A MODEL SELECTION PARADIGM FOR MODELING RECURRENT ADENOMA DATA IN POLYP PREVENTION TRIALS

By

Christopher L Davidson

A Thesis Submitted to the Faculty of the

MEL AND ENID ZUCKERMAN COLLEGE OF PUBLIC HEALTH

In Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE
WITH A MAJOR IN BIOSTATISTICS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2012

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Christopher L Davidson

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Chiu Hsieh (Paul) Hsu — May 7, 2012

Chiu-Hsieh (Paul) Hsu, PhD — Date

Associate Professor of Biostatistics

ACKNOWLEDGEMENTS

I offer my sincere gratitude to my committee members. First I wish to thank Dr Paul Hsu for his thoughtful instruction, support, and advice not only with this project, but throughout my graduate career. I will always value the instruction I received from Dr Hsu. I am particularly grateful to Dr Denise Roe for her continued support throughout my years as a graduate student and her generosity in obtaining a position for me in the Health Promotion Sciences division as a Data Analyst while pursing my graduate studies. She too has provided invaluable guidance in statistical analysis and reporting of statistical results. Thank you to Dr Elizabeth Jacobs for taking time to be part of my thesis committee and providing valuable feedback so that I am successful in completing my degree.

I would like to thank the many wonderful students of the University of Arizona who are on a similar path as I in pursuit of advanced training in statistics and biostatistics. I will always value the many interactions and exchanges of ideas we shared and will continue to share in the future. I would also like to thank my family and friends outside of school. Without their support and encouragement this project would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES....5
LIST OF TABLES....6
ABSTRACT....7
INTRODUCTION....8
METHODS....11
Models for Count Data....11
Generalized Linear Models....12
Overdispersion....13
Robust Standard Errors for Handling Overdispersion....14
Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models....15
Assessment of Overdispersion, Zero-Inflation, and Model Selection....20
Model Selection Paradigm....23
The Ursodeoxycholic Acid Clinical Trial....25
Tests Employed in Model Selection....26
Exploratory Data Analysis....27
DATA ANALYSIS....29
Results....29
Model Fitting and Selection....37
Interpretation of the ZINB Model....41
DISCUSSION....44
Rationale for Approach....44
Limitations of the Approach....45
Misclassification....47
Limitations of the Example Data....47
APPENDIX 1....49
Models (No Offset)....49
Models with Offset....52
APPENDIX 2....56
REFERENCES....63

LIST OF FIGURES

Figure 1 Relationship of Models based on Moments 23

Figure 2 Model Selection Paradigm 25

Figure 3 Adenoma Counts by Treatment Group 30

Figure 4 Distributions of Continuous Variables 33

Figure 5 Recurrent Adenoma Count by Levels of Binary Covariates 34

Figure 6 Assessment of Linearity for Recurrent Adenoma Count with Age and BMI 35

Figure 7 Comparison between Observed Probabilities and Univariate Poisson Probabilities 36

Figure 8 Observed vs Predicted Proportions for all Models (with Offset) 41

LIST OF TABLES

Table 1 Example of Interpreting a Zero-inflated model with Recurrent Adenoma Count as the Outcome and Sex=Male as a Covariate after Exponentiating the Model 19

Table 2 Distribution of Recurrent Adenoma Count by Treatment Group 29

Table 3 Summary Statistics 32

Table 4 Results of Model Selection 39

Table 5 Zero-Inflated Negative Binomial Model (with Offset) 43

Table A1 Poisson GLM with Robust Sandwich estimators for SE – No Offset 49

Table A2 Negative Binomial GLM – No Offset 49

Table A3 Zero-Inflated Poisson – No Offset 50

Table A4 Zero-Inflated Negative Binomial – No Offset 51

Table A5 Poisson GLM with Robust Sandwich estimators for SE 52

Table A6 Negative Binomial GLM 52

Table A7 Zero-Inflated Poisson 53

Table A8 Summary Statistics (Full Dataset) 54

Table A9 Recurrent Adenoma Count by Treatment Group (Full Dataset) 55

ABSTRACT

Colorectal polyp prevention trials (PPTs) are randomized, placebo-controlled clinical trials that evaluate some chemo-preventive agent and include participants who will be followed for at least 3 years to compare the recurrence rates (counts) of adenomas. A large proportion of zero counts will likely be observed in both groups at the end of the observation period. Poisson general linear models (GLMs) are usually employed for estimation of recurrence in PPTs. Other models, including the negative binomial (NB2), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) may be better suited to handle zero-inflation or other forms of overdispersion that are common in count data. A model selection paradigm that determines a statistical approach for choosing the best fitting model for recurrence data is described. An example using a subset from a large Phase III clinical trial indicated that the ZINB model was the best fitting model for the data.

INTRODUCTION

Colorectal polyp prevention trials (PPT) are generally designed to evaluate some potential chemo-preventative agent that may prevent or reduce the number of adenomas (a benign tumor that has a probability of becoming malignant over time) that would otherwise recur in patients who suffer from colorectal diseases. PPTs are typically randomized, placebo-controlled clinical trials designed to include participants who will be followed for at least 3 years to determine the recurrence rate of colorectal polyps that are adenomas (1-3). The number of recurrent colorectal adenomas is measured by performing a colonoscopy within 6 months of the 3-year anniversary of the baseline colonoscopy to remove all incident polyps and then analyze them to determine if any are adenomas. In some PPTs, patients need to undergo a colonoscopy 6 months to a year after being randomized to treatment in order to clear any polyps that may have gone undetected. Therefore, any follow-up colonoscopy occurring within 6 months after randomization is considered as part of the baseline colonoscopy in this project.

Variation in the length of follow-up times adds to the challenge of analyzing data collected in PPTs. Most patents receive the follow-up colonoscopy as scheduled within 6 months of the 3-year follow-up end point. However, some patients will have follow-up colonoscopy early because of family history of colorectal cancer or previous history of polyps. The variability in the follow-up times means that patients who have a zero count for a follow-up time of 1 year (3-year count is censored) should not be treated the same as a patient who has a zero count for follow-up time of 3 years. Hsu (4) discussed biases associated with censored patients and shows that results obtained from methods such as logistic regression and survival analysis may be inferior when trying to estimate recurrence rates in a PPT population. Hsu used a weighted zero-

inflated Poisson model where the weight function was derived from a nonparametric survival curve in order to adjust for the variable lengths of follow-up.

The most popular statistical methodology currently employed for the analysis of count data and estimation of adenoma recurrence in PPTs is generalized linear models (GLM). The outcome is usually modeled as independent and identically distributed Poisson random variables when the actual counts will be analyzed or as the log odds of recurrence (logistic regression) when only recurrence is of interest (3, 5). Modeling the outcome as a binary variable when the count level data are available is not taking advantage of all the information in the sample. The limitation with Poisson regression is that the distribution is constrained to having the mean and variance being equal. Additionally, the Poisson distribution assumes a limited number of zeros that may render an inappropriate fit to most PPT count data structures. The negative binomial distribution can sometimes provide a better fit to count data where the variance is greater than the mean, but may not provide the best fit when a large proportion of zeros is included. Hence, it is important to systematically study the performances of the existing statistical models for analyzing PPT count data.

The objectives of this project are to (one) describe statistical models that are appropriate for modeling PPT count data including Poisson, negative binomial, zero-inflated Poisson (ZIP), and zero-inflated negative binomial models (ZINB), (two) describe a model selection paradigm based on score- and likelihood-based tests for selecting the model among the Poisson, negative binomial, ZIP, or ZINB models that best fits the data, and (three) apply the model selection paradigm to data collected in a large PPT. The primary hypothesis is that the best fitting statistical model for PPT count data is one that can accommodate the zero-inflated structure of the outcome variable.

This thesis is organized as follows: All applicable models for count data are defined and reviewed in the Methods section, and model selection procedures are described. In the Data Analysis section, a large PPT, studying ursodeoxycholic acid (UDCA) is briefly described, and the model fitting and selection procedures for count data described in the Methods section are demonstrated with the UDCA data. The performance and limitations of the models and model selection methods are addressed in the Discussion section.

METHODS

Models for Count Data

Count data are traditionally modeled using the Poisson distribution, which is defined as follows:

$$\Pr(Y_i = y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{\Gamma(y_i + 1)}, \qquad 0 \leq \lambda < \infty,$$

where y_i is the outcome, $E(Y_i) = \lambda_i$ and $Var(Y_i) = \lambda_i$. This model is constrained to having the variance of the outcome equal to the mean, a condition sometimes referred to as equidispersion. It is not common in practice for data to be equi-dispersed; however, the Poisson model can be used when slight deviations from equidispersion are observed.

When the Poisson model provides an inadequate fit to the data, the negative binomial distribution is usually employed. The most widely used parameterization of the negative binomial is

$$\Pr(Y_i = y_i|\lambda_i, \alpha) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)}\left(\frac{1}{1 + \alpha\lambda_i}\right)^{1/\alpha}\left(\frac{\alpha\lambda_i}{1 + \alpha\lambda_i}\right)^{y_i}, \qquad 0 \leq \lambda < \infty, \qquad \alpha \geq 0$$

where y_i is the outcome, $E(Y_i) = \lambda_i$ and $VAR(Y_i) = \lambda_i + \alpha\lambda_i^2$. The Poisson distribution can be thought of as a special case of the negative binomial distribution when comparing the first 2 moments of each distribution. Statisticians refer to this parameterization as NB2 because of the quadratic term in the variance function (6), and this nomenclature will be used throughout this paper. The parameter α is the heterogeneity parameter, which is directly proportional to the amount of overdispersion in the data. As $\alpha \to 0$, the NB2 distribution converges to the Poisson, thus the Poisson and NB2 models are nested. The NB2 can accommodate data that are over-dispersed relative to the Poisson, which

makes it a popular alternative to the Poisson model when the data exhibit overdispersion.

Generalized Linear Models

Because the Poisson and negative binomial distributions are members of the exponential family of distributions, they can be formulated into generalized linear models (GLM). In GLMs, the dependence of the conditional mean $\mathrm{E}(\mathrm{y_i}|\mathrm{x_i}) = \mu_\mathrm{i}$ on a vector of predictors x_i is written as

$$\mathrm{g}(\mu_\mathrm{i}) = \beta_0 + \mathrm{x_i^T}\boldsymbol{\beta}$$

where g() is a link function, β_0 is the intercept, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\mathrm{x_i}$ is the row vector of covariates for the i^{th} row of the design matrix. The canonical link for both the Poisson and NB2 GLMs is the natural logarithm, $\ln(\mu_\mathrm{i}) = \eta_\mathrm{i} = \mathrm{x_i^T}\boldsymbol{\beta}$. Taking the inverse of the link solves $\mu_\mathrm{i} = \exp\left(\mathrm{x_i^T}\boldsymbol{\beta}\right)$, which gives the incidence rate ratio (IRR) relative to the reference group. The incidence rate is the rate at which counts enter the time period or space being modeled.

