Nicotine & Tobacco Research, 2020, 1390–1398
doi:10.1093/ntr/nty072
Original investigation
Received June 27, 2017; Editorial Decision April 5, 2018; Accepted April 16, 2018

Original investigation

Models for Analyzing Zero-Inflated and Overdispersed Count Data: An Application to Cigarette and Marijuana Use

Brian Pittman MS,[1] Eugenia Buta PhD,[2] Suchitra Krishnan-Sarin PhD,[1] Stephanie S. O'Malley PhD,[1] Thomas Liss BS,[1] Ralitza Gueorguieva PhD[1,2]

[1]Department of Psychiatry, Yale School of Medicine, New Haven, CT; [2]Department of Biostatistics, Yale School of Public Health, New Haven, CT

Corresponding Author: Brian Pittman, MS, Department of Psychiatry, Yale University School of Medicine, Connecticut Mental Health Center, 3rd Floor CNRU, 34 Park Street, New Haven, CT 06519, USA. Telephone: 203-974-7789; Fax: 203-974-7662; E-mail: brian.pittman@yale.edu

Abstract

Introduction: This article describes different methods for analyzing counts and illustrates their use on cigarette and marijuana smoking data.

Methods: The Poisson, zero-inflated Poisson (ZIP), hurdle Poisson (HUP), negative binomial (NB), zero-inflated negative binomial (ZINB), and hurdle negative binomial (HUNB) regression models are considered. The different approaches are evaluated in terms of the ability to take into account zero-inflation (extra zeroes) and overdispersion (variance larger than expected) in count outcomes, with emphasis placed on model fit, interpretation, and choosing an appropriate model given the nature of the data. The illustrative data example focuses on cigarette and marijuana smoking reports from a study on smoking habits among youth e-cigarette users with gender, age, and e-cigarette use included as predictors.

Results: Of the 69 subjects available for analysis, 36% and 64% reported smoking no cigarettes and no marijuana, respectively, suggesting both outcomes might be zero-inflated. Both outcomes were also overdispersed with large positive skew. The ZINB and HUNB models fit the cigarette counts best. According to goodness-of-fit statistics, the NB, HUNB, and ZINB models fit the marijuana data well, but the ZINB provided better interpretation.

Conclusion: In the absence of zero-inflation, the NB model fits smoking data well, which is typically overdispersed. In the presence of zero-inflation, the ZINB or HUNB model is recommended to account for additional heterogeneity. In addition to model fit and interpretability, choosing between a zero-inflated or hurdle model should ultimately depend on the assumptions regarding the zeros, study design, and the research question being asked.

Implications: Count outcomes are frequent in tobacco research and often have many zeros and exhibit large variance and skew. Analyzing such data based on methods requiring a normally distributed outcome are inappropriate and will likely produce spurious results. This study compares and contrasts appropriate methods for analyzing count data, specifically those with an over-abundance of zeros, and illustrates their use on cigarette and marijuana smoking data. Recommendations are provided.

Introduction

In the field of addiction research, outcomes are often represented as the count of a particular event. For example, in smoking cessation trials, cigarettes smoked per day or the number of smoke-free days is frequently of interest. The number of side effects owing to a test drug or the number of symptoms related to nicotine withdrawal are other examples.

Distributions of counts commonly exhibit *overdispersion*, where the variance is considerably greater than expected under an assumed distribution (eg, Poisson). Additionally, count outcomes are often *zero-inflated*, where excessive zeros beyond what would be expected under a given probability distribution are observed. Zero counts can be generated from one of two separate processes: (1) *sampling* zeros due to sampling variability (eg, a safe drug producing few adverse events) or (2) *structural* zeros, among subjects not at risk for the event (eg, smoking frequency among nonsmokers). That is, the event *might* occur for one segment of a population but, necessarily, *never* occur for another. In practice, counts are generally overdispersed or zero-inflated and, indeed, commonly both. Because normality assumptions are usually violated, standard regression models are inappropriate for analyzing counts.

Generalized linear models allow for analyzing non-normal count data within a regression framework.[1,2] Poisson regression is a popular choice for modeling counts, but assumes the variance and mean of the distribution are equal, which is atypical in practice. The negative binomial (NB) model includes a dispersion parameter allowing for the variance to exceed the mean and is a popular alternative. Advances in statistical software have allowed for employing zero-inflated[3] and hurdle[4] Poisson and NB regression, which allow for modeling zero-generating processes.

A number of recent studies in the field of tobacco and addiction research have utilized count-specific models including NB,[5] zero-inflated Poisson (ZIP),[6] and NB hurdle[7] regression. Another report[8] compared some of these methods when analyzing smoking cessation over two time points, while others have contrasted some of these methods in other scientific fields.[9–14] The purpose of this article is to describe, compare and contrast, and provide recommendations on available methods for analyzing count outcomes including Poisson and NB regression and their zero-inflated and hurdle counterparts. Each model is illustrated with cigarette and marijuana smoking outcome data in young e-cigarette users recruited for an experimental study by our group. The current report does not test any hypothesis proposed in the parent study. Its goal, rather, is to present an overview of each method and illustrate their use on cross-sectional smoking data. Considering the nature of the data when choosing an appropriate model and interpreting effects is also emphasized.

