance when using OLS. In practice, count data models are relatively easy to interpret using Stata. However, we are aware that these techniques are not widely used. Therefore, this study of count data strategies guides and encourages the appropriate use of models in healthcare utilisation studies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01733-3.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

We thank the Director General of Health Malaysia for permission to publish this article. We are thankful for all research team members for their contributions and to all respondent's kind cooperation. We would also grateful to Dr Zulkarnain and Adilius Manual comments for improvement.

Authors contribution

NSAB contributes in overall conceptualization, methodology, writing and final editing of the paper. JAH contributes in the methodology, formal analysis of the data, writing and final editing of the paper.JS contributes to the writing of the paper.ASJ contributes in the methodology and analysis of the data. All authors agreed to be responsible for all aspect of the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by Ministry of Health Malaysia (NMRR-18-3085-44207).

Data Availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request upon permission from Director General of Health, Malaysia.

Declarations

Ethics approval and consent to participate

Data source form this study obtained from National Health and Morbidity Survey 2019. This survey obtained ethical approval from Medical Research & Ethical Committee (MREC), Ministry of Health Malaysia (NMRR-18-3085-44207) and all methods were in accordance to the Declarations of Helsinki. Informed Consents from respondents were obtained prior to the interviews.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 March 2022 / Accepted: 22 September 2022
Published online: 05 October 2022

References

1. Colin CA, Pravin T. Regression analysis of count data, Second edition. Regres. Anal. Count Data, Second Ed. 2013.
2. Lee J-H, Han G, Fulp J, Giuliano R. Analysis of overdispersed count data:application to the Human. Epidemiol Infect. 2012;140:1–7.
3. Speedie SM, Park YT, Du J, Theera-Ampornpunt N, Bershow BA, Gensinger RA, et al. The impact of electronic health records on people with diabetes in three different emergency departments. J Am Med Informatics Assoc. 2014;21.
4. Hu M-C, Pavlicova M, Nunes EV, Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. Am J Drug Alcohol Abuse [Internet]. Taylor & Francis; 2011;37:367–75. Available from: https://doi.org/10.3109/00952990.2011.597280.
5. Preisser SJ, Das K. L LD, K D. Marginalized Zero-inflated Negative Binomial Regression with Application to Dental Caries. Stat Med. 2016;35:1722–35.
6. Long DL, Preisser JS, Herring HA, Golin EC. A marginalized Zero-inflated Poisson Regression Model with Random Effects. J R Stat Soc Ser C Appl Stat. 2015;64:815–30.
7. Rodríguez-Avi J, Conde-Sánchez A, Sáez-Castillo AJ, Olmo-Jiménez MJ, Martínez-Rodríguez AM. A generalized Waring regression model for count data. Comput Stat Data Anal [Internet]. Elsevier B.V.; 2009;53:3717–25. Available from: https://doi.org/10.1016/j.csda.2009.03.013.
