Published in final edited form as:

Am J Drug Alcohol Abuse. 2011 September ; 37(5): 376–382. doi:10.3109/00952990.2011.568080.

Some considerations for excess zeroes in substance abuse research

Dipankar Bandyopadhyay, PhD¹, Stacia M. DeSantis, PhD¹, Jeffrey E. Korte, MSPH, PhD¹, and Kathleen T. Brady, MD, PhD^{2,3}

¹Department of Medicine, Division of Biostatistics and Epidemiology, Medical University of South Carolina, Charleston, SC

²Ralph H. Johnson VA Medical Center, Charleston SC

³Department of Psychiatry and Behavioral Sciences, Division of Clinical Neuroscience, Medical University of South Carolina, Charleston, SC

Abstract

Background—Count data collected in substance abuse research often comes with an excess of ‘zeroes’, which are typically handled using zero-inflated regression models. However, there is a need to consider the design aspects of those studies before using such a statistical model in order to ascertain the sources of zeroes.

Objectives—We sought to illustrate hurdle models as alternatives to zero-inflated models to validate a two-stage decision making process in situations of ‘excess zeroes’.

Methods—We use data from a study of 45 cocaine dependent subjects where the primary scientific question was to evaluate whether study participation influences drug-seeking behavior. The outcome, ‘the frequency (count) of cocaine use days per week’, is bounded (ranging from 0 to 7). We fit and compare binomial, Poisson, negative binomial, and the hurdle version of these models to study the effect of gender, age, time and study participation on cocaine use.

Results—The hurdle binomial model provides the best fit. Gender and time are not predictive of use. Higher odds of use vs. no use are associated with age; however once use is experienced, odds of further use decreases with increase in age. Participation was associated with higher odds of no-cocaine use; once there is use participation reduced the odds of further use.

Conclusion—Age and study participation are significantly predictive of cocaine use behavior.

Scientific Significance—The two-stage decision process as modeled by a hurdle binomial model (appropriate for bounded count data with excess zeroes) provides interesting insights into the study of covariate effects on count responses of substance use, when all enrolled subjects are believed to be ‘at-risk’ of use.

Keywords

Cocaine addiction; excess zero; hurdle model; zero-inflated model

1. Introduction

The last decade has seen significant productivity and progress in drug abuse research, in part due to the initiatives of the National Drug Abuse Clinical Trials Network (CTN) of the National Institute on Drug Abuse (NIDA). A wide variety of responses (i.e. continuous, binary, count, ordinal, nominal, etc) are collected not only in substance abuse clinical trials but also in observational/survey studies like the National Survey on Drug Use and Health (NSDUH), which provides national and state-level data on the use of tobacco, alcohol, illicit drugs, etc in the United States. In this context, observed count responses might exhibit 'over-dispersion' when fitting the data to a Poisson distribution, in the sense that the sample variance is larger than the sample mean, and hence the well known Poisson assumption of 'unit variance to mean ratio' is violated (1). Over-dispersion (2) can be attributed to several factors, like unobserved heterogeneity, missing covariates, or correlation among repeated/longitudinal measures, which are common features of substance use trials. These count responses are also sometimes characterized by excessive observations at one end of the ordering, typically, zeroes. For example, one might be interested in the question: '*How often do you use a particular drug (or the frequency of intake) within the last 30 days?*' If Y is the response variable, then Y is expected to have a preponderance of zero observations, highlighting the consumption levels of 'never/not recent' use, i.e. $Y = 0$. This phenomenon is often termed as 'zero-inflation' (2) and can also contribute to over-dispersion.