Methods

Subjects and Study Design

Data were obtained from eligible subjects who participated in an NIH-funded study entitled, "Flavors and E-cigarette Effects in Adolescent Smokers," performed by our group. Briefly, participants were all e-cigarette users, between 16 and 20 years old, and were recruited online via Facebook and Craigslist, as well as at local high schools and colleges, to participate in a three-session experimental study examining the effects of e-cigarettes containing different levels of nicotine and menthol. In the current report, we focused on baseline interview data which included daily recording of e-cigarette, cigarette, and marijuana use over the past 28 days using the Time Line Follow Back.[15]

Outcomes

For the purpose of illustration, the reported number of cigarettes and the number of marijuana joints smoked on the day before study intake (1-day back on the Time Line Follow Back) served as count outcomes. We also show how these models can be fit to rounded average daily counts over the entire 28-day period in Supplement S1.

Predictors

Three predictors were considered: gender (categorical), age (continuous), and e-cigarette use (continuous). Daily e-cigarette use was quantified as the product of the number of times the device was used and the average number of puffs for each use. Daily e-cigarette use was log-transformed to achieve normality. The choice of predictors was guided by both statistical and substantive considerations. From a statistical perspective, the relatively small sample size precluded testing additional predictors. Substantively, there is a broad literature concerning youth smoking with respect to gender, age, and e-cigarette use.[16–19]

Models Considered

Poisson Regression

Poisson regression is a widely considered method for analyzing counts. Log of expected (mean) counts is modeled as a linear function of predictors, constraining predicted responses to be non-negative. Estimated coefficients represent the expected change in the log of the mean for a one unit change in the corresponding predictor. To facilitate interpretation, the inverse of the log link is applied by exponentiating model coefficients to estimate rate ratios (RR), similar to exponentiating coefficients in logistic regression to estimate odds ratios. Poisson regression assumes that the count is Poisson distributed, with its mean, μ, equaling its variance—also known as *equidispersion*. However, this assumption is too restrictive as count data are often overdispersed.[20] An alternative model is the overdispersed Poisson model where a dispersion parameter, k, is introduced to the relationship between the mean and variance, such that $\text{var}(y) = k\mu$. McCullagh and Nelder[2] suggested using the ratio of deviance or the Pearson chi-square to its associated degrees of freedom as an estimate of k. The overdispersed Poisson model does not affect parameters and predicted values compared to those estimated by Poisson, but appropriately increases standard errors by a factor of k when data are overdispersed. Because the dispersion parameter is artificial, requiring a quasi-likelihood approach for inference without full knowledge of the probability distribution, the overdispersed Poisson model is not considered as a competing model in this report.

NB Regression

The NB regression model is a favorable alternative to the Poisson model when data are overdispersed. Specifically, the NB model allows for overdispersion since it has an additional dispersion parameter, k, built in the distribution. Because the variance of the NB distribution, $\mu + k\mu^2$, is also a function of the dispersion parameter, the model is less restrictive than the Poisson model. When the dispersion parameter converges to zero (ie, $k \to 0$), the NB model approaches a Poisson model. Interpretation of regression coefficients is the same as in Poisson regression.

Zero-Inflated Models

As noted earlier, count data are zero-inflated when extra zeros exist above what would be expected under a given probability distribution (eg, Poisson, NB). Zero-inflated regression[3] was developed to analyze such data and assumes the zeros come from two latent sub-classes. Specifically, one population consists of observations that *might* contain zero counts (at risk class) due to sampling while another population consists of observations that *always* contain zeros (not at risk class). Conceptually, the ZIP model makes sense when the population consists of, say, for example, in a smoking study, nonsmokers (no risk) and smokers (at risk), some of whom may produce zeros cigarette counts due to sampling—for example, smokers trying to quit. The model typically uses logistic regression (logit link) to discriminate between structural *always* zeros versus sampling zeros and positive counts; other link functions could also be used (eg, probit, complementary log-log). A count model (eg, Poisson, NB) is simultaneously used to model the counts among observations that *might* contain zeros. Predictors used in each model are not constrained to be the same, adding a level of flexibility. Odds ratios (OR) and rate ratios can be estimated from the logistic and count portions of the model, respectively.

Hurdle Models

While both structural and sampling zeros are modeled simultaneously in the zero-inflated model, the hurdle model[4] posits the entire population is at risk for the event under study and that all zeros are generated from a single structural process. Specifically, logistic regression is first used to discriminate zero counts from nonzero counts. Conditioned on crossing the nonzero "hurdle," the remaining positive counts are then modeled using a truncated (at zero) probability distribution (eg, truncated Poisson or truncated NB). Like zero-inflated models, predictors of choice are used in each model component. Conceptually, the hurdle model makes sense when the entire population is at risk for an event (eg, the number of nonsmoking days among smokers).

Longitudinal Analyses of Daily Cigarette Use

Longitudinal analyses of daily cigarette use were also performed using generalized estimating equation analysis, and generalized linear mixed models with random intercepts only. Details of these models are described in Supplement S2.