When parameterized as rate models, both the Poisson and NB2 are modified to incorporate the time or space variable as an offset (i.e. each observation may have a different exposure to the conditions of the experiment). With an offset included, the log-linear GLM is then constructed as

$$\log(\mu_\mathrm{i}/\mathrm{t_i}) = \beta_0 + \mathrm{x_i^T}\boldsymbol{\beta}$$

$$\log(\mu_\mathrm{i}) - \log(\mathrm{t_i}) = \beta_0 + \mathrm{x_i^T}\boldsymbol{\beta}$$

$$\log(\mu_\mathrm{i}) = \beta_0 + \mathrm{x_i^T}\boldsymbol{\beta} + \log(\mathrm{t_i})$$

Thus, the log-offset term $\log(\mathrm{t_i})$ is entered directly into the log-linear model, and its

coefficient is constrained to be 1. The intercept term is usually incorporated into $x_i^T \boldsymbol{\beta}$ so that $\log(\mu_i/t_i) = x_i^T \boldsymbol{\beta}$. When the model is exponentiated, $\mu_i = \exp\left(x_i^T \boldsymbol{\beta} + \log(t_i)\right)$, which shows that the conditional mean will depend on variable follow-up times. When no offset is included, $t_i = 1$ for all i (i.e. the assumption is made that all observations had the same exposure).

Overdispersion

Overdispersion in linear models for discrete random variables has been discussed at length in the statistical literature (7-11). Although not formally defined, overdispersion is generally recognized when the variance exceeds that of the predicted variance estimated by the fitted model, or more generally when the variance is larger than the mean relative to the assumed distribution. In the case of the Poisson model however, the most widely used model for count data, it would be unrealistic to expect that the predicted variance would be exactly equal to the observed variance. In some cases, the Poisson model will adequately fit slightly over-dispersed count data. Since some degree of overdispersion will almost always exist in models for count data, it is the extent of overdispersion, as well as the source of overdispersion, that become important points of consideration when choosing an appropriate model or modifications to an existing model to accommodate overdispersion. Furthermore, if a statistical model fails to adequately account for overdispersion, the resulting standard errors of the coefficients estimated in GLMs will be underestimated (6, 10, 12, 13), even though the parameter estimates for the coefficients are not affected, which may lead to concluding that a predictor is statistically significant when in fact it is not (i.e. inflation of the Type I error rate).

The extent of overdispersion in a GLM for count data (ie binomial, multinomial, Poisson, or negative binomial) is usually assessed using the Pearson X^2 goodness-of-fit statistic and the model residual degrees of freedom (6). The ratio of the sum of the Pearson residuals, $X^2 = \sum_i (O_i - E_I)^2/E_i$, to the residual degrees of freedom (Res DF) should be close to 1 in a model that adequately accounts for the amount of overdispersion in the data (7). It has been argued that values of the dispersion statistic ≥ 1.25 in moderate sized models may warrant modification, and models with a large number of observations may be over dispersed if the value of the dispersion statistic is as low as 1.05 (6). Overdispersion in linear models for counts may arise from several sources, but all sources are generally thought to be some form of clustering in the data (6, 7, 11). However, with a limited number of predictors the clustering is considered as unobserved heterogeneity.

Robust Standard Errors for Handling Overdispersion

Several methods have been proposed for dealing with overdispersion in count models, particularly in Poisson models including quasi-Poisson fitting and robust "sandwich" estimators (14, 15). The sandwich estimators are used to adjust standard errors when a Poisson model exhibits overdispersion and the Poisson distribution continues to be the distribution of choice for the model. The estimators are called sandwich estimators because of the form of the resulting matrix outer product used to compute the estimates. The outer product is constructed so that the gradient (first derivative(s)) of the model likelihood is "sandwiched" between two copies of the Hessian matrix (matrix of partial second derivatives of the model likelihood). After computation, the sandwich estimators of the standard errors replace the standard errors computed by

ML, but the parameter estimates remain as they were in the original model. The adjusted standard errors are inflated, thus decreasing the effect of overdispersion (and the Type I error rate) and allowing a more robust statistical inference to be made with respect to the parameter estimates. However, with PPT count data (ie data that may contain excess zeros), the fit may not be optimal given current statistical methods such as ZIP or ZINB, or hurdle models (2-part truncated-at-zero models).

Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models

Because overdispersion in count data is sometimes the result of more zeros than would be predicted using standard log-linear GLMs, ZIP and ZINB were investigated as part of modeling paradigm for PPT count data. The problem of more zeros than would be predicted by a Poisson log-linear model was initially investigated with development of the ZIP model (16) and later extended to the ZINB model (17) as a way of accounting for unobserved heterogeneity likely caused by the presence of the excess zeros in the response. The ZIP and ZINB models are best understood as a methodological approach to a specific application, which in this case will be accounting for the high number of patients in PPTs who do not have recurring adenomas within the observation time of the study. Thus we assume that the zero component in the distribution of recurrent adenoma counts arises from a 2-part zero-generating process: 1 part assumes that some subjects have zero probability of recurrence (ie the certain zero group), and the other part assumes that the remaining patients have a probability of recurrence dictated by a discrete count model such as the Poisson or NB2.

Given n subjects in a polyp prevention trial, each subject i will have Y_i observable recurrent adenomas such that $i = 1, \ldots, n$. The zero-inflated models can be formulated

as a missing data problem where Ψ_i, an unobservable latent variable, represents the true number of recurrent adenomas for each subject i. Because over counting of adenomas is not assumed to occur with colonoscopy methodology (4) when there are no adenomas, it can be assumed that $\Pr(Y_i = 0|\Psi_i = 0) = 1$. It is further assumed that $[Y_i \mid \Psi_i > 0]$ is distributed as either Poisson (λ) for ZIP models or negative binomial (λ, α) for ZINB models. If $\Delta_i = I(Y_i = 0)$, where $I(.)$ is the indicator function, then the observable random vector for subject i is $\mathbf{O} = (Y_i, \Delta_i)$, and the observed vector is $O = (y_i, \delta_i)$, where $\delta_i = I(y_i = 0)$ is the indicator for a zero count. It is assumed that the n subjects are a random sample from the population in which the results will be generalized, and each subject i is independent. Let $\omega = \Pr(\Psi_i = 0)$ represent the probability of an "always zero" count and $1 - \omega = \Pr(\Psi_i > 0)$ is the probability of being in the population that may experience a recurrent adenoma or simply the true recurrence rate. The marginal distribution of Y_i in the ZIP model is expressed as follows:

$$\Pr(Y_i = y_i|\omega, \lambda) = \Pr(Y_i = y_i|\Psi_i = 0) * \Pr(\Psi_i = 0) + \Pr(Y_i = y_i|\Psi_i > 0) * \Pr(\Psi_i > 0)$$

$$= \begin{cases} \omega_i + (1 - \omega_i)e^{-\lambda_i}, & \text{if } y_i = 0 \\ (1 - \omega_i)\, e^{-\lambda_i}\lambda_i^{y_i}/\Gamma(y_i + 1), & \text{if } y_i > 0 \end{cases}$$

where $\boldsymbol{\lambda}$ is modeled as a log-linear Poisson model, $\log(\lambda_i) = x_i^T\beta$, conditional on some set of covariates (**X**), $\boldsymbol{\omega}$ is modeled as the log odds (logit) of being a zero count, $\ln(\omega_i/(1 - \omega_i)) = z_i^T\gamma_i$,for some set of covariates (**Z**). **X** may or may not be equal to **Z**.

The expected value of the ZIP model is $E(Y_i) = (1 - \omega_i)\lambda_i$ and the variance is $Var(Y_i) = (1 - \omega_i)(\lambda_i + \omega_i\lambda_i^2)$. Substituting the negative binomial distribution for the Poisson distribution yields the ZINB model. With the NB2 parameterization, the ZINB model can be expressed as

$$\Pr(Y_i = y_i|\omega, \lambda, \alpha)$$

$$= \begin{cases} \omega_i + (1-\omega_i)(1+\alpha\lambda_i)^{-1/\alpha} & y_i = 0, \\ (1-\omega_i)\dfrac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}(1+\alpha\lambda_i)^{-1/\alpha}(\alpha\lambda_i/(1+\alpha\lambda_i))^{y_i} & y_i > 0 \end{cases}$$

The expected value of the ZINB is $E(Y_i) = (1-\omega_i)\lambda_i$, and the variance is $\mathrm{Var}(Y_i) = (1-\omega_i)\{\lambda_i(1+\alpha\lambda_i) + \omega_i\lambda_i^2)\}$. Thus, the variance of the ZINB can accommodate overdispersion due to sources other than excess zeros as evidenced by the parameterization of the variance function, which is analogous to the parameterization in the GLMs. Also analogous to the GLM construction is the fact that the ZINB reduces to the ZIP model as $\alpha \to 0$, which shows that ZIP is nested within ZINB. The zero-inflated models however are not members of the exponential family (6) and thus are not GLMs, but they can be estimated by maximum likelihood via the Newton-Raphson or EM algorithms.

When a zero-inflated model is fit to a dataset, 1 set of parameters is estimated for the count component, and 1 set of parameters is estimated for the binary (probability of a zero count) component. The covariates may or may not be the same in both components depending on the interest of the analyst. The interpretation of zero-inflated models is similar to their GLM counterparts, but one needs to be careful, particularly with respect to the binary component. If the model is parameterized such that the binary component is estimating the probability of a certain zero count, then the interpretation is the opposite of what it would be in a traditional logistic regression model, which usually models the log odds of having an event ($\Pr(Y_i = 1)/(1-\Pr(Y_i = 1))$). Additionally, a parameter estimate may be statistically significant in 1 component and not in the other. For example, suppose a zero-inflated model is fitted to recurrent adenoma data from a

PPT with sex (female as the reference group) as a covariate in both components. If the model is exponentiated, and we can assume no false negatives ($\Pr(Y_i = 0|\Psi_i = 0) = 1$ when there are no adenomas), then the interpretation would fall into 1 of the categories described in Table 1 assuming that the covariate was significant in at least 1 component. When covariates are not significant in either component, then their interpretation has little practical value.

Table 1. Example of Interpreting a Zero-inflated model with Recurrent Adenoma Count as the Outcome and Sex=Male as a Covariate after Exponentiating the Model.