The addiction literature abounds with studies in which the data contains excess zeroes (3), often derived from two distinct sources. Some may be from subjects who choose not to use the drug within the last 30 days and thereby contribute to 'sampling zeroes' while some might never exhibit drug using behavior and are hence considered as 'structural zeroes'. It is likely that the two types of zeros are driven by measured/unmeasured confounders. Additionally the same (or different) set of explanatory variables might have varying effects on the two types of zeroes. For example, a higher income bracket might represent health awareness leading to genuine non-users (structural zeros); however subjects in lower/other income categories might currently behave as non-users (sampling zeros) and would have the potential to be abusers depending on price changes as well as changes in income. Traditional count regression models like binomial (where counts have an upper bound), Poisson or negative binomial (to control over-dispersion) may not be appropriate here, because they cannot account for these two sources of zeroes arising from two separate data generation processes simultaneously, or separately. Fitting these models might result in inflated variance estimates of model parameters or loss of power to determine a possible intervention effect. When the data has an excess of both structural and sampling zeroes, the recommended model is the 'zero-inflated'-binomial (ZIB)/-Poisson (ZIP)/-negative binomial (ZINB) model. In contrast, the data might exhibit only an excess of sampling zeros and in such cases, a Hurdle-binomial (HB)/-Poisson (HP)/-negative binomial (HNB) model would be more appropriate.

For count data, the ZI model was introduced in Lambert's celebrated 1992 publication (9) and had been extensively used in other fields, such as econometrics (10), dentistry and public health (11,12), etc. The H model was first conceptualized in the context of Tobit (censored normal) models in (13), however the general form applicable for count regression first appeared in (14), with applications to model daily beverage consumptions. Since then, it has been exploited in many other disciplines, such as health-care utilization (15, 16), population studies (17), etc.

In the context of substance abuse, consider the NIDA-CTN Protocol 0015 focusing on women's treatment for trauma and substance use disorders. In order to assess sexual/HIV risk behavior defined as the 'number of unprotected sexual-activity, or USO in the past 30

days, prior to baseline assessments' using *Risk Behavior Survey* (RBS, 4), the authors considered a ZINB distribution (5). In related studies pertaining to reductions in drug use among young people living with HIV (6) or assessing risk for marijuana-related problems among college students (7), the (longitudinal) count response of substance use frequency exhibited zero-inflated profiles, and were modeled using a ZIP and a ZINB model, respectively.

Although both the zero-inflated (ZI) and hurdle (H) models can be viewed as finite mixture models (7) and they often produce indistinguishable fits revealed through goodness-of-fit measures, one model may be more applicable than the other based on the objectives and design of the study, i.e., the underlying mechanism by which 'excess zeroes' were generated. ZI models are more appropriate if one can clearly identify two underlying 'substance-use' processes: one which puts the subjects at risk and the other which influences the outcome in the at-risk population (8). For example in the marijuana-related study (7) described above, many college students have never been exposed, or do not currently use marijuana despite its high prevalence on college campuses. They contribute to 'current non-users' or structural zeroes, while those who do use the drug and thus engage in the risk behavior may report a wide range of their consumption levels, sometimes including zero. The zeros from the 'at-risk' subjects are the sampling zeros. Hence, a ZINB regression model was used, which involves two separate regressions related to appropriate probabilities and parameters through well-defined 'link functions' to identify covariates which might predict, (i) non-users (structural zeroes) and (ii) users (with possible sampling zeroes).

The H model is conceptually different and considers a two-step decision process. If all the subjects in a study can be considered to be at-risk, then the first step involves moving through a 'zero' realization stage (the sampling zeroes), i.e. whether there is 'no use' or 'some use'. Once the hurdle is crossed (i.e., 'use' has been established), the second stage determines the number of subsequent events/use. Thus, the H model assumes two separate processes determined through two separate regressions, where one generates the (i) sampling zeros, i.e. subjects who are believed to be temporarily use-free (with the potential for future use), and the other generating (ii) counts strictly > 0 . Note that, for both the ZI and H models, same (or different) sets of predictors/co-variables might determine (i) and (ii). In addition, subject-specific random effects are included in both the regressions, which might be connected through some variance-covariance structures to explain associations between (i) and (ii).