Statistical Methods

Cigarette and marijuana counts were modeled as a function of gender, age, and e-cigarette use using Poisson, NB, ZIP, zero-inflated NB (ZINB), hurdle Poisson (HUP), and hurdle NB (HUNB) regression. The same predictors were included in the logit and count components of the zero-inflated and hurdle models. Exponentiated coefficients and 95% confidence intervals were estimated for each model. Model fit was assessed using the Akaike information criterion (AIC), the Schwartz-Bayesian information criterion (BIC), and −2 log-likelihood statistics where for each, smaller values indicate better fit. Models were tested pair-wise for equivalence using the likelihood ratio-based Vuong[21] test, which produces a z-statistic where a value >1.96 supports the alternative that the first model fits the data better and a value <−1.96 indicates the second model fits better. Data were analyzed using PROC FMM in SAS, version 9.4 (SAS Institute, Inc., Cary, NC). The SAS code for the different models is included in the Supplementary Appendix.

Results

A total of 69 subjects were available for analysis. Subjects were 40% female and between 16 and 20 years old (average ± standard deviation = 18.5 ± 0.95). Among respondents, 36% reported zero cigarettes smoked while 64% reported zero marijuana use on the day before intake suggesting each outcome may contain excess zeros. Distributions for both cigarettes (average ± standard deviation = 4.8 ± 6.5) and marijuana joints (average ± standard deviation = 1.4 ± 3.4) exhibited high positive skew with each variance exceeding its mean by a factor of ~8.5, suggesting each outcome is overdispersed relative to the Poisson distribution.

Cigarette Smoking Results

For each model, predicted probabilities superimposed on the observed distribution of cigarette frequencies are shown in Figure 1. The Poisson model grossly underestimated the observed zeros (1% vs. observed 36%) and overestimated observed frequencies between two and seven cigarettes. The NB model predicted the zeros better (31% vs. 36%), but also overestimated relatively few cigarettes. By contrast, the ZIP, HUP, ZINB, and HUNB models all perfectly predicted no smoking. As shown in Figure 1, the ZIP mirrored the performance of the HUP, and the ZINB was nearly indistinguishable from HUNB. The ZINB and HUNB models matched the observed data more closely than their Poisson counterparts.

Rate ratios (count component), odds ratios (logistic component), and corresponding 95% confidence limits for each model predicting smoking are shown in Table 1. Estimated dispersion (k) for each NB model along with fit statistics and Vuong's tests are also shown. Each effect was significant in the Poisson model, but the model had the poorest fit according to the AIC and BIC and was statistically inferior to all models (all Vuong, $p < .01$). While the NB model estimated significant dispersion ($k = 1.9$, $p < .0001$), null effects were observed for each predictor. After accounting for excess zeros, null effects were observed for each predictor in the count components of the ZIP and HUP and, according to fit statistics, had inferior fits compared to the NB. Significant dispersion, although reduced compared to NB (0.54 vs. 1.9), persisted in the count components of the ZINB and HUNB. After accounting for both zero-inflation and significant dispersion, the ZINB and HUNB models fit the data best

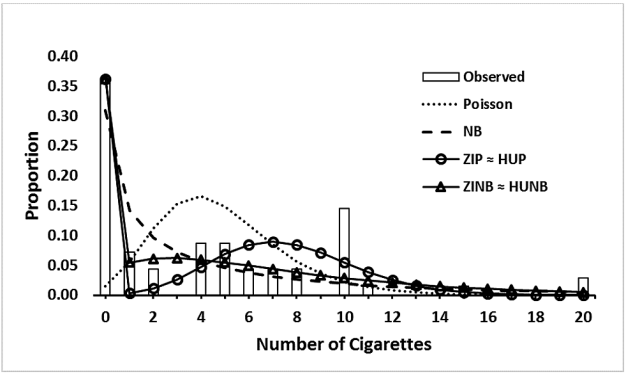

Figure 1. Observed versus predicted cigarette use reported on the day just before study intake. NB = negative binomial; ZIP = zero-inflated Poisson; HUP = hurdle Poisson; ZINB = zero-inflated negative binomial; HUNB = hurdle negative binomial; cigarettes use was truncated at 20 for clarity. Not shown: a single endorsement of 40 cigarettes. Note: ZIP and ZINB are modeling structural and sampling zeros.

Table 1. Comparison of Parameter Estimates Predicting Cigarette Smoking Among the Six Models