8. Sáez-Castillo AJ, Conde-Sánchez A. A hyper-Poisson regression model for overdispersed and underdispersed count data. Comput Stat Data Anal [Internet]. Elsevier B.V.; 2013;61:148–57. Available from: https://doi.org/10.1016/j.csda.2012.12.009.
9. Islam MM, Alam M, Tariquzaman M, Kabir MA, Pervin R, Begum M, et al. Predictors of the number of under-five malnourished children in Bangladesh: Application of the generalized poisson regression model. BMC Public Health. 2013;13.
10. Ateca-Amestoy V, Prieto-Rodriguez J. Forecasting accuracy of behavioural models for participation in the arts. Eur J Oper Res [Internet]. 2013;229:124–31. Available from: https://www.sciencedirect.com/science/article/pii/S0377221713001239.
11. Sarma S, Simpson W. A microecenometric analysis of Canadian health care utilization. Health Econ. 2006;15:219–39.
12. Deb P, Trivedi PK. The structure of demand for health care: latent class versus two-part models. J Health Econ [Internet]. 2002;21:601–25. Available from: https://www.sciencedirect.com/science/article/pii/S0167629602000085.
13. Gerdtham U-G. Equity in Health Care Utilization: Further Tests Based on Hurdle Models and Swedish Micro Data. Health Econ [Internet]. 1997;6:303–19. Available from: https://econpapers.repec.org/RePEc:wly:hlthec:v:6:y:1997:i:3:p:303-319.
14. Gamme C, Morin J. Health determinants that influence the seeking and utilization of health care A qualitative study among non-natives. 2009.
15. Garcia-Subirats I, Vargas Lorenzo I, Mogollón-Pérez AS, De Paepe P, da Silva MRF, Unger JP, et al. Determinantes del uso de distintos niveles asistenciales en el Sistema General de Seguridad Social en Salud y Sistema Único de Salud en Colombia y Brasil. Gac Sanit [Internet]. 2014;28:480–8. Available from: https://www.sciencedirect.com/science/article/pii/S0213911114001629.
16. Andersen RM, Newman JF. Societal and Individual Determinants of Medical care Utilization in the United States. Milbank Q. 2005;83:1–28.
17. Babitsch B, Gohl D, von Lengerke T. Re-revisiting Andersen's Behavioral Model of Health Services Use: a systematic review of studies from 1998–2011. Psychosoc Med. 2012;9:Doc11.
18. Andersen RM. National health surveys and the behavioral model of health services use. Med Care United States. 2008;46:647–53.
19. Chen AW, Kazanjian A, Wong H. Determinants of mental health consultations among recent Chinese immigrants in British Columbia, Canada: implications for mental health risk and access to services. J Immigr Minor Heal United States. 2008;10:529–40.
20. IPH. National Health and Morbidity Survey 2019 - Technical Report (MOH/S/IKU/144.20(TR)-e) [Internet]. Minist. Heal. Malaysia. 2019. Available from: http://www.iku.gov.my/nhms-2019.
21. Tan YR, Tan EH, Jawahir S, Mohd Hanafiah AN, Mohd Yunos MH. Demographic and socioeconomic inequalities in oral healthcare utilisation in Malaysia: evidence from a national survey. BMC Oral Health [Internet]. 2021;21:34. Available from: https://doi.org/10.1186/s12903-020-01388-w.
22. Hofstetter H, Dusseldorp E, Zeileis A, Schuller AA. Modeling Caries Experience: Advantages of the use of the hurdle model. Caries Res. 2016;50:517–26.
23. Jaafar S, Mohd Noh K, Muttalib KA, Othman NH, Healy J, Maskon K, et al. Malaysia Health System Review. Health Syst Transit [Internet]. 2013;3:1–103.