1.1. Statistical Framework

We first explore the statistical framework of ZI and H models. Let Y be a random variable (r.v.) which denotes the count responses (i.e. numbers of days of drug use per week) and y be the realized value of Y . Define $f(y_{ij}) = P(Y_{ij} = y_{ij})$ to be the probability mass function for the r.v. Y corresponding to the i th subject at the j th time-point (week), distributed as binomial (n, θ_{ij}), Poisson (λ_{ij}) or negative binomial (λ_{ij}, a), where a denotes a dispersion parameter. Note that as $a \rightarrow 0$, the Negative Binomial (NB) model converges to the Poisson (P) model. Then, the ZI distribution is defined as

$$P(Y_{ij}=y_{ij}) = \begin{cases} p_{ij} + (1 - p_{ij})f(0), & \text{if } y_{ij}=0 \\ (1 - p_{ij})f(y_{ij}), & \text{if } y_{ij}>0 \end{cases} \quad (1)$$

where, p_{ij} is the probability of excess 'structural' zeroes. ZI models put greater emphasis on the probability of observing a zero, which is determined as the sum of the probabilities of observing a structural and a sampling zero (see the expression corresponding to $y_{ij} = 0$ in (1)). Thus, the ZI model has the ability to pick up two different regimes (1) from which zero

counts arise. When p_{ij} equals 0, the ZI distribution reduces to the standard B, P or NB distributions; as p_{ij} approaches 1, the variance increases and the data exhibit greater over-dispersion (8).

The Hurdle model can be considered as a mixture of two distributions: a point mass at zero and a truncated-at-zero distribution (B, P or NB) for the positive counts. Hence, it is a modified count model (8) that conceptualizes two separate processes generating the zero and positive counts, the positive counts resulting after crossing the zero threshold or hurdle. Thus, the Hurdle distribution is defined as

$$P(Y_{ij}=y_{ij}) = \begin{cases} p_{ij}, & \text{if } y_{ij}=0 \\ (1 - p_{ij}) \frac{f(y_{ij})}{1-f(0)}, & \text{if } y_{ij}>0 \end{cases} \quad (2)$$

where, $1 - p_{ij}$ is defined as the probability of ‘crossing the hurdle’.

In general, the Hurdle model is an alternative way to model zero modifications (both inflation and deflation), whereas the ZI model can handle only zero-inflations. For a ZI model, we have

$$P(Y_{ij}=0) = p_{ij} + (1 - p_{ij})f(0) > f(0) \quad (3)$$

whereas, for a Hurdle model,

$$P(Y_{ij}=0) = p_{ij} \leq f(0) \quad (4)$$

In a ZI framework, there is no selection process leading to a zero or non-zero values; in contrast, within the Hurdle framework, there is a clear hierarchical process leading to the choice of $Y_{ij} = 0$ vs. $Y_{ij} > 0$, and afterwards a process that follows accounting for $Y_{ij} > 0$. Now in order to study the effect of potential covariates on the count outcomes, two separate regression models connect p_{ij} and λ_{ij} (or θ_{ij}) using respective link functions, to similar (or different) sets of covariates, primarily determined by the study experts. The HNB model allows more flexibility as compared to the HP model, given that the HNB model has an additional ‘dispersion’ parameter (a) which accounts for heterogeneity in addition to excess zeroes, whereas the HP model only accounts for excess zeroes. As earlier, when $a \rightarrow 0$, the HNB model converges to the HP model.

To our knowledge, there is no prior drug abuse literature explaining the similarities and differences in the conceptual framework of the ZI and H models. We motivate this paper using an observational study with a longitudinal follow-up component performed on 45 cocaine-dependent participants (18) at the Medical University of South Carolina (MUSC). The main objective of this paper is to facilitate understanding of these two types of ‘excess zero’ models to enable substance abuse researchers to choose between them while analyzing their data.