Count component	Poisson	NB	ZIP	ZINB	HUP	HUNB
	RR	RR	RR	RR	RR	RR
	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
	p-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value
Female	1.42	1.55	1.10	1.07	1.10	1.09
	(1.15% to 1.76%)	(0.73% to 3.31%)	(0.88% to 1.38%)	(0.63% to 1.83%)	(0.88% to 1.38%)	(0.64% to 1.85%)
	.001	.26	.42	.80	.42	.75
Age	0.89	0.77	1.10	1.16	1.10	1.14
	(0.79% to 1.00%)	(0.47% to 1.28%)	(0.96% to 1.26%)	(0.81% to 1.66%)	(0.96% to 1.26%)	(0.80% to 1.61%)
	.04	.32	.15	.41	.15	.47
eCig	0.90	0.91	0.95	0.91	0.95	0.93
	(0.84% to 0.97%)	(0.68% to 1.21%)	(0.87% to 1.04%)	(0.73% to 1.13%)	(0.87% to 1.04%)	(0.76% to 1.15%)
	.007	.51	.24	.38	.25	.51
Dispersion (*k*)	—	1.93	—	0.54	—	0.55
	—	(1.09% to 2.78%)	—	(0.17% to 0.91%)	—	(0.17% to 0.93%)
	—	.0001	—	.005	—	.005
Logit component (modeling cig = 0)	OR	OR	OR	OR	OR	OR
	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
	p-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value
Female	—	—	0.43	0.38	0.43	0.43
	—	—	(0.14% to 1.35%)	(0.10% to 1.52%)	(0.14% to 1.35%)	(0.14% to 1.35%)
	—	—	.15	.17	.15	.15
Age	—	—	2.34	2.78	2.34	2.34
	—	—	(1.17% to 4.70%)	(1.14% to 6.79%)	(1.17% to 4.68%)	(1.17% to 4.68%)
	—	—	.02	.03	.02	.02
eCig	—	—	1.31	1.26	1.31	1.31
	—	—	(0.89% to 1.93%)	(0.82% to 1.93%)	(0.89% to 1.93%)	(0.89% to 1.93%)
	—	—	.17	.30	.17	.17
Goodness-of-fit						
−2 Log-likelihood	624.1	354.4	426.0	336.6	426.0	336.5
AIC	632.1	364.4	442.0	354.6	442.0	354.5
BIC	641.0	375.6	459.9	374.7	459.9	374.6
Vuong test[a]	Poisson	NB	ZIP	ZINB	HUP	HUNB
Poisson	—	−2.90	−4.04	−3.17	−4.04	−3.17
NB		—	1.17	−2.03	1.17	−2.10
ZIP			—	−1.60	0.25	−1.60
ZINB				—	1.60	−0.19
HUP					—	−1.60

RR = rate ratio; OR = odds ratio; CI = confidence interval; NB = negative binomial; ZIP = zero-inflated Poisson; HUP = hurdle Poisson; ZINB = zero-inflated negative binomial; HUNB = hurdle negative binomial.
[a]Vuong statistics are *z*-scores where values <−1.96 indicate better fit among the column model compared to row model and vice-a-versa for scores >1.96.

per fit statistics and were marginally superior to their ZIP and HUP (both Vuong tests, $p < .11$) counterparts. Therefore, both ZINB and HUNB were chosen as candidate models for interpretation.

Interpretation between the ZINB and HUNB are similar but with an important distinction, particularly with respect to the logit component. Estimates from the HUNB logit suggests increased age is significantly associated with smoking zero cigarettes (OR = 2.34, 95% confidence interval [CI]: 1.17% to 4.68%), whereas the ZINB logit model suggests age is positively associated with being a structural zero—that is, belonging to the risk-free (nonsmokers) latent class (OR = 2.78, 95% CI: 1.14% to 6.79%). Predictors in the count components from both models were not statistically significant.

Marijuana Smoking Results

Predicted probabilities and observed frequencies are depicted in Figure 2. The Poisson underestimated zeros in the data and overestimated use between 1 and 4 joints. Performance of ZIP, ZINB, and HUP coincided, each perfectly predicting zeros while underestimating a single joint and overestimating 2–4 joints. Predictions based on the NB and HUNB models were indistinguishable and appeared to reflect the observed data best.

Results of the models predicting marijuana use are shown in Table 2. The Poisson had the poorest fit as indicated by the largest AIC and BIC values and was inferior to all models (Vuong, all $p < .001$). Based on fit statistics and Vuong tests, the NB, ZINB, and HUNB were candidates for interpretation. The significant dispersion was estimated in the NB model. Estimates from both the count and logit components of the ZINB and HUNB suggest e-cigarette use was associated with marijuana use. From the ZINB count component, for example, increased e-cigarette use was positively associated with intensity of use (RR = 1.41, 95% CI: 1.13% to 1.76%) in the at-risk latent class and, conversely, associated with increased odds of belonging to the risk-free (nonmarijuana user) latent class (OR = 2.14, 95% CI: 1.11% to 4.14%). Similar effects of e-cigarettes were observed in each component of the HUNB, while the effect of e-cigarettes was null in the NB (RR = 1.13, 95% CI: 0.84% to 1.51%). The female gender effect in the NB (RR = 0.25, 95% CI: 0.08% to 0.78%) appeared spread across both model components of the HUNB: a nonsignificant gender effect was observed in the HUNB count component (RR = 0.46, 95% CI: 0.13% to 1.66%), but a borderline trend ($p = .10$) effect was observed in the logit (OR = 2.68, 95% CI: 0.83% to 8.67%), suggesting females are more likely to avoid marijuana altogether. Likewise, increasing age was associated with diminished use (RR = 0.52, 95% CI: 0.26% to 1.02%) in the NB model, while the HUNB model suggested distinguished age users from nonusers (OR = 1.84, 95% CI: 0.98% to 3.47%) but was not associated with intensity of use among those that smoke the drug (RR = 0.96, 95% CI: 0.45% to 2.05%).