Significance of Sex Covariate in Components	Sign of Male Covariate	Interpretation*
Significant in Count and Binary	(+) in Count (+) in Binary	Males had a higher rate of recurrent adenomas and were more likely to be in the certain zero group (ie the group with zero probability of recurrence) compared with females.
	(+) in Count (-) in Binary	Males had a higher rate of recurrent adenomas and were less likely to be in the certain zero group compared with females.
	(-) in Count (+) in Binary	Males had a lower rate of recurrent adenomas and were more likely to be in the certain zero group compared with females.
	(-) in Count (-) in Binary	Males had a lower rate of recurrent adenomas and were less likely to be in the certain zero group compared with females.
Significant in Count and Not in Binary	(+) in Count (+) in Binary	Males had a higher rate of recurrent adenomas compared with females.
	(+) in Count (-) in Binary	
	(-) in Count (+) in Binary	Males had a lower rate of recurrent adenomas compared with females.
	(-) in Count (-) in Binary	
Significant in Binary and Not in Count	(+) in Count (+) in Binary	The odds of being male and being in the certain zero group were higher than the odds of being female and being in the certain zero group.
	(-) in Count (+) in Binary	
	(+) in Count (-) in Binary	The odds of being male and being in the certain zero group were lower than the odds of being female and being in the certain zero group.
	(-) in Count (-) in Binary	

*Assumes no false negatives (ie $\Pr(Y_i = 0 | \Psi_i = 0) = 1$) when there are no adenomas

Assessment of Overdispersion, Zero-Inflation, and Model Selection

The extent of overdispersion in the Poisson models may be assessed using the model dispersion statistic (i.e. sum of Pearson residuals divided by the model residual degrees of freedom). The score test for overdispersion in the Poisson model (18) is constructed as follows:

$$Z_i = \frac{(y_i - \mu_i)^2 - y_i}{\mu_i\sqrt{2}}$$

where **Z** is the vector of z-scores distributed standard normal under the null hypothesis, y_i is the recurrent adenoma count for the i^{th} observation, and μ_i is the fitted value from the Poisson model for the i^{th} observation. The vector of z-scores is then regressed as a simple linear model using an intercept-only model and the p-value of the resulting Wald test (H0: Z=0) is used to make inferences regarding the presence of overdispersion. The test is equivalent to a 2-sided t-test at the 0.05 level of significance.

The score test for zero-inflation in a Poisson model (19) may be used to assess zero-inflation, and determine the need for fitting zero-inflated models. The score statistic, $S(\widehat{\boldsymbol{\beta}})$, for this test is constructed as

$$S(\widehat{\boldsymbol{\beta}}) = \frac{\left\{\sum_{i=1}^{n}\left(\frac{I_{(y_i=0)}}{e^{-\hat{\lambda}_i}} - 1\right)\right\}^2}{\left\{\sum_{i=1}^{n}\left(\frac{1}{e^{-\hat{\lambda}_i}} - 1\right)\right\} - \hat{\boldsymbol{\lambda}}^{\mathbf{T}}\mathbf{X}\left[\mathbf{X}^{\mathbf{T}}\mathbf{diag}(\hat{\boldsymbol{\lambda}})\mathbf{X}\right]^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\hat{\boldsymbol{\lambda}}}$$

where $\widehat{\boldsymbol{\beta}}$ is the vector of parameter estimates, and $\hat{\boldsymbol{\lambda}}$ is the vector of fitted values under the null hypothesis that $\boldsymbol{\omega} = 0$ in the alternative ZIP model. Under H0, $S(\widehat{\boldsymbol{\beta}}) \sim X_1^2$.

Because score tests are local tests and are usually less conservative than global tests, the LRT was also used to compare the Poisson and the NB2 as these are nested models through their variance functions. The test statistic (τ) for the LRT is $\tau = -2(\mathrm{LL}_{\mathrm{Poison}} - \mathrm{LL}_{\mathrm{NB2}})$, where $\mathrm{LL}_{\mathrm{Poison}}$ is the log-likelihood for the Poisson model, and $\mathrm{LL}_{\mathrm{NB2}}$ is the log-likelihood for the NB2 model. The null hypothesis for the LRT is that $\alpha = 0$ in $\mathrm{V}(\mathrm{Y}) = \mu + \alpha\mu^2$ (the variance function for the NB2 model), which would equate to the Poisson variance under the null hypothesis. The test statistic has a chi-squared (df=1) distribution under the null hypothesis.

For non-nested models (i.e. Poisson and ZIP, NB2 and ZIP, or NB2 and ZINB), the Vuong test (20) is used to test for zero-inflation and preferred fit. The Vuong test is based on the Kullback-Leibler information criteria, which is a measure of "distance" between an estimated model and the true model. The definition of non-nested in this paradigm is such that 2 models are non-nested if one model cannot be reduced to the other model by imposing a set of linear restriction on the parameter vector (21). The test is not valid for partially nested or overlapping models. The null hypothesis for the Vuong test can be stated such that neither model is closer to the true model. The alternative hypothesis is that 1 of the models is closer to the true model. Which model is closest to the true model depends on the sign of the resulting test statistic. The test statistic (τ_{Vuong}) is constructed as follows:

Under H0:

$$\tau_{\mathrm{Vuong}} = \frac{\mathrm{LR}_{\mathrm{n}}(\hat{\beta}_{\mathrm{n}}, \hat{\gamma}_{\mathrm{n}})}{(\sqrt{\mathrm{n}})\hat{\omega}_{\mathrm{n}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$$

where

$$\mathrm{LR}_{\mathrm{n}}(\hat{\beta}_{\mathrm{n}}, \hat{\gamma}_{\mathrm{n}}) \equiv \mathrm{L}_{\mathrm{n}}^{\mathrm{f}}(\hat{\beta}_{\mathrm{n}}) - \mathrm{L}_{\mathrm{n}}^{\mathrm{g}}(\hat{\gamma}_{\mathrm{n}})$$

With this formulation, $\mathrm{LR_n}(\cdot)$ is a log-likelihood ratio such that $\mathrm{L_n^f}\left(\hat{\beta}_\mathrm{n}\right)$ are the predicted probabilities for the more complex model (e.g. ZIP), and $\mathrm{L_n^g}(\hat{\gamma}_\mathrm{n})$ are the predicted probabilities for the less complex model (e.g. Poisson GLM), and

$$\hat{\omega}_\mathrm{n}^2 \equiv \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[\ln\frac{\mathrm{f}\left(\mathrm{Y_i}|\mathrm{X_i};\hat{\beta}_\mathrm{n}\right)}{\mathrm{g}(\mathrm{Y_i}|\mathrm{Z_i};\hat{\gamma}_\mathrm{n})}\right]^2 - \left[\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\ln\frac{\mathrm{f}\left(\mathrm{Y_i}|\mathrm{X_i};\hat{\beta}_\mathrm{n}\right)}{\mathrm{g}(\mathrm{Y_i}|\mathrm{Z_i};\hat{\gamma}_\mathrm{n})}\right]^2$$

Large positive values of τ_{Vuong}favor the more complex model and large negative values favor the less complex model. Values close to zero in absolute value favor neither model. In a test comparing ZIP with the Poisson GLM, for example, if the test statistic ≥ 1.96, then it can be concluded that the ZIP model is favored over the Poisson GLM at the 0.05 level of significance. If the test statistic is ≤ -1.96, then the Poisson GLM is favored over the ZIP model at the 0.05 level of significance.

The Akaike information criteria (AIC) and Bayesian information criteria (BIC) are calculated as supplemental data to compare the fit for each model after assessing overdispersion and zero-inflation. The AIC statistic (22) is defined as $-2\mathrm{LL}+2\mathrm{k}$, where LL is the log-likelihood for the model and k is the number of predictors including the intercept. The BIC statistic is similar to the AIC statistic and is defined as $-2\mathrm{LL}+\mathrm{k}*\mathrm{Ln}(\mathrm{n})$, where the symbols have the same interpretation as in the formula for AIC. Given a pool of models, the model with the minimum AIC and BIC is usually considered optimal in terms of fit; however, these measures penalize the criteria for number of predictors. Thus, zero-inflated models can have slightly higher AIC and BIC statistics when they are providing the optimal fit. The information criteria are provided as added information to the model selection paradigm, and should be used to compare models within a GLM

series and models within a zero-inflation series, but not across series as twice as many parameters may need to be fit to zero-inflated models compared with GLMs.

Model Selection Paradigm

The tests described above were used to construct a model selection paradigm based on the relationship between the four competing models through their first 2 central moments (Figure 1). The Poisson and NB2 models are nested as described at the beginning of the Methods section of this paper. One can take advantage of the fact that ZIP and ZINB models are analogously nested in the limit as $\alpha \rightarrow 0$. Likelihood ratio tests can be used for model selection in the case of nested models. For non-nested models, as $\omega \rightarrow 0$, the ZIP model converges to the Poisson, and the ZINB model converges to the NB2, thus the Vuong test can be used.

	Poisson $E(Y) = \lambda$ $VAR(Y) = \lambda$	**NB2** $E(Y) = \lambda$ $VAR(Y) = \lambda+\alpha\lambda^2$
MODELS NOT NESTED ↕	**ZIP** $E(Y) = (1-\omega)\lambda$ $VAR(Y) = (1-\omega)(\lambda+\omega\lambda^2)$	**ZINB** $E(Y) = (1-\omega)\lambda$ $VAR(Y) = (1-\omega)\{\lambda+\alpha\lambda^2+\omega\lambda^2\}$

⟵ NESTED MODELS ⟶

Figure 1. Relationship of Models based on Moments

A graphical depiction of the model selection paradigm for PPT data is provided in Figure 2. Beginning with the Poisson GLM, the score test (18) for testing if the

heterogeneity parameter (α) is equal to zero is constructed to determine the presence of overdispersion. The score test for zero-inflation in the Poisson model (19) may also be constructed. Score tests have the advantage of not having to fit the alternative model to actually carry out the test. The NB2 model can also be tested using a likelihood ratio test (LRT), testing H0: $\alpha = 0$, as the variance function of the Poisson is nested within the variance function of the NB2 model. The Vuong test (20) for non-nested models can be used to test which model is closer to the true model among the Poisson and ZIP models; however, the test is limited if the null hypothesis is not rejected. If a ZIP model fits the data better than a Poisson, then the ZINB model can be tested using an LRT (variance functions nested) to determine if a better fit achieved with the ZIP or ZINB. The Vuong test can be used to test the NB2 against the ZINB analogous to the Poisson and ZIP models. This testing paradigm was used for selecting the best model using a subset of the UDCA clinical trial data as an example.