2. Materials and Methods

2.1. Participants

The 45 participants in this study were part of a larger, non-treatment study (18) examining the relationship between stress reactivity, hypothalamic-pituitary axis functioning, and cocaine dependence. They met DSM-IV criteria for cocaine dependence. Participants were recruited in South Carolina via newspaper and other media advertisements. Exclusion criteria included psychiatric conditions known to affect HPA axis functioning, pregnancy,

obesity, and other major medical disorders that could affect HPA axis. The study was approved by the MUSC Institutional Review Board (IRB).

2.2. Procedures and measures

Following baseline assessment, participants underwent an overnight stay at the General Clinical Research Center (GCRC) at MUSC where all were exposed to pharmacological and psychosocial laboratory stress tests. For more details on these tests, please refer to (18). Among the study participants, 44.4% were females and the mean [range] of age was 38 [20–53] years. The Structured Clinical Interview for DSM-IV was used to diagnose cocaine dependence and Time-Line-Follow-Back (TLFB) was used to assess the dollar value and frequency of cocaine use from approximately 11 weeks before to 5 weeks after the study, thereby creating a longitudinal profile of 16 time points for each subject. ‘Entry into the study’ was considered as an ‘intervention’ and a binary indicator of participation (coded as 1 if time $\geq 12^{\text{th}}$ week and 0 if time is $< 12^{\text{th}}$ week) was recorded. The main research question is whether study participation alters drug seeking behavior. About 30% of the reported cocaine intake frequency is 0, suggesting an ‘excess zero’ model for analysis.

3. Results

Since the outcome/response (number of days of drug abuse per week) in our motivating data is a non-negative inter-valued variable bounded with an upper-count of 7 (the maximum number of drug abuse days in a week), we started with a Binomial (B) model. We also considered Poisson (P) and negative binomial (NB) models, assuming the Poisson approximation to be appropriate for the binomial counts. A density histogram plot of the raw counts (Figure 1, left panel) reveal a zero-inflated nature, with also some inflation at 7 since our population consisted of cocaine-heavy participants who reported use for all 7 days in a week. The study participants were ‘at risk’ of using cocaine from the onset (i.e. observed zeros arising only from sampling), so a hurdle model was more appropriate than a ZI model. However, this does not imply that the HB, HP or HNB models are more appropriate than the B, P or NB models. Covariates considered in the B, P or NB models include an intercept, gender, age, time (in weeks) and a binary study participation indicator. A random effect term U_i following Normal $(0, \sigma^2)$ was added to control for subject specific heterogeneity, similar in spirit to a random intercept regression model. For the H models, the same set of covariates were considered in both the regressions on p_{ij} (logit link) and λ_{ij} (log link) for HP and HNB models, and p_{ij} and θ_{ij} (logit links for both). Two subject-specific random effects U_{1i} and U_{2i} were set to follow jointly a bivariate-normal density with mean vector $(0,0)$ and variance-covariance matrix given by $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

To proceed with the data analysis, we first fit various competing models, specifically B, P, NB, HP, HNB and HB, and used popular model selection criteria (lower is better) like Deviance, AIC and BIC, as well as Vuong’s test (19) to determine the best-fitting model.

The Vuong statistic is defined as $V = \bar{m} \sqrt{n} / S_m$ where $m_i = \ln[P_C(Y_i)/P_S(Y_i)]$, and P_C/P_S is the ratio of two competing model likelihoods with P_C as the comparison model and P_S as the standard model. The statistic m_i has mean $\bar{m}$ and standard deviation S_m . The statistic V follows a standard normal distribution, such that if $|V| < 1.96$, then the test does not favor one model over the other. If $|V| > 1.96$, then large positive values of V favor the comparison model while large negative values favor the standard model. The advantage of the Vuong’s statistic over other goodness-of-fit measures is that it considers the entire distribution of counts and not only the zero responses. Finally, we presented visual checks of goodness-of-fit by plotting the observed proportion minus the mean (expected) probability at each count for the

competing models. All the models were fitted using the NLMIXED procedure available in SAS Version 9.1.3. The code for our best-fitting model (HB) is available in the Appendix.