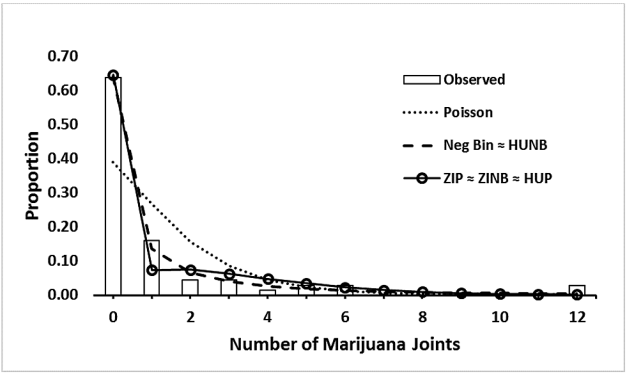

Figure 2. Observed versus predicted marijuana use reported on the day just before study intake. NB = negative binomial; ZIP = zero-inflated Poisson; HUP = hurdle Poisson; HUNB = hurdle negative binomial; marijuana use was truncated at 12 for clarity. Not shown: a single endorsement for 21 marijuana joints. *Note:* ZIP and ZINB are modeling structural and sampling zeros.

Discussion

We compared and contrasted six different count models for analyzing cigarette and marijuana count data: Poisson, NB, ZIP, ZUP, HUP, and HUNB. The observed distributions shown in Figures 1 and 2 clearly indicate the outcomes were not normally distributed. With each variance exceeding its mean by a factor of 8.5 and notable clumping of zero counts, both outcomes were overdispersed and likely zero-inflated. While Poisson regression is often used as a baseline model for counts, in practice most count outcomes are overdispersed relative to the Poisson distribution, even when considering only positive counts.[22] Ignoring overdispersion could lead to dramatically smaller standard errors and false positive results.[20] For these reasons, and because as the NB dispersion parameter (k) converges to zero the model approaches the Poisson, we recommend avoiding Poisson regression in most practical cases and advocate for using NB regression as a baseline model for count data, even when the data contain many zeros.

With the inclusion of a dispersion parameter, the NB model has more flexibility to capture additional variability and therefore fit the highly variable cigarette data better than the Poisson. However, the model underestimated zeros in the data, suggesting a model appropriate for zero-inflated data may be indicated. As shown in Figure 1, both the ZIP and HUP perfectly predicted the zero counts. However, dispersion parameters from the ZINB and HUNB were significant, suggesting the equidispersion assumption in the count portions of ZIP and HUP was violated. Therefore, after considering both the zero counts and the enduring overdispersion, the ZINB and HUNB fit the data best according to the AIC and BIC and trend-level Vuong tests.

For cigarette smoking, choosing between the ZINB and HUNB should ultimately depend on the assumptions regarding the zeros. In the current study, smoking status was ascertained from the Time Line Follow Back (ie, no smoking status variable per se) so the nature of the zeros was unknown. From one perspective, researchers might consider subjects not smoking on a daily basis as, ostensibly, nonsmokers. In this case, it's reasonable to report covariate effects on the odds of being a nonsmoker and their effects on smoking intensity among subjects that do smoke from the HUNB model. However, a researcher might consider consistent daily cigarette use as too stringent in determining smoking status. From this perspective, the observed zeros might emanate from both nonsmokers (structural zeros) and smokers who chose not to smoke that particular day (sampling zeros). In this case, reporting results from the ZINB model makes better conceptual sense.

The NB model accurately estimated zero marijuana use with fit statistics commensurate to those from the ZINB and ZUNB models. This finding reinforces the notion that count data with many zeros don't necessarily require a zero-inflated or hurdle-type model, as Xie et al.[8] point out. However, the ZINB and HUNB results offer an alternative interpretation. For example, a nonsignificant effect for

Table 2. Comparison of Parameter Estimates Predicting Marijuana Smoking Among the Six Models

Count component	Poisson	NB	ZIP	ZINB	HUP	HUNB
	RR	RR	RR	RR	RR	RR
	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
	p-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value
Female	0.26	0.25	0.54	0.51	0.51	0.46
	(0.14% to 0.47%)	(0.08% to 0.78%)	(0.26% to 1.12%)	(0.19% to 1.41%)	(0.23% to 1.13%)	(0.13% to 1.66%)
	.0001	.02	.10	.19	.10	.23
Age	0.60	0.52	0.87	0.83	0.93	0.96
	(0.46% to 0.78%)	(0.26% to 1.02%)	(0.63% to 1.21%)	(0.48% to 1.43%)	(0.65% to 1.34%)	(0.45% to 2.05%)
	.0001	.06	.41	.50	.70	.91
eCig	1.39	1.13	1.43	1.41	1.45	1.52
	(1.23% to 1.56%)	(0.84% to 1.51%)	(1.30% to 1.57%)	(1.13% to 1.76%)	(1.31% to 1.60%)	(1.11% to 2.08%)
	.0001	.42	.0001	.002	.0001	.01
Dispersion (*k*)	—	3.23	—	0.55	—	0.97
	—	(1.28% to 5.19)	—	(−0.21% to 1.31)	—	(−0.92% to 2.86)
	—	.001	—	.16	—	.32
Logit component (modeling Marij = 0)	OR	OR	OR	OR	OR	OR
	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
	p-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value	*p*-Value
Female	—	—	2.51	2.47	2.68	2.68
	—	—	(0.65% to 9.645%)	(0.51% to 11.95%)	(0.83% to 8.675%)	(0.83% to 8.675%)
	—	—	.18	.26	.10	.10
Age	—	—	1.71	1.80	1.84	1.84
	—	—	(0.79% to 3.70%)	(0.69% to 4.66%)	(0.98% to 3.47%)	(0.98% to 3.47%)
	—	—	.18	.23	.06	.06
eCig	—	—	1.98	2.14	1.51	1.51
	—	—	(1.09% to 3.62%)	(1.11% to 4.14%)	(1.00% to 2.28%)	(1.00% to 2.28%)
	—	—	.03	.02	.051	.051
Goodness-of-fit						
−2 Log-likelihood	293.8	186.2	185.2	175.6	184.3	173.0
AIC	301.8	196.2	201.2	193.6	200.3	191.0
BIC	310.7	207.4	219.0	213.7	218.2	211.1
Vuong test[a]	Poisson	NB	ZIP	ZINB	HUP	HUNB
Poisson	—	−3.34	−3.54	−3.52	−3.88	−3.75
NB		—	0.32	−1.49	−0.79	−2.46
ZIP			—	−1.18	−0.58	−1.07
ZINB				—	0.28	−0.95
HUP					—	−0.87