Available from: http://www.wpro.who.int/asia_pacific_observatory/hits/series/Malaysia_Health_Systems_Review2013.pdf.
24. Du J, Park YT, Theera-Ampornpunt N, McCullough JS, Speedie SM. The use of count data models in biomedical informatics evaluation research. J Am Med Informatics Assoc. 2012;19:39–44.
25. Samsudin S, Jamil N, Zulhaid N. Health care utilisation in Kedah: A micro-econometric analysis. OIDA Int J Sustain Dev. 2012;4:45–52.
26. Mohammed EA, Naugler C, Far BH. Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics [Internet]. Emerg. Trends Comput. Biol. Bioinformatics, Syst. Biol. Algorithms Softw. Tools. Elsevier Inc.; 2015. Available from: https://doi.org/10.1016/B978-0-12-802508-6.00032-6.
27. Desmarais BA, Harden JJ. Testing for zero inflation in count models: Bias correction for the Vuong test. Stata J. 2013;13:810–35.
28. Cummings TH, Hardin JW. Modeling count data with marginalized zero-inflated distributions. Stata J. 2019;19:499–509.
29. Le DD, Gonzalez RL, Matola JU. Modeling count data for health care utilization: an empirical study of outpatient visits among Vietnamese older people. BMC Med Inform Decis Mak [Internet]. BioMed Central; 2021;21:265. Available from: https://doi.org/10.1186/s12911-021-01619-2.
30. Chan YH. Biostatistics 101. Singapore Med J [Internet]. 2003;44:280–5. Available from: https://medicine.nus.edu.sg/rsu/wp-content/uploads/sites/15/2020/02/biostat101_resources3.pdf.
31. Safurah J, Kamaliah MH, Khairiyah AM, Nour HO, Healy J. Ministry of Health, Malaysia Kamaliah Mohd Noh, Ministry of Health, Malaysia Khairiyah Abdul Muttalib, Ministry of Health, Malaysia Nour Hanah Othman, Ministry of Health, Malaysia Kalsom Maskon, Ministry of Health, Malaysia Abdul Rahim Abdullah, Ministry. Malaysia Heal Syst Rev [Internet]. 2013;3:103. Available from: http://apps.who.int/iris/bitstream/handle/10665/206911/9789290615842_eng.pdf;j.
32. Nagaraj S, Nai-Peng T, Chiu-Wan N, Kiong-Hock L, Pala J. Counting Ethnicity in Malaysia: The Complexity of Measuring Diversity BT - Social Statistics and Ethnic Diversity: Cross-National Perspectives in Classifications and Identity Politics. In: Simon P, Piché V, Gagnon AA, editors. Soc Stat Ethn Divers [Internet]. Cham: Springer International Publishing; 2015. p. 143–73. Available from: https://doi.org/10.1007/978-3-319-20095-8_8.
33. Hidayat B, Pokhrel S. The selection of an appropriate count data model for modelling health insurance and health care demand: Case of Indonesia. Int J Environ Res Public Health. 2010;7:9–27.
34. Ekman B. The impact of health insurance on outpatient utilization and expenditure: Evidence from one middle-income country using national household survey data. Heal Res Policy Syst. 2007;5:1–15.
35. Hansen AH, Halvorsen PA, Ringberg U, Førde OH. Socio-economic inequalities in health care utilisation in Norway: a population based cross-sectional survey. BMC Health Serv Res [Internet]. 2012;12:336. Available from: https://doi.org/10.1186/1472-6963-12-336.
36. Models for Count Data. Chapter 4POISSON Model COUNT DATA. 2008. p. 277–86.
37. Hofstetter H, Dusseldorp E, Zeileis A, Schuller AA. Modeling Caries Experience: Advantages of the Use of the Hurdle Model. Caries Res [Internet]. 2016;50:517–26. Available from: https://www.karger.com/DOI/10.1159/000448197.
38. Neelon B, O'Malley AJ, Smith VA. Modeling zero-modified count and semi-continuous data in health services research Part 1: background and overview. Stat Med. 2016;35:5070–93.
39. Mouatassim Y, Ezzahid EH. Poisson regression and Zero-inflated Poisson regression: application to private health insurance data. Eur Actuar J [Internet]. 2012;2:187–204. Available from: https://doi.org/10.1007/s13385-012-0056-2.
40. Yang SA Comparison of Different Methods Of Zero-Inflated Data Analysis and Its Application in Health Surveys [Internet]. University Rhode Island; 2014. Available from: https://digitalcommons.uri.edu/theses/345.
41. Awoke MA, Negin J, Moller J, Farell P, Yawson AE, Biritwum RB, et al. Predictors of public and private healthcare utilization and associated health system responsiveness among older adults in Ghana. Glob Health Action. 2017;10:1301723.
42. Abu Bakar NS, Manual A, Ab Hamid J. Socioeconomic status affecting inequity of healthcare utilisation in Malaysia. Malaysian J Med Sci Penerbit Universiti Sains Malaysia. 2019;26:79–85.
43. Thangiah N, Majid HA, Su TT. Comparing income inequalities in healthcare utilization in the low income community in suburban Kuala Lumpur and Malaysia. BMC Health Serv Res [Internet]. 2014;14:P144. Available from: https://doi.org/10.1186/1472-6963-14-S2-P144.
44. Busato A, Dönges A, Herren S, Widmer M, Marian F. Health status and health care utilisation of patients in complementary and conventional primary care in Switzerland - An observational study. Fam Pract. 2006;23:116–24.

Publisher's Note