Tables 1 and 2 present various model comparison criteria and Vuong's statistics, respectively, for the competing models. Note that our goal is to determine the model that best fits the data, as determined by these criteria. From the AIC, BIC and deviance measures, it is clear that the B model provides the worst fit, hence it is not considered further. NB model provides an improved fit over the P model, however Vuong's statistic suggests no difference between them. The HP is chosen over both P and NB models and HNB over NB in Tables 1 and 2. Although HNB provides a better fit as compared to HP in Table 1, the Vuong's statistic suggests otherwise. From both Tables 1 and 2, the HB model seems to provide the best fit. Finally, from plots of observed minus expected probabilities for all counts (Figure 1, panel b), it is clear that the P model results in the worst fit, followed by the NB model. Both of these models demonstrate substantial underestimation at the two extremes (counts 0 and 7). Clearly, these models are not adapted to handle 'excess zeroes', or heterogeneity together. The HB model provides the best fit ('Observed – Expected' values lying closer to the $y = 0$ line), and substantially better than the HP or the HNB models, though all of them provides some overestimation at 0 and underestimation at 7. Henceforth, we choose the HB model as our model of choice, and discuss parameter estimates and results using this model.

Table 3 presents parameter estimates of fixed effects covariates, robust standard errors, variance components and p-values for both the 'Hurdle' and 'Binomial' parts of the HB model. Age and study participation are statistically significant at 0.05 level of significance for both regressions. The estimates of the variance components σ_1 , σ_2 and correlation ρ are 1.688, 1.726 and -0.319 , respectively. Table 4 contrasts results for the HB model. Parameters for this model are evaluated in terms of the 'odds ratio' (OR) of 'no cocaine use vs. use' for the hurdle part and the 'odds ratio' of 'experiencing one more day of cocaine use (given that use has already taken place)' for the conditional binomial part. If the [LCL,UCL] interval includes 1, the effect is non-significant at 5% alpha level. The results reveal that there is no difference in the odds (OR = 0.910, 95% CI = [0.293, 2.832]) of 'no cocaine use' for males vs. females; even after a subject has experienced cocaine use, there is also no difference by gender (OR = 0.957, 95% CI = [0.320, 2.862]) for further days of cocaine use. Similarly, time is also not a significant predictor in both the regressions. From the hurdle part, the OR of 'no cocaine use' is 0.527 for age, i.e., elderly subjects are more likely to experience cocaine use, which although contrary to some published findings, was true in this sample. Interestingly, conditional that use has already occurred, a 1 unit increase in age is associated with a reduced odds (OR = 0.611, 95% CI = [0.277, 0.946]) of further use. The hurdle part also reveals that the odds of no cocaine use is 6.31 times higher (95% CI = [1.388, 11.227]) after the participation in the study versus before. Once there is cocaine use, the odds of further use is lower (OR = 0.474, 95% CI = [0.313, 0.636]) after versus before participation.

4. Discussion

In this paper, we have described the Hurdle model as an alternative to the widely used zero-inflated model to handle the 'excess zero' situation often encountered in substance use studies. The decision criteria for the choice between these two models rest upon the assumption of the 'origin of zeroes' dictated by the study design. Often in substance use observational studies or clinical trials when the response exhibits an excess of zeroes, the origin of the zeroes are not considered in depth. For example, in the study of drug abuse reduction in HIV infected young people (6), the authors fitted a longitudinal ZIP model to assess intervention effects despite the fact that drug abuse was an entry criterion for this

study and thus all participants were active users at the time of recruitment. Hurdle models would have been more appropriate here.

In our study, the subjects that were recruited were cocaine dependent both before, and at the time of study. Hence, the study design lends itself to consider a Hurdle model for its analysis, because the excess zeroes appearing in the dataset can be convincingly believed to be free of any 'structural' zeroes. The zeroes here are due to random chances due to cocaine-free behavior during the study period, with the potential of future cocaine use.