RR = rate ratio; OR = odds ratio; CI = confidence interval; NB = negative binomial; ZIP = zero-inflated Poisson; HUP = hurdle Poisson; ZINB = zero-inflated negative binomial; HUNB = hurdle negative binomial.
[a]Vuong statistics are *z*-scores where values <−1.96 indicate better fit among the column model compared to row model and vice-a-versa for scores >1.96.

e-cigarette use was observed in the NB model. In contrast, ZINB and HUNB count components show a positive association between e-cigarettes and marijuana. At the same time, odds ratios from the logits indicate increased e-cigarette use is associated with avoiding marijuana (HUNB) or being classified as a nonuser (ZINB). That is, e-cigarette use might be protective against marijuana initiation but also may reinforce use among those that smoke the drug. It is possible this is because youth may be using e-cigarettes to vape marijuana as has been shown by our earlier work.[23] In their analysis of the number hospital stays from the 1987 National Medical Expenditure Survey, Liu and Cela[10] encountered a similar situation where according to AIC/BIC criteria an NB model fit their data best, yet a zero-inflated model provided additional interpretation.

While the NB and HUNB models had a similar fit to the marijuana data as shown in Figure 2, the Vuong test indicated the HUNB was superior ($p = .01$). However, the HUNB was statistically equivalent to the ZINB ($p = .34$). More importantly, substantive considerations suggest the ZINB model might naturally model the zeros better. For example, like the cigarette data, marijuana use was determined based on counts from a single day. Unlike the cigarette data, however, it seems less justified to classify subjects reporting zero use on a particular day as nonusers, especially since marijuana users don't necessarily smoke the drug every day and because access to marijuana is more restricted compared to cigarettes. Therefore, it is reasonable to assume the zeros were generated from both nonusers (structural) and from users (sampling) who chose not to smoke (or didn't have access to) marijuana on the day prior to study intake. Conceptually, reporting result from the ZINB makes the most sense given the assumed latent nature of the zeros.

The considerations discussed above in choosing each final model reflect the notion that model selection is, as Rose et al.[9] elegantly frames, "often as much art as science." The "art" relies on qualitative assessment of the model in terms of appropriateness and interpretation. If the researcher chooses to treat all zeros the same, then the hurdle model is appropriate, whereas if the researcher believes the zeros can result from a mixture of populations, then the zero-inflated model is appropriate. The "science" of the selection relies on quantitative attributes such as goodness-of-fit criteria and prediction. In discussing the importance of distribution choice when analyzing counts, Wagner et al.[24] and others[25] recommend choosing theoretically appropriate distributions based on characteristics of the outcome, followed by a comparison of fit statistics and visual confirmation. As often the case,[9] no quantitative advantage between ZINB and HUNB were observed when modeling cigarettes, but based on substantive considerations, the HUNB was chosen because it was assumed all zeros represented nonsmokers. For the marijuana data, the NB and HUNB offered quantitative advantages (ie, better fit), but conceptually it seemed natural to assume the zeros were derived from two subclass populations: those at risk (users) and nonusers.

It is worth reminding that, for the purpose of illustrating the application of count models to cross-sectional data, the current study considered cigarette and marijuana use drawn from a single day. In an effort to aggregate information, total counts and the total number of days used over the 28 days were also considered as outcomes. Unfortunately, the distributions of these outcomes did not reflect the count distributions considered in this report. For example, a total number of days of use had both floor (at zero) and ceiling (at 28) effects. The distribution of total counts did not provide a sufficient number of zeros, particularly for cigarette use. Although not a count per se, the rounded average daily use was also considered for both cigarette and marijuana. The best-fitting models for rounded average cigarettes per day yielded results similar to those shown above for cigarette use reported on a single day prior and provided smooth reflections on the data (Supplement S1). Note that in our data example rounded averages had less of a heaping[26] problem than smoking on a particular day (eg, individuals reported smoking, eg, 10 cigarettes on a particular day much more frequently than 9 and 11 cigarettes). Although rounded averages have been modeled as counts previously, both in the subject-matter[27] and in the statistical literature,[28] this approach may not be ideal in some situations, especially when changes over time are expected, and there is substantial intra-individual variability. In the current study, however, cigarette and marijuana use were reported during the baseline period when no systematic change was observed.