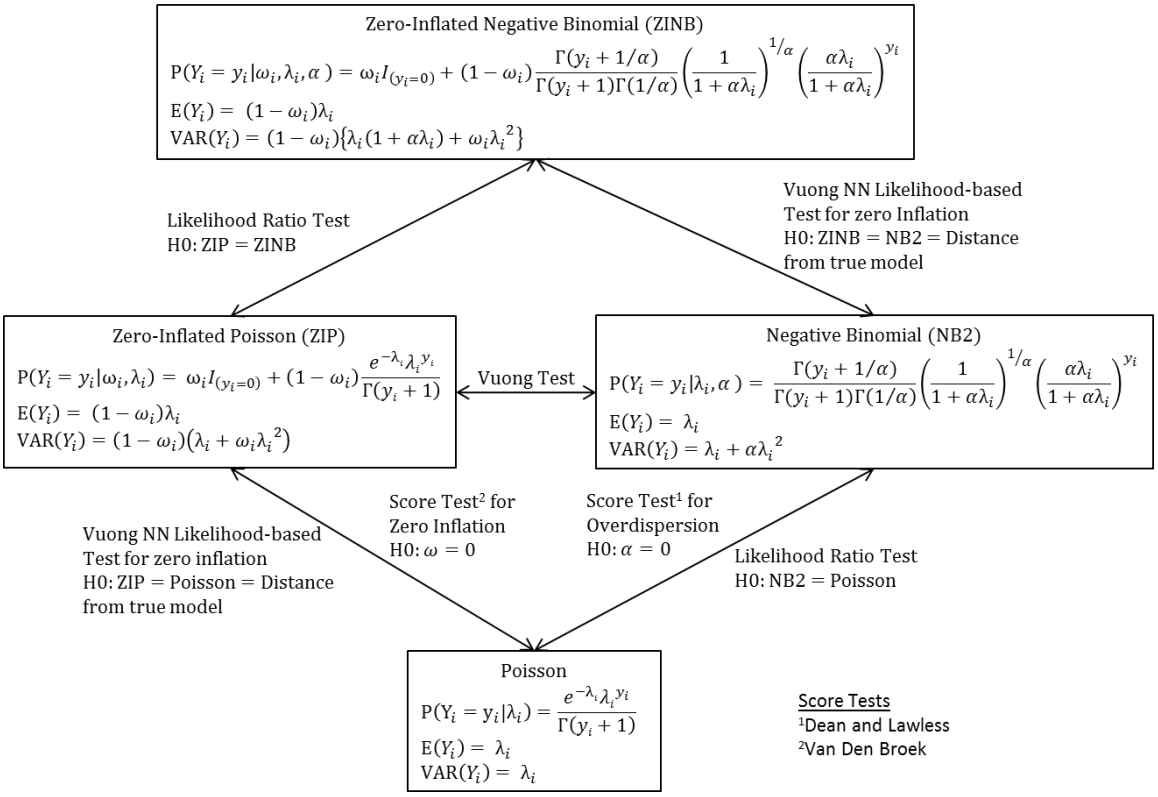

Figure 2. Model Selection Paradigm

The Ursodeoxycholic Acid Clinical Trial

An example of employing the model selection paradigm is demonstrated. The data to be analyzed in this report were collected in a phase III, randomized, double-blind, placebo-controlled clinical trial designed to test the efficacy of ursodeoxycholic acid (UDCA) for the prevention of colorectal adenomas. This study has been analyzed extensively in the literature (3, 5, 23-25). Patients who were eligible for the study had at least 1 colorectal adenoma ≥ 3 mm in diameter removed during a colonoscopy within the 6-month period before registration. Patients had no clinical evidence of disease and no invasive cancer in the previous 5 years. All other colorectal neoplasms (except for diminutive polyps) were completely removed. Patients were randomized to receive

UDCA (8-10 mg/kg/day) or placebo. The duration of the treatment period was to be approximately 3 years or until completion of the follow-up colonoscopy within 6 months of the 3-year anniversary.

The primary outcome of the study was the sum of the recurrent colorectal adenomas recorded during the treatment period (baseline colonoscopy to the end of the follow-up). Some patients had more than 1 colonoscopy during the treatment period; however, this information was not used for this demonstration. The rates of recurrence (number of incident adenomas / year) were compared between UDCA and placebo in a Poisson model that used robust variance estimators for the standard errors. A non-statistically significant 12% reduction in the recurrence rate in patients who received UDCA compared with placebo was observed in the study (3). The UDCA trial dataset contains a total of 1192 observations with exposure times ranging from 2.3 months to approximately 7 years; however 871 observations (73%) of the observations have exposure times ≤ 3.5 years. The methods described in this paper will be applied to the subset of subjects who had an exposure time ≤ 3.5 years.

Tests Employed in Model Selection

The count of recurrent adenomas was modeled as Poisson with robust (sandwich) estimators of the standard errors, NB2, ZIP, and ZINB with covariates treatment group (placebo as reference), sex (female as reference), family history of colorectal cancer (“no” as reference), age (years) and Ln (BMI) for the subset of 871 subjects in the UDCA data set who had an exposure of 0 to 3.5 years on study. Each model was constructed with and without an offset term for the follow-up time to determine how influential an offset would be in terms of inference and model selection.

The model selection paradigm described in Figure 2 was applied to select the model that best fit the data. Covariates that were statistically significant in at least 1 component of the zero-inflated models were to remain in the final model. Poisson and negative binomial models were constructed as GLMs using iteratively reweighted least squares (IRLS) algorithms to compute maximum likelihood estimates of the parameters. All statistical calculations were carried out using R version 2.11.1 statistical software, and all zero-inflated models were estimated using the Newton-Raphson algorithm as implemented in the R statistical software PSCL package (26). The R code used to fit the models with the offset is provided in Appendix 2.

Exploratory Data Analysis

Descriptive statistics were summarized for each variable in the dataset. The total number of adenomas recorded for each subject in the trial was tabulated as a list, summarized with the mean, median, standard deviation, minimum and maximum values, and graphed as a histogram. BMI was explored as untransformed and as the natural log transformation of the variable, and summarized with the mean, standard deviation, minimum value, and maximum value. Age was summarized untransformed analogously to BMI. Histograms were constructed for each variable. Binary variables (treatment group, previous history of polyps, family history of colon cancer, and sex) were summarized with the number and percentage of each level of the variable. The follow-up time in days was transformed to units of years in order to be consistent with the methodology used in the original paper (3). Histograms were constructed for the follow-up time and natural log of the follow-up time.

Bivariate graphs were constructed to determine a preliminary relationship between the outcome (count of adenomas) and each independent variable. Because count data are discrete random variables, scatterplots produce gaps that make visualizing linearity difficult. Thus, age and BMI were categorized and plotted against the mean of the count data for each category with locally weighted scatterplot smoothing (LOWESS) through the points (27). For exploratory purposes, age was categorized in 5-year intervals and BMI was categorized as follows: <22, 22 to 23.5, 23.5 to 25, 25 to 27.5, 27.5 to 30, 30 to 35, and ≥ 35. Box plots were constructed for each level of the binary predictors against the number of recurrent adenomas.

DATA ANALYSIS

Results

The distribution of recurrent adenoma counts for the subset of subjects in the UDCA clinical trial that had 0 to 3.5 years of follow-up by treatment group is provided in Table 2 and Figure 3. A total of 430 subjects were randomized to the placebo group and 441 subjects were randomized to receive UDCA. Counts of recurrent adenomas ranged from 0 to 6 in both groups. A slightly higher proportion of zero counts were observed in the UDCA group (65.5%) compared with the placebo group (61.4%).

Table 2. Distribution of Recurrent Adenoma Count by Treatment Group

Adenoma Count	Placebo N (%)	UDCA N (%)
0	264 (61.4)	289 (65.5)
1	106 (24.7)	99 (22.4)
2	33 (7.6)	32 (7.3)
3	16 (3.7)	11 (2.5)
4	7 (1.6)	6 (1.4)
5	2 (0.5)	3 (0.7)
6	2 (0.5)	1 (0.2)
Total	430	441

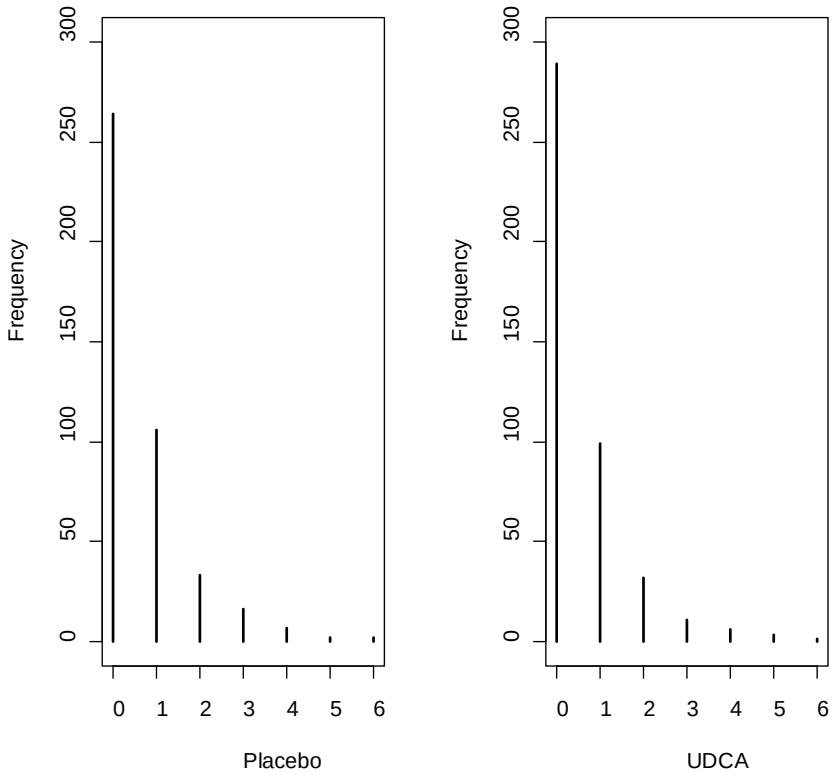

Figure 3. Adenoma Counts by Treatment Group

Summary statistics by treatment group for the variables that were included in the dataset are provided in Table 3, and the distributions of continuous variables by treatment group are provided in Figure 4. Overall, the demographics in both treatment groups were similar. Most subjects in both groups were overweight males over the age of 66 years. The mean follow-up time in both groups was 2.6 years, which is typical for a PPT. Boxplots of recurrent adenoma count by each binary covariate are provided in Figure 5. A total of 128 subjects (29.8%) in the placebo group and 96 subjects (21.8%) in the UDCA group had a family history of colorectal cancer. One hundred, seventy-two subjects (40.0%) in the placebo group and 197 subjects (44.7%) in the UDCA group had a history of polyps. There were 29 missing values in the placebo group and 21 missing

values in the UDCA group for the history-of-polyps variable, thus history of polyps was not included in any models for this analysis.

Sensitivity analyses at the 0.05 level of significance were conducted to compare summary statistics for covariates in the subset with those in the full dataset. One-sample t-tests were used to compare means of continuous covariates assuming the respectful mean in the full dataset is the true value, and one-sample proportion tests were used to compare proportions for categorical covariates assuming the respectful proportion in the full dataset is the true value. All tests were 2-sided. There was a higher proportion of patients in the placebo group of the subset who had a family history of CRC compared with the proportion of the placebo group in the full dataset who reported a family history of CRC ($p=0.0089$). The mean total adenoma count was lower in each treatment group in the subset compared with the mean total adenoma count for each treatment group in the complete dataset ($p<0.0001$). No other covariates were significantly different from their respectful covariates in the complete dataset. Summary statistics for the complete dataset are provided in Appendix 1 (Tables A8 and A9) of this paper.

Table 3. Summary Statistics

	Placebo (N = 430)							UDCA (N = 441)						
	Mean	SD	Med	Min	Max	Missing	P*	Mean	SD	Med	Min	Max	Missing	P*
Outcome														
Recurrent Adenomas	0.6	1.0	0	0	6	0	<0.0001	0.5	1.0	0	0	6	0	<0.0001
Continuous Variables														
Age (Years)	66.5	8.2	68.0	42.0	81.0	0	0.9999	66.2	8.7	68.0	40.0	80.0	0	0.0541
BMI (kg/m^2)	27.7	4.8	26.9	13.1	44.9	0	0.3881	27.2	4.3	26.7	15.0	46.1	0	0.6255
Follow-time (Years)	2.6	0.6	2.8	0.2	3.5	0		2.6	0.6	2.8	0.5	3.5	0	
Binary Variables	Yes		Percent					Yes		Percent				
Sex = Male	281		65.3			0	0.8426	314		71.2			0	0.3185
Fam Hx of CRC	128		29.8			0	0.0089	96		21.8			0	0.0760
History of Polyps	172		40.0			29	0.2648	197		44.7			21	0.4035

P*=p-value compared with complete data: one-sample t-test for continuous covariates and one-sample test of proportion for binary covariates

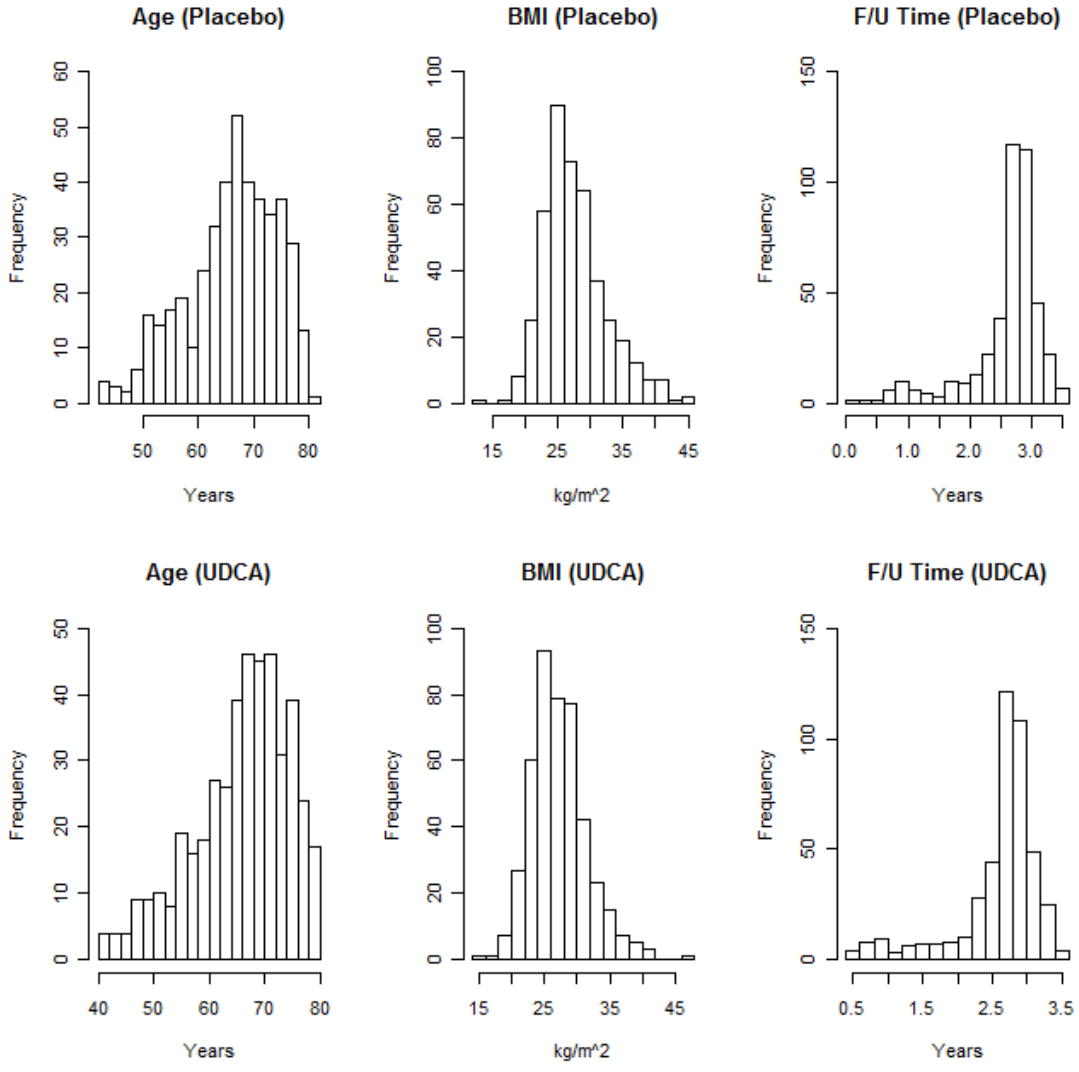

Figure 4 Distributions of Continuous Variables

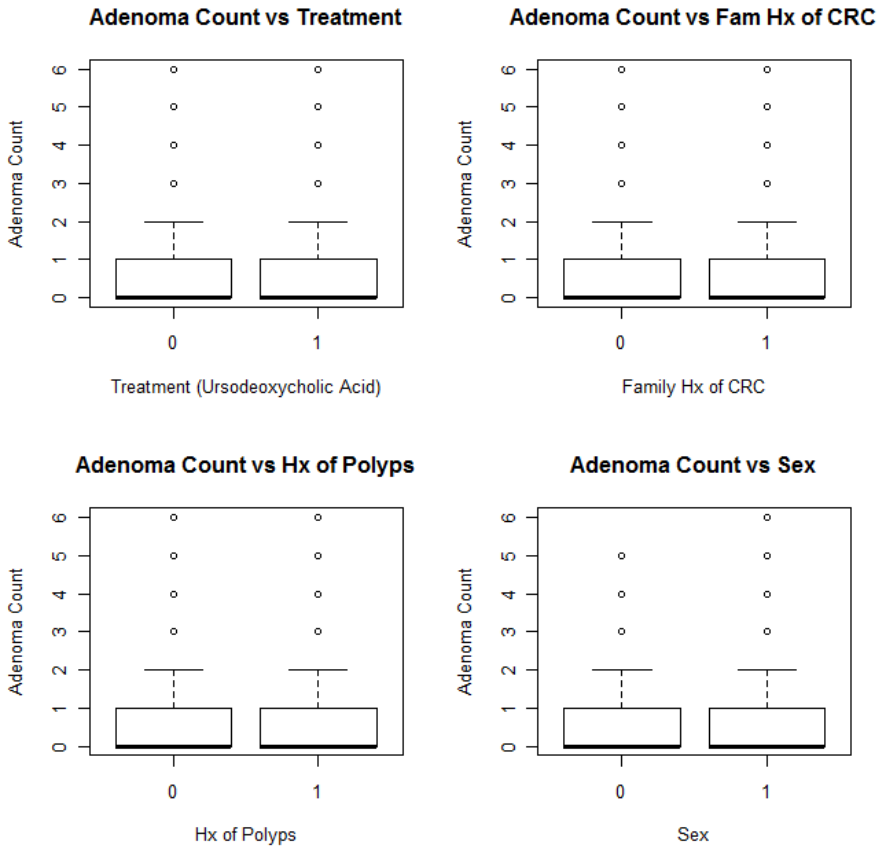

Figure 5 Recurrent Adenoma Count by Levels of Binary Covariates

Scatterplots were constructed to assess a possible linear relationship between recurrent adenoma count and continuous covariates age and BMI (Figure 6). However, the scatterplots were not useful for visualizing the linearity of the adenoma counts with the covariates because the count data has a relatively high variability. Hence age and BMI were categorized and plotted against the mean count of recurrent adenomas in each category. The LOWESS lines indicate an approximate linear relationship between recurrent adenomas with age and BMI. The excess zeros in the outcome of each category presented a scaling problem since most of the mass of the count distribution

was in the 0 responses. However, linearity may be detected when the data are plotted this way.

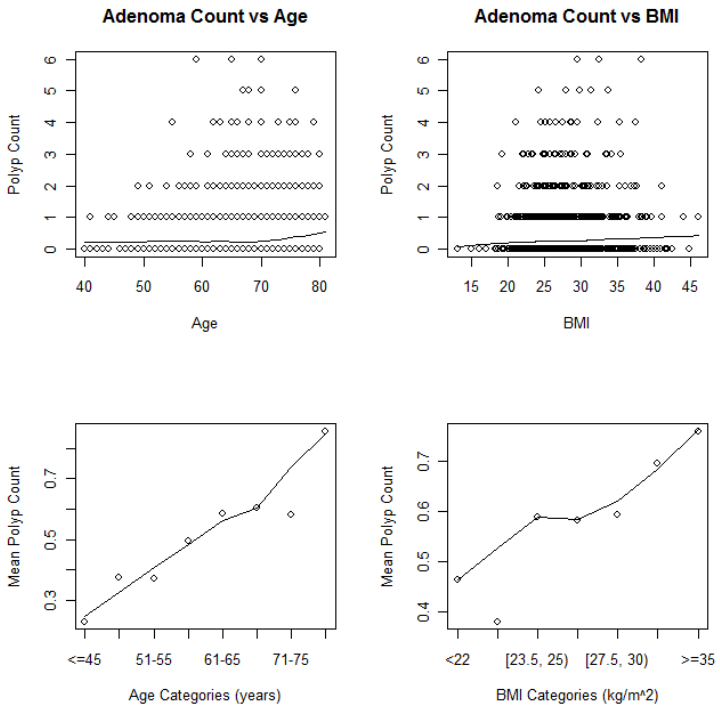

Figure 6 Assessment of Linearity for Recurrent Adenoma Count with Age and BMI

The mean recurrent adenoma count in the example dataset, not taking into consideration the variable follow-up times, was 0.5867. Figure 7 displays the observed and predicted probabilities assuming a Poisson distribution with $\lambda = 0.5867$. It is clear that a Poisson model would likely underestimate zero counts, overestimate counts of 1 and 2, and slightly underestimate the remaining counts. Thus it is likely that a Poisson model will not provide the optimal fit for the data.

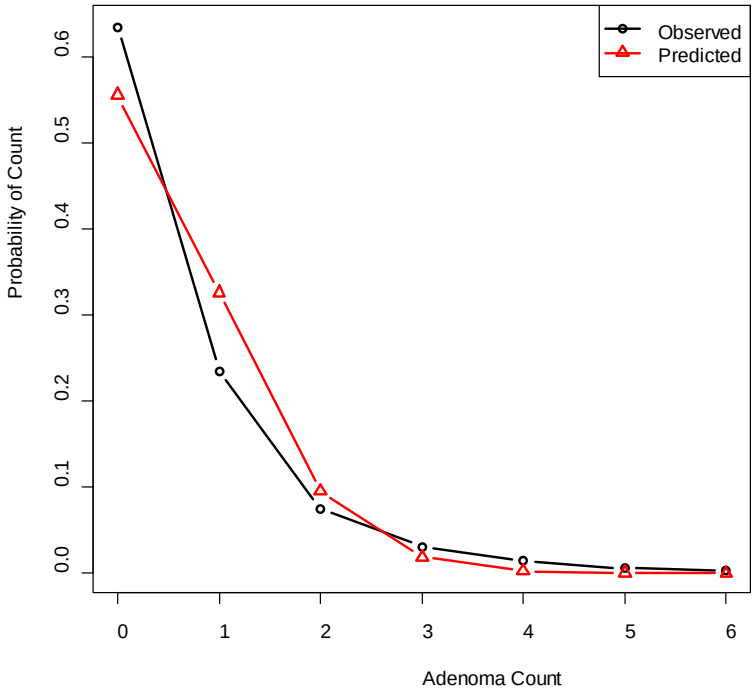

Figure 7 Comparison between Observed Probabilities and Univariate Poisson ($\hat{\lambda}$ = 0.5867) Probabilities

Model Fitting and Selection

The results of applying the model selection paradigm for the series of models fitted to the subset of UDCA data are provided in Table 4, and detailed parameter estimates for each model are provided in Appendix 1. Two sets of 4 models (1 set with no offset and 1 set with an offset for the natural log of the exposure time) were fitted as in Figure 2 beginning with the Poisson GLM. In both sets of models, the NB2 estimates the dispersion statistic closest to 1; however, the dispersion statistics in ZINB models are acceptably close to those estimates. The score tests for overdispersion in the Poisson indicate that the models with and without the offset are overdispersed ($p < 0.001$). The score test for zero-inflation in the Poisson indicates that at least some of the overdispersion can be attributed to zero-inflation ($p < 0.001$). The null hypothesis for the likelihood ratio test that $\alpha = 0$ in $V(Y) = \mu + \alpha\mu^2$ for the NB2 is rejected, and thus we would conclude that the NB2 provides a superior fit to the data compared with the Poisson ($p < 0.001$). The reduction in AIC and BIC for the NB2 also provide evidence for selecting the NB2 model rather than the Poisson.

The zero-inflated models provided superior fit to the data compared with the Poisson. The Vuong test for testing the fit of the ZIP model compared with the Poisson model when no offset was included estimates the test statistic at 3.56, which indicates a significantly better fit to the data for the ZIP model. The test statistic for the Vuong test between ZIP and NB2 was -1.83, which indicates that neither model is closer to the true model. The ZINB model, however, provided the best fit of the four possible models to the data, and this result was consistent when the model contained an offset. The null hypothesis for the LRT (ZIP compared with ZINB) was rejected indicating that the ZINB

model is superior to the ZIP model ($p < 0.001$), and the Vuong test favored the ZINB over the NB2 model. Similar results were observed when an offset was included.

The AIC and BIC statistics were consistently lower for the NB2 models compared with the Poisson models, and were lower for the ZINB models compared with the ZIP models. Because information criteria statistics penalize the values for more covariates, it may not be reasonable to compare AIC and BIC for a zero-inflated model with a GLM, since twice as many covariates are needed for the zero-inflated model; however, comparisons within zero-inflated models are valid just as are comparisons within GLMs. Thus, within the zero-inflated models, the AIC and BIC for the ZINB model provide evidence than the fit is better than for the ZIP model, and lowest for the ZINB model that includes an offset for the follow-up time. Overall, the results of applying the model selection paradigm to the subset of UDCA data truncated at a follow-up time of 3.5 years suggest that the ZINB model that includes an offset for the follow-time is the model best suited for the data compared with the other 3 choices.

Table 4 Results of Model Selection

w/o Offset	LL	RDF	Pearson Dispersion	D&L Score Test Overdispersion	VDB Score Test Zero-inflation	LRT chi-square p-value	Vuong 1	Vuong 2	AIC	BIC
Poisson	-933.41	865	1.54	Z = 0.377 H0: Z = 0 p < 0.001	64.20 p < 0.001				1878.81	1907.44
NB2	-892.12	865	0.99			Poisson vs NB2 82.57 p<0.001			1798.25	1824.87
ZIP	-900.65	859	1.13				ZIP vs Poisson T = 3.56 p < 0.001 ZIP > Poisson	ZIP vs NB2 T = -1.83 p = 0.034 NB2 = ZIP	1825.30	1882.54
ZINB	-883.07	858	1.04			ZIP vs ZINB 35.15 p<0.001 ZINB > ZIP		ZINB vs NB2 T = 2.44 p = 0.007 ZINB > NB2	1790.15	1847.38
w/Offset										
Poisson	-928.80	865	1.56	Z = 0.348 H0: Z = 0 p < 0.001	61.53 p < 0.001				1869.59	1898.21
NB2	-890.37	865	1.04			Poisson vs NB2 76.86 p<0.001			1794.73	1821.35
ZIP	-895.81	859	1.12				ZIP vs Poisson T = 3.53 p < 0.001 ZIP > Poisson	ZIP vs NB2 T = -1.36 p = 0.088 NB2 = ZIP	1815.62	1872.86
ZINB	-882.38	858	1.09			ZIP vs ZINB 26.85 p<0.001		ZINB vs NB2 T = 2.16 p = 0.015 ZINB > NB2	1788.77	1846.00

Graphs of the observed and predicted proportions of recurrent adenoma counts for all models fitted with the offset for the Ln of the follow-up time are provided in Figure 8. The fit of the NB2, ZIP, and ZINB models is improved compared with the Poisson. The NB2 model slightly over estimates the proportion of subjects who had 2 recurrent adenomas, but provides an acceptable fit to the data overall. The ZIP model predicts zero counts well, but does not account for any other sources of overdispersion. Consequently, the ZIP model over predicts counts of 2 and 3. The ZINB estimates the counts similar to the NB2, but the model is capable of accounting for zero-inflation as well as other sources of overdispersion.

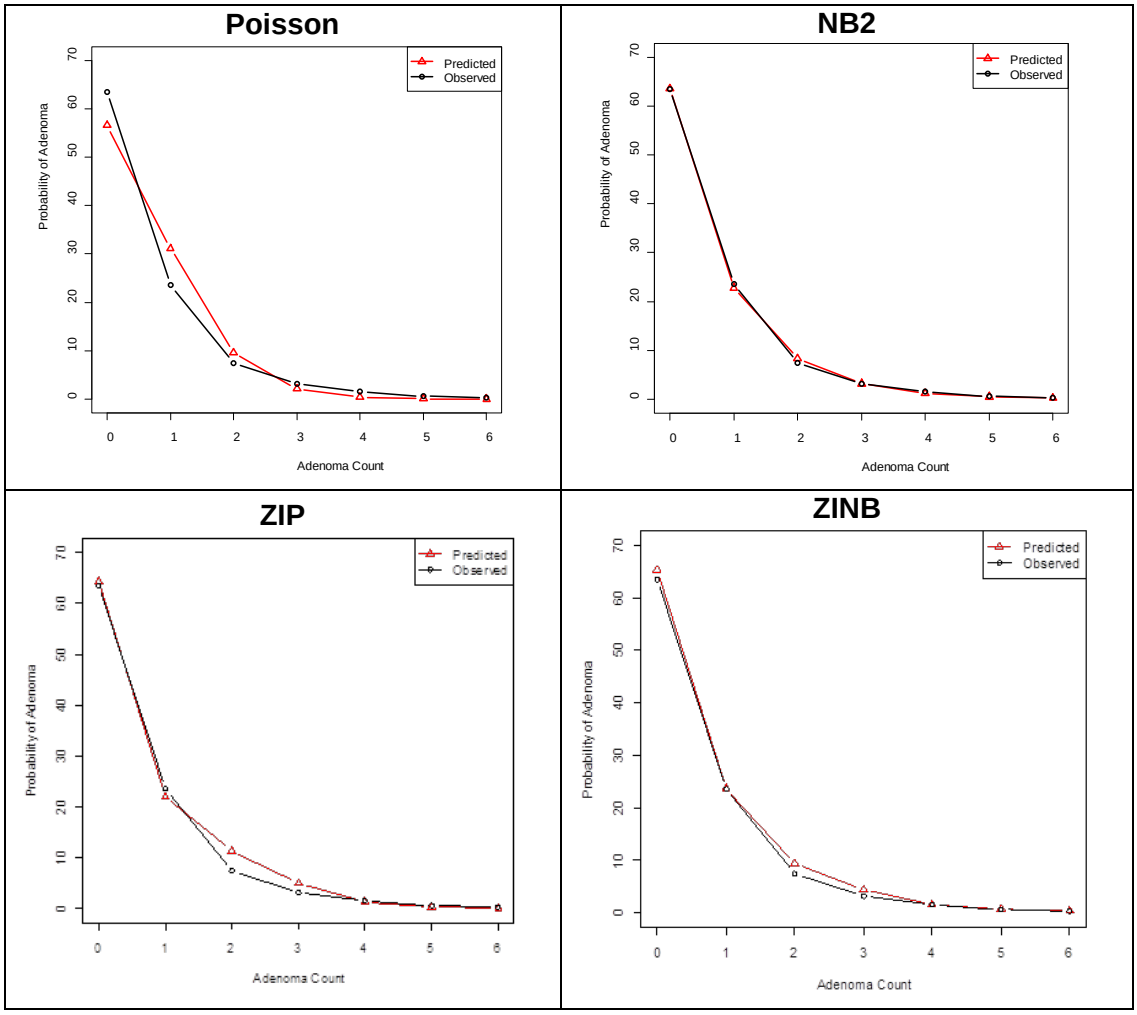

Figure 8. Observed vs Predicted Proportions for all Models (with Offset)

Interpretation of the ZINB Model

The ZINB model was demonstrated to be the best of the 4 distributions in the model selection paradigm to model the UDCA data. The model suggests that the rate of adenoma recurrence in male patients was exp(0.376) = 1.46 times the rate of adenoma recurrence in female patients holding all other predictors constant (p=0.0046). The rate of recurrence for patients who had a family history of CRC was exp(0.404) = 1.50 times

the rate of recurrence in patients who had no family history of CRC holding all other predictors constant (p=0.0166).

Ln(BMI) is negatively associated with being in the “certain zero” group. The odds ratio for any 1-unit increase in Ln(BMI) is exp(-9.67) = 6.31X10-5 , which is a large (approximately 99,000%) decrease in the odds of being in the certain zero group for every 1-unit increase in Ln(BMI). Another way to interpret this estimate is as a change in the average log odds of being in the certain zero group for a given percentage, say 20%, increase in BMI. For example, the average log odds of being in the certain zero group will increase Ln(1.20)(-9.67) = -1.76; or decrease 1.76, for every 20% increase in the BMI within the range of BMIs in the sample and holding all other predictors constant (p=0.0286).

The model fitting algorithm in the PSCL package estimates the heterogeneity parameter ($\theta = 1/\alpha$) as the natural logarithm (Ln(θ)), and the null hypothesis for the resulting Wald test is a test of $\theta = 1$. We fail to reject the null hypothesis at the 0.05 level of significance, and conclude that the heterogeneity parameter is not significantly different from 1 (p = 0.1234). When the heterogeneity parameter in the NB2 distribution is equal to 1, the distribution reduces to the geometric distribution (28). Thus, this particular model is statistically equivalent to a zero-inflated geometric model.

Table 5 Zero-Inflated Negative Binomial Model (with Offset)

LL	Res DF	Pearson Dispersion	AIC	BIC
-882.38	858	1.09	1788.77	1846.00

Count Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	-5.939984	1.381811	-4.299	<0.0001	-8.648284	-3.231685
Group = UDCA	-0.182996	0.114833	-1.594	0.1110	-0.408064	0.042072
Family hx of CRC	0.404039	0.168684	2.395	0.0166	0.073425	0.734653
Sex = Male	0.375742	0.132426	2.837	0.0046	0.116191	0.635292
Age	0.035844	0.007273	4.928	<0.001	0.021588	0.050099
Ln(BMI)	0.549296	0.376290	1.460	0.1444	-0.188219	1.286812
Ln(θ)	0.320209	0.207809	1.541	0.1234	-0.087097	0.727515

Binary Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	10.58521	27.22229	0.389	0.6974	-42.769506	63.939920
Group = UDCA	-1.02199	1.33980	-0.763	0.4456	-3.647959	1.603976
Family hx of CRC	10.39135	24.24322	0.429	0.6682	-37.124499	57.907192
Male	0.06359	0.92247	0.069	0.9450	-1.744424	1.871603
Age	0.12760	0.09886	1.291	0.1968	-0.066168	0.321372
Ln(BMI)	-9.67297	4.41824	-2.189	0.0286	-18.332567	-1.013380

DISCUSSION

Rationale for Approach

The model selection paradigm presented in this paper was designed to unify the Poisson, NB2, ZIP, and ZINB models into a collection of models and hypothesis tests that can be used to achieve optimal fit of data collected in PPTs. It is likely, but not certain, that PPT data will be overdispersed and possibly contain more zeros than would be predicted by the Poisson GLM. Excess zeros are a source of overdispersion in Poisson and negative binomial GLMs, but it is the nature and extent of overdispersion that becomes important when selecting a model for PPT data. The Poisson distribution can be used reliably in the presence of slight overdispersion; however, the fit of a Poisson GLM degrades rapidly as the dispersion statistic deviates from unity. In the absence of a known mechanism of overdispersion in the study design, apparent overdispersion in the Poisson GLM may sometimes be resolved by including additional statistically significant covariates, transforming the outcome variable, constructing interaction terms, or excluding known outliers (6). The overdispersion in PPT data is assumed to be real overdispersion, thus methods for handling apparent overdispersion were not discussed in this paper.

The NB2 GLM is traditionally the alternative model when a Poisson GLM is overdispersed; however, zero-inflated models may provide superior fit to overdispersed data that contain excess zeros. The NB2 GLM could be sufficient to model the rate of recurrent adenomas in PPT data that contain excess zeros, but using these models fails to account for the probability of zero counts separately. The ZIP model is useful for modeling data that are overdispersed if the overdispersion is only the result of excess zeros. However, when the data contain unobserved heterogeneity resulting from

sources of overdispersion other than excess zeros, then the ZIP model will likely be inferior to the NB2 or ZINB models. The ZINB model can accommodate overdispersion resulting from excess zeros and other sources of overdispersion. Some modelers may be hesitant to use a zero-inflated model because interpreting the coefficients can be challenging, particularly in the case when the estimates for a parameter in both the count and binary components appears to conflict each other. It is important to take the parameterization of the model into account when interpreting such models (e.g. the log odds of being in the certain zero group for some level of a covariate in the binary component). If the estimates in each component conflict each other for the same covariate, the interpretation is challenging.

Limitations of the Approach

The likelihood-based tests presented in this paper demand several assumptions for the results to be valid. The Vuong test assumes that the models are non-nested and that the outcome vectors, and the cardinality of the vectors, are equal for both models being tested. This means that missing values for covariates present a problem for covariate selection using this paradigm. For example missing values for a covariate will exclude the corresponding values of the outcome from the model that includes that covariate, which would violate the assumption that that the outcomes are equal in both models. In fact, it is recommended that no covariate selection be determined by the approach presented in this paper. The model selection paradigm is helpful only in determining the best probability mass function to model the count structure of the outcome. All models were constructed with all covariates in the dataset, except history of polyps because of missing data, and assessed using the model selection paradigm.

The AIC and BIC statistics can be used to select between Poisson and negative binomial GLMs or between ZIP and ZINB models without regard for the penalty of adding predictors. When comparing zero-inflated models with GLMs however, some consideration should be given regarding the penalty for additional covariates, particularly with the BIC statistic because zero-inflated models will have twice as many predictors as their analogous GLM counterparts if the full models are constructed in both components of the zero-inflated models. For example, the ZINB models constructed for the UDCA data have slightly lower log-likelihoods but higher BIC values compared with the NB2 models, which is consistent with a model that has twice as many parameters to estimate.

Other methods for modeling excess zeros in count data including zero-inflated binomial and hurdle models (29) were excluded from the approach. The zero-inflated binomial would be a poor choice for modeling counts because some of the information contained in the sample would be lost by collapsing the counts >0 into a single category. The hurdle model was excluded because the hurdle model separates zero counts from the remaining counts by modeling the rate data as a truncated-at-1 Poisson (or negative binomial), which may be appropriate if a proportion of the population had no probability of recurrence at the beginning of the study. However, all subjects in PPTs have had at least 1 adenoma before beginning the study and have a chance of recurrence that should be modeled with no truncation in the count component.

Additionally, the modeling paradigm presented in this paper was designed to accommodate overdispersion as expected in PPTs. However, count data can be underdispersed as well as overdispersed (12), and an entirely different modeling approach is employed in the event of underdispersion. Thus, the work in this paper is

intended only for count data suspected of being overdispersed, and preferably when an influential source of overdispersion resulted from excess zero counts in the data.

Misclassification

The potential for misclassification of recurrent adenomas in polyp prevention trials provides a compelling reason to model the outcome with a distribution that can account for the mechanism that generates excess zero counts relative to the Poisson distribution. Because of the limits of detection of colonoscopy, misclassification of polyps (false negatives) is possible (4). Misclassification is most likely to occur when there are diminutive polyps (< 1 mm diameter) that are not detected, thereby leading to biased results due to measurement error. Miss rates of 27% for adenomas ≤ 5 mm in diameter in back to back colonoscopies have been reported (30). However, misclassification was not explicitly modeled because the emphasis for model selection was based on goodness-of-fit. Thus, while misclassification was possible with this study design, the observed proportion of misclassified adenomas may have been less than expected for a ZIP model to provide optimal fit. The conclusion that the treatment effect was not statistically significant remained unchanged with all of the models fitted in this analysis.

Limitations of the Example Data

The analysis of the data provided for the example was used to show how the model selection paradigm could be used to select the best fitting model that accounts for the excess zeros in a PPT. The selection of the ZINB model and subsequent model fitting was not intended to make a statement regarding the study as no covariate

selection methods were employed. Of note, an interaction between treatment and sex was found in a logistic regression model of the full dataset (31), and this interaction would likely need to be further analyzed as part of a comprehensive covariate selection method before a valid inference could be made for the example data. The interpretation for the final model is limited to the subset of data used for the example conditional on the covariates used in the example and should not be generalized to either the complete UDCA trial dataset or the population from which the study sample was drawn.

The purpose of this project was to describe a model selection paradigm that could be applied when fitting statistical models to recurrent adenoma data collected in a PPT study. Overdispersion and specifically zero-inflation are expected to exist in recurrent adenoma data. There are several ways to model overdispersion, but because excess zero-counts are expected in PPTs, zero-inflated models for count data (ZIP and ZINB) should be considered as part of an appropriate family of modes for such data. The model selection paradigm described here relies on well-known statistical tests for model selection, and these methods that have been documented extensively in the literature. Additionally, this selection methodology can be expanded to include recurrent adenoma data with random effects and thus provides a valuable tool for investigators interested in constructing optimal models for recurrent adenoma count data collected in PPTs conditional on covariates collected in the study.

APPENDIX 1

Models (No Offset)

Table A1. Poisson GLM with Robust Sandwich estimators for SE – No Offset

LL	Res DF	Pearson	Pearson/Res DF	AIC	BIC
-933.4086	865	1335.047	1.543407	1878.817	1907.435

	Estimate	SE	z	P	CILower	CIUpper
Intercept	-6.180305	1.289065	-4.794	<0.0001	-8.706872	-3.653738
Group = Urso	-0.129037	0.111100	-1.161	0.2456	-0.346793	0.088719
Fam hx of CRC	0.109813	0.125026	0.878	0.3799	-0.135238	0.354864
Male	0.390356	0.124795	3.128	0.0018	0.145758	0.634954
Age	0.029886	0.006488	4.606	<0.0001	0.017169	0.042602
Ln(BMI)	1.023063	0.339649	3.012	0.0026	0.357351	1.688775

Table A2. Negative Binomial GLM – No Offset

LL	Res DF	Pearson	Pearson/Res DF	AIC	BIC
-892.125	865	858.224	0.9921	1798.25	1824.87

	Estimate	SE	z	P	CILower	CIUpper
Intercept	-6.075382	1.348726	-4.505	<0.0001	-8.748627	-3.437787
Group = Urso	-0.143477	0.112455	-1.276	0.2020	-0.364665	0.077284
Fam hx of CRC	0.111097	0.127296	0.873	0.3828	-0.143042	0.363229
Male	0.396966	0.128355	3.093	0.0020	0.145968	0.652126
Age	0.030266	0.007168	4.223	<0.0001	0.016352	0.044457
Ln(BMI)	0.984212	0.355823	2.766	0.0057	0.279302	1.693245
1/α	1.064	0.181	5.878	<0.0001	0.70924	1.41876

Table A3. Zero-Inflated Poisson – No Offset

LL	Res DF	Pearson	Pearson /Res DF	AIC	BIC
-900.650	859	974.169	1.13	1825.301	1882.536

Count Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	-3.724032	1.703192	-2.187	0.0288	-7.062226	-0.385838
Group = Urso	-0.069120	0.132358	-0.522	0.6015	-0.328538	0.190297
Fam hx of CRC	0.340040	0.190705	1.783	0.0746	-0.033736	0.713816
Male	0.382993	0.170237	2.250	0.0244	0.049335	0.716651
Age	0.031395	0.008598	3.652	<0.001	0.014544	0.048246
Ln(BMI)	0.370398	0.505106	0.733	0.4634	-0.619591	1.360387

Binary Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	4.075026	4.249390	0.959	0.338	-4.253626	12.403678
Group = Urso	0.188607	0.308805	0.611	0.541	-0.416639	0.793853
Fam hx of CRC	0.542518	0.439672	1.234	0.217	-0.319223	1.404258
Male	-0.016464	0.396126	-0.042	0.967	-0.792857	0.759929
Age	0.004805	0.020578	0.233	0.815	-0.0355283	0.045138
Ln(BMI)	-1.552599	1.346118	-1.153	0.249	-16.819613	-2.279001

Table A4. Zero-Inflated Negative Binomial – No Offset

LL	Res DF	Pearson	Pearson /Res DF	AIC	BIC
-883.07	858	894.29	1.04	1790.15	1847.38

Count Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	-4.837887	1.390609	-3.479	0.0005	-7.563430	-2.112344
Group = Urso	-0.186285	0.114773	-1.623	0.1046	-0.411236	0.038666
Fam hx of CRC	0.438223	0.168757	2.597	0.0094	-0.107465	0.768981
Male	0.406402	0.132131	3.076	0.0021	0.147430	0.665374
Age	0.035965	0.007276	4.943	<0.001	0.021703	0.050226
Ln(BMI)	0.495739	0.377572	1.313	0.1892	-0.244289	1.235766
Ln(theta)	0.272922	0.200304	1.363	0.1730	-0.119674	0.665518

Binary Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	12.24267	36.87382	0.332	0.7399	-60.028694	84.514043
Group = Urso	-0.89191	1.17472	-0.759	0.4477	-3.194326	1.410497
Fam hx of CRC	10.88155	34.64869	0.314	0.7535	-57.028637	78.791741
Male	0.06107	0.89193	0.068	0.9454	-1.687073	1.809209
Age	0.11656	0.09030	1.291	0.1968	-0.060429	0.293550
Ln(BMI)	-9.78188	4.14794	-2.358	0.0184	-17.911698	-1.652055

Models with Offset

Table A5. Poisson GLM with Robust Sandwich estimators for SE

LL	Res DF	Pearson	Pearson/Res DF	AIC	BIC
-928.80	865	1350.55	1.56	1869.59	1898.21

	Estimate	SE	z	P	CILower	CIUpper
Intercept	-7.138190	1.280337	-5.575	<0.0001	-9.647651	-4.628729
Group = Urso	-0.125981	0.110433	-1.140	0.2543	-0.342430	0.090468
Fam hx of CRC	0.101717	0.124250	0.818	0.4134	-0.141813	0.345247
Male	0.361928	0.123892	2.921	0.0035	0.119100	0.604756
Age	0.030421	0.006505	4.677	<0.0001	0.017671	0.043171
Ln(BMI)	1.017585	0.338264	3.008	0.0026	0.354588	1.680582

Table A6. Negative Binomial GLM

LL	Res DF	Pearson	Pearson/Res DF	AIC	BIC
-890.37	865	896.31	1.04	1794.73	1821.25

	Estimate	SE	z	P	CILower	CIUpper
Intercept	-0.705283	1.342681	-5.253	<0.0001	-9.719199	-4.421690
Group = Urso	-0.140217	0.112013	-1.252	0.2107	-0.361018	0.080151
Fam hx of CRC	0.095466	0.126940	0.752	0.4520	-0.158072	0.346861
Male	0.366761	0.128159	2.862	0.0042	0.115615	0.622086
Age	0.030123	0.007142	4.218	<0.0001	0.016215	0.044300
Ln(BMI)	1.000730	0.354756	2.821	0.0048	0.296593	1.708921
1/α	1.116	0.195	5.723	<0.0001	0.7338	1.4982

Table A7. Zero-Inflated Poisson

LL	Res DF	Pearson	Pearson /Res DF	AIC	BIC
-895.81	859	963.15	1.12	1815.62	1872.86

Count Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	-6.00276	1.74281	-3.444	0.0006	-9.418613	-2.586912
Group = Urso	-0.05499	0.13242	-0.415	0.6779	-0.314519	0.204539
Fam hx of CRC	0.13003	0.20783	0.626	0.5315	-0.277302	0.537359
Male	0.36427	0.16945	2.150	0.0316	0.032145	0.696392
Age	0.02304	0.00877	2.627	0.0086	0.005848	0.040225
Ln(BMI)	0.95927	0.52898	1.813	0.0698	-0.077515	1.996055

Binary Component	Estimate	SE	z	P	CILower	CIUpper
Intercept	-0.22537	4.07223	-0.055	0.956	-8.206786	7.756052
Group = Urso	0.20992	0.30274	0.693	0.488	-0.383441	0.803277
Fam hx of CRC	0.08766	0.48300	0.181	0.856	-0.859014	1.034326
Male	-0.01255	0.39508	-0.032	0.975	-0.786902	0.761799
Age	-0.01466	0.01893	-0.775	0.439	-0.051759	0.022438
Ln(BMI)	-0.11114	1.25536	-0.089	0.929	-2.571602	2.349328

Table A8. Summary Statistics (Full Dataset)

	Placebo (N = 579)						UDCA (N = 613)					
	Mean	SD	Med	Min	Max	Miss	Mean	SD	Med	Min	Max	Miss
Outcome												
Recurrent Adenomas	0.8	1.4	0	0	15	0	0.7	1.1	0	0	8	0
Continuous Variables												
Age (Years)	66.5	8.3	68.0	41.0	81.0	0	67.0	8.6	67.0	40.0	80.0	0
BMI (kg/m^2)	27.5	4.7	26.7	13.1	44.9	0	27.3	4.5	26.7	15.0	46.1	0
Follow-time (Years)	3.2	1.2	2.9	0.2	6.7	0	3.2	1.1	2.9	0.5	6.9	0
Binary Variables	Yes		Percent				Yes		Percent			
Sex = Male	381		65.8			0	423		69.0			0
Fam Hx of CRC	141		24.3			0	156		25.4			0
History of Polyps	247		45.5			36	286		46.7			29

Table A9. Recurrent Adenoma Count by Treatment Group (Full Dataset)

Adenoma Count	Placebo N (%)	UDCA N (%)
0	325 (68.0)	362 (81.9)
1	144 (30.1)	140 (31.7)
2	56 (11.7)	63 (14.3)
3	27 (5.6)	29 (6.6)
4	17 (3.6)	11 (2.5)
5	3 (0.6)	5 (1.1)
6	4 (0.8)	2 (0.5)
8	1 (0.2)	1 (0.2)
11	1 (0.2)	
15	1 (0.2)	
Total	579	613

APPENDIX 2

R Code – Models With Offset

```
rm(list=ls())
library(MASS)
library(COUNT)
library(sandwich)
library(pscl)
urso<-read.csv("Path to CSV file")

urso$lnFUtime<-log(urso$FUtime)
urso$FUtimeyrs<-urso$FUtime/365.25
urso$lnFUtimeyrs<-log(urso$FUtimeyrs)
urso$lnBMI<-log(urso$bmi)
urso$polyphx[urso$histpol=="Y"]<-1
urso$polyphx[urso$histpol=="N"]<-0
summary(urso$FUtimeyrs)
hist(urso$FUtimeyrs)
ursoal3<-urso
ursoal3$yrs23[ursoal3$FUtimeyrs<=3.5]<-1
ursoal3$yrs23[ursoal3$FUtimeyrs>3.5]<-0
ursoall3<-subset(ursoal3, yrs23==1)


#---------------------------Poisson GLM---------------------------#
Mo5<-glm(tnumaden~factor(grp) + factor(famcolon) + factor(sex) + age +
lnBMI + offset(lnFUtimeyrs), family="poisson", data=ursoall3)
summary(Mo5)

#Sandwich stimates
robustSEsMo5<-sqrt(diag(vcovHC(Mo5, type="HC0")))
robustSEsMo5
confint(Mo5)      #Confidence Intervals

Mo5rdf<-Mo5$df.residual       #Extract Residual Degrees of Freedom
Mo5rdf
prMo5<-sum(residuals(Mo5, type="pearson")^2)    #Extract Pearson Resids
prMo5
dispMo5<-prMo5/Mo5$df.residual      #Compute Dispersion Statistic
dispMo5
Mo5N<-length(Mo5$y)     #Number of obs in model
Mo5N
```

```
Mo5k<-dim(vcov(Mo5))[1]       #Number of predictors k in model
Mo5k
LLMo5<-(Mo5$aic-2*Mo5k)/-2    #Compute Log-Likelihood
LLMo5

mfitMo5<-modelfit(Mo5) #Extract Information Criteria (COUNT Pkg)
mfitMo5

Mo5count<-predict(Mo5) #same as type = "link" - probs on log scale
Mo5count[1:20]


####P(Yi|Xi) same as fitted.values y-hat
Mo5response<-predict(Mo5, type="response") #P(Yi|Xi) same as y-hat

###No5response is the predicted value of lambdai!###
Mo5response[1:10]
Mo5$fitted.values[1:10]

Mo5predprob<-predprob(Mo5) #Predicted probability of each count
Mo5predprob[1:10,]

###Dean and Lawless Score Test
Mo5Zi<-((ursoall3$tnumaden - Mo5response)^2 -
  ursoall3$tnumaden)/(sqrt(2)*Mo5response)
Mo5Z<-lm(Mo5Zi ~ 1)
summary(Mo5Z)

#-----------------------------------------------------------------#
###Van Den Broek Score Test
Mo5mui<-as.vector(predict(Mo5, type="response"))
Mo5Iyzero<-as.vector(ifelse(Mo5$y==0,1,0))
Mo5aux2<-exp(-Mo5mui)
Mo5intcpt<-rep(1, length(Mo5$y))
Mo5N<-length(Mo5$y)
Mo5X<-matrix(c(Mo5intcpt, as.vector(ursoall3$grp),
  as.vector(ursoall3$famcolon), as.vector(ursoall3$sex),
  as.vector(ursoall3$age), as.vector(ursoall3$lnBMI)), nrow=Mo5N,
ncol=6)
Mo5aux3<-t(Mo5mui) %*% Mo5X %*% solve(t