Interestingly, we compared our HB fit to that of a zero-inflated Binomial (ZIB) model to this dataset (which assumes presence of some structural zeroes). The deviance, AIC and BIC values were 2305, 2331, and 2354.4, which is suggestive of the superior performance of the HB model over the ZIB. Thus, it is recommended to properly consider the design aspects (viz. 'origin of zeroes', etc) while analyzing 'excess zero' situations, and seek a model that provides the best fit to the dataset, although the parameter estimates from both the models might represent closely related facts.

Although the ZIP model appears to be quite sophisticated (with an underlying latent framework depicting two different sources of zeros), detailed simulation studies comparing H and ZI models for Poisson regression (20) reveal the unstable nature, and the potential problems associated with the ZIP formulation, primarily because there is no distinct selection process leading to zero or non-zero values. On the contrary, the H model motivated by a distinct hierarchical decision process (1) has a very stable behavior/performance. Whether this observation is similar for bounded 'Binomial' counts (as in our case) warrants detailed simulation studies which is beyond the scope of this paper. However, we believe that the results from those simulations will be similar, and will be considered elsewhere. Nevertheless, the methodological issues discussed here should be a guiding force while considering analysis of 'excess zero' situations for bounded or unbounded counts in clinical trials such as the CTN, as well as longitudinal studies on substance use like the one presented.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by grant 5U10-DA013727, "South Carolina Consortium of the Clinical Trials Network", from the NIH/NIDA. Bandyopadhyay acknowledges support from grant P32-RR017696 from NIH/NCRR. We thank Amy Wahlquist for her assistance in creating plots.

References

1. Alfö M, Maruotti A. Two-part regression models for longitudinal zero-inflated count data. *The Canadian Journal of Statistics*. 2010; 38(2):197–216.
2. Cameron, AC.; Trivedi, PK. *Regression analysis of count data*. Cambridge: Cambridge University Press; 1998.
3. Delucchi KL, Bostrom A. Methods for analysis of skewed data distributions in psychiatric clinical studies: working with many zero values. *American Journal of Psychiatry*. 2004; 161:1159–1168. [PubMed: 15229044]
4. Booth RE, Watters JK, Chitwood DD. HIV risk-related sex behaviors among injection drug users, crack smokers, and injection drug users who smoke crack. *American Journal of Public Health*. 1993; 83:1144–1148. [PubMed: 8342724]

5. Hien DA, Campbell ANC, Killen T, Hu M-C, Hansen C, Jiang H, Hatch-Mailette M, Miele GM, Cohen LR, Gan W, Resko SM, DiBono M, Wells EA, Nunes EV. The impact of trauma-focused group therapy upon HIV sexual risk behaviors in the NIDA clinical trials network "Women and Trauma" multi-site study. *AIDS and Behavior*. 2010; 14(2):421–430. [PubMed: 19452271]
6. Comulada WS, Weiss RE, Cumberland W, Rotheram-Borus MJ. Reductions in drug use among young people living with HIV. *The American Journal of Drug and Alcohol Abuse*. 2007; 33:493–501. [PubMed: 17613977]
7. McLachlan, G.; Peel, D. *Finite Mixture Models*. New York: John Wiley and Sons, Inc; 2000.
8. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*. 2006; 16:463–481. [PubMed: 16892908]
9. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992; 34:1–14.
10. Harris MN, Zhao X. A zero-inflated ordered probit model, with an application to modeling tobacco consumption. *Journal of Econometrics*. 2007; 141:1073–1099.
11. Böhning D, Dietz E, Schlattmann P, Mendonca L, Kirchner U. The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society, Series A*. 1999; 162(2):195–209.
12. Moulton LH, Curriero FC, Barroso PF. Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research*. 2002; 11:317–325. [PubMed: 12197299]
13. Cragg JG. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*. 1971; 39:829–844.
14. Mullahy J. Specification and testing of some modified count data models. *Journal of Econometrics*. 1986; 33:341–365.
15. Brown CJ, Pagan JA, Rodriguez-Oreggia E. The decision-making process of health care utilization in Mexico. *Health Policy*. 2005; 72(1):81–91. [PubMed: 15760701]
16. Liu T-C, Chen C-S. An analysis of private health insurance purchasing decisions with national health insurance in Taiwan. *Social Science and Medicine*. 2002; 55:755–774. [PubMed: 12190269]
17. Arulampalam W, Booth A. Who gets over the training hurdle? A study of the training experiences of young men and women in Britain. *Journal of Population Economics*. 1997; 10:197–217.
18. DeSantis SM, Bandyopadhyay D, Back SE, Brady KT. Non-treatment clinical studies of cocaine and methamphetamine users are protective against follow-up substance abuse. *Drug and Alcohol Dependence*. 2009; 105:227–233. [PubMed: 19717245]
19. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 1989; 57(2):307–333.
20. Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*. 2005; 5:1–19.