While zero-inflated models are attractive for modeling zero-inflated data, Preisser et al.[29] found that interpretations of these models provided in the literature are often imprecise or misleading, particularly with respect to the excess zero latent class. An important advantage of hurdle models is the straightforward interpretation of parameter estimates they provide. Hurdle models represent a two-part decision-making process common in human behavior.[10] In the context of the present study, one must first make the decision to smoke or not smoke (logit portion). Once the decision has been made to smoke—that is, the hurdle has been crossed—the quantity of use is then modeled in the zero-truncated count component. The predictors in each model component, therefore, have a straightforward interpretation. Zero-inflated model are more complex to fit than hurdle models as the logit and count components are fit simultaneously.[22] In fact, zero-inflated models have been shown through simulation studies to be unstable,[30] if not unreliable especially in the presence of *zero-deflation* (ie, fewer zeros than expected) at any level of a covariate, even for cross-sectional data.[31] Unlike zero-inflated models, hurdle models are well-suited to handle both zero-inflated and zero-deflated outcomes, although this advantage is less relevant in substance use research where data are usually zero-inflated. Despite the characteristic advantages of both types of models, whether to use a hurdle or zero-inflated model should be ultimately guided by the study design, whether the nature of the zeros is known, and the question being asked.[9,12] Conceptually, a zero-inflated model is more appropriate when the zeros can be thought to have been generated from a mixture of populations (at risk, not at risk), while a hurdle model is more appropriate when the entire population under study is at risk for the event, with realization of the event representing the hurdle having been crossed.[32]

The models illustrated in this article represent some of the commonly used methods for analyzing count data. However, there are numerous other models and approaches that can be considered. First, one might try normalizing the outcome through transformation (eg, log) and then model the outcome using multiple linear regression. However, such an approach is futile when data are zero-inflated because zero-inflation can't be corrected by a transformation. Through simulations studies, O'Hara and Kotze[33] present pitfalls associated with this approach and argue strongly against transforming count data, especially given the wide variety of count-specific generalized linear models available today. Second, as described previously, the overdispersed Poisson model can be used to adjust standard errors by a factor of an estimated dispersion parameter. Third, employing a nonparametric approach might provide inferential utility, but such an approach precludes effect size estimation. Min and Agresti[22] summarize other approaches for analyzing zero-inflated count data including methods based on finite mixture models[34] and those based on the Neyman type A distribution.[35]

Each of the models presented in this report were based on cross-sectional data. However, each method presented can be extended to handle correlated or clustered count data. Although beyond the scope of this article, results from longitudinal analyses of daily cigarette use applied to the current data are described in detail in Supplement S2. The results are substantively similar to those obtained from the cross-sectional analyses, with higher age being associated with higher probability of abstaining from smoking. Methods for analyzing correlated count data are discussed extensively in the statistical literature,[31,32,36,37] including a review and application using SAS software with available code.[8]

This study has limitations, with the relatively small sample size the most glaring. A larger sample size would have allowed for including more predictors, perhaps important ones which may have explained additional variability. Future research could consider reanalysis using a large sample and simulation studies to assess how each model performs under different scenarios, such as varying proportions of zeros in the data, different levels of skew, and varying levels of overdispersion caused by either zero-inflation, extra variability, or both. Absent rigorous simulation studies, caution should be applied not to generalize our model performance results to other data sets. When analyzing alternative data sets, all models need to be compared and the best model selected based on prespecified statistical and substantive criteria.

The results from the analyses presented here make clear the importance of considering both overdispersion and zero-inflation of count outcomes. Failure to do so can produce biased effect estimates and false positive results. In the absence of zero-inflation, we recommended that the NB model will be of primary consideration when analyzing count outcomes, as count outcomes are usually overdispersed. In the presence of zero-inflation, a zero-inflated or hurdle mode may be more suitable depending on the status of the zeros, and the NB version of each is preferred to capture additional overdispersion due to heterogeneity. When analyzing count outcomes, researchers should carefully examine the distribution of the outcome under study, consider all relevant predictors, and ultimately select a model based on an appropriate balance between model fit, interpretability, and nature of the zero counts.

Supplementary Material

Supplementary data are available at *Nicotine and Tobacco Research* online.

Funding

Research reported in this publication was supported by grant number P50DA036151 from the National Institute on Drug Abuse and FDA Center for Tobacco Products (CTP). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Food and Drug Administration.

Declaration of Interest

None declared.

References

1. Agresti A. *An Introduction to Categorical Data Analysis*. Hoboken, NJ: Wiley; 2007.
2. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall; 1989.
3. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992;34(1):1–14.
4. Mullahy J. Specification and testing of some modified count data models. *J Econom*. 1986;33(3):341–365.
5. van der Sluijs W, Haseen F, Miller M, et al. "It looks like an adult sweetie shop": point-of-sale tobacco display exposure and brand awareness in Scottish secondary school students. *Nicotine Tob Res*. 2016;18(10):1981–1988.
6. Barondess DA, Meyer EM, Boinapally PM, Fairman B, Anthony JC. Epidemiological evidence on count processes in the formation of tobacco dependence. *Nicotine Tob Res*. 2010;12(7):734–741.
7. Sanjuan PM, Rice SL, Witkiewitz K, Mandler RN, Crandall C, Bogenschutz MP. Alcohol, tobacco, and drug use among emergency department patients. *Drug Alcohol Depend*. 2014;138:32–38.
8. Xie H, Tao J, McHugo GJ, Drake RE. Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: an example of smoking cessation. *J Subst Abuse Treat*. 2013;45(1):99–108.
9. 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(4):463–481.
10. Liu W, Cela J. *Count Data Models in SAS*. In: Proceedings SAS Global Forum; March 16–19, 2008; San Antonio, Texas; paper 371–2008.
11. Goulet JL, Buta E, Carroll C, Gueorguieva R, Bathulapalli H, Brandt CA. Statistical modelling approaches for the analysis of pain intensity numeric rating scale data. *J Pain*. 2016;18(3):340–348.
12. Hu MC, 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*. 2011;37(5):367–375.
13. Khan A, Ullah S, Nitz J. Statistical modelling of falls count data with excess zeros. *Inj Prev*. 2011;17(4):266–270.
14. Swartout KM, Thompson MP, Koss MP, Su N. What is the best way to analyze less frequent forms of violence? The case of sexual aggression. *Psychol Violence*. 2015;5(3):305–313.
15. Sobell LC, Sobell MB. Timeline followback: a technique for assessing self-reported ethanol consumption. In: Allen J, Litten R, eds. *Techniques to Assess Alcohol Consumption*. Totowa, NJ: Humana Press; 1993:41–72.
16. Bunnell RE, Agaku IT, Arrazola RA, et al. Intentions to smoke cigarettes among never-smoking US middle and high school electronic cigarette users: National Youth Tobacco Survey, 2011–2013. *Nicotine Tob Res*. 2015;17(2):228–235.
17. Burt RD, Peterson AV Jr. Smoking cessation among high school seniors. *Prev Med*. 1998;27(3):319–327.
18. Demissie Z, Everett Jones S, Clayton HB, King BA. Adolescent risk behaviors and use of electronic vapor products and cigarettes. *Pediatrics*. 2017;139(2):e20162921.
19. Spindle TR, Hiler MM, Cooke ME, Eissenberg T, Kendler KS, Dick DM. Electronic cigarette use and uptake of cigarette smoking: a longitudinal examination of U.S. college students. *Addict Behav*. 2017;67:66–72.
20. Cameron AC, Trivedi PK. *Regression Analysis of Count Data*. Cambridge: University Press; 1998.
21. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 1989;57(2):307–333.
22. Min Y, Agresti A. Modeling nonnegative data with clumping at zero: a survey. *JIRSS*. 2002;1(1):7–33.
23. Morean ME, Kong G, Camenga DR, Cavallo DA, Krishnan-Sarin S. High school students' use of electronic cigarettes to vaporize cannabis. *Pediatrics*. 2015;136(4):611–616.
24. Wagner B, Riggs P, Mikulich-Gilbertson S. The importance of distribution-choice in modeling substance use data: a comparison of negative binomial, beta binomial, and zero-inflated distributions. *Am J Drug Alcohol Abuse*. 2015;41(6):489–497.
25. Gorelick DA, McPherson S. Improving the analysis and modeling of substance use. *Am J Drug Alcohol Abuse*. 2015;41(6):475–478.
26. Wang H, Heitjan DF. Modeling heaping in self-reported cigarette counts. *Stat Med*. 2008;27(19):3789–3804.

27. DeSantis SM, Bandyopadhyay D, Baker NL, Randall PK, Anton RF, Prisciandaro JJ. Modeling longitudinal drinking data in clinical trials: an application to the COMBINE study. *Drug Alcohol Depend*. 2013;132(1–2):244–250.
28. Zhu H, Luo S, DeSantis SM. Zero-inflated count models for longitudinal measurements with heterogeneous random effects. *Stat Methods Med Res*. 2017;26(4):1774–1786.
29. Preisser JS, Stamm JW, Long DL, Kincade ME. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Res*. 2012;46(4):413–423.
30. Baggio S, Iglesias K, Rousson V. Modeling count data in the addiction field: some simple recommendations. *Int J Methods Psychiatr Res*. 2017; 27(1):e1585.
31. Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. *Stat Model*. 2005;5:1–19.
32. Buu A, Li R, Tan X, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Stat Med*. 2012;31(29):4074–4086.
33. O'Hara RB, Kotze DJ. Do not log-transform count data. *Methods Ecol Evol*. 2010;1:118–122.
34. Aitkin M, Rubin DB. Estimation and hypothesis testing in finite mixture models. *J R Stat Soc Series B Stat Methodol*. 1985;47:67–75.
35. Dobbie M, Welsh AH. Models for zero-inflated count data using the Neyman type a distribution. *Stat Model*. 2001;1:65–80.
36. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991;59(1):25–35.
37. Kong M, Xu S, Levy SM, Datta S. GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Comput Stat Data Anal*. 2015;85:54–66.