Figure 1. Plots of cocaine intake data. The left panel displays density histogram of the raw counts (frequency of cocaine intake). The right panel compares plots of “Observed-Expected predicted probabilities” for each of the counts for the 5 competing models.

Table 1

Model-comparison measures (Deviance, AIC and BIC) for the competing models.

Models	Fit Statistics		
	Deviance	AIC	BIC
Binomial (B)	2942.0	2954.0	2964.8
Poisson (P)	2742.0	2754.0	2764.8
Negative Binomial (NB)	2711.8	2725.8	2738.5
Hurdle Poisson (HP)	2502.0	2528.0	2551.5
Hurdle Negative Binomial (HNB)	2434.5	2462.5	2487.8
Hurdle Binomial (HB)	2305.0	2331.0	2354.4

Table 2

Table of Vuong's Statistics. In the pairwise comparisons, the first component is the comparison model, while the second component is the standard model.

Pairwise comparison	Vuong Statistic (V)	p-value
B vs P	-3.428	0.0006
NB vs B	2.753	0.0058
P vs NB	-0.43001	0.667
P vs HP	-9.528	< 0.0001
NB vs HP	-12.189	< 0.0001
NB vs HNB	-12.012	< 0.0001
HP vs HNB	6.3885	< 0.0001
HNB vs HB	-3.8347	0.00012

Table 3

Parameter estimates, robust standard errors (SE) and associated p-values for our best-fitting model: Hurdle Binomial (HB) model. 's' denotes significant covariates.

	Parameter estimate	Robust SE	p-value
<i>Hurdle Part</i>			
Intercept ^s	-1.984	0.495	0.0002
Gender	-0.094	0.562	0.868
Age ^s	-0.64	0.286	0.031
Time	0.011	0.039	0.785
Participation ^s	1.84	0.386	< 0.0001
<i>Binomial part</i>			
Intercept	0.151	0.417	0.718
Gender	-0.043	0.543	0.936
Age ^s	-0.491	0.234	0.042
Time	-0.007	0.016	0.654
Participation ^s	-0.746	0.168	<0.0001
σ_1^s	1.688	0.254	< 0.0001
σ_2^s	1.726	0.217	< 0.0001
ρ^s	-0.319	0.156	0.047

Table 4

Result for selected contrasts using the Hurdle Binomial (HB) Model. The estimates of both regressions are in terms of Odds Ratio (OR), though they would represent different events. LCL and UCL stand for lower and upper 95% confidence limits respectively. 's' denotes significance and 'Part.' denotes the Participation covariate.

	Estimates	LCL	UCL
<i>Hurdle Part (OR)</i>			
Male vs Female	0.910	0.293	2.832
Age ^s	0.527	0.222	0.832
Time	1.010	0.931	1.089
Part. vs No Part.. ^s	6.308	1.388	11.227
<i>Binomial part (OR)</i>			
Male vs Female	0.957	0.320	2.862
Age ^s	0.611	0.277	0.946
Time	0.992	0.961	1.025
Part. vs No Part.. ^s	0.474	0.313	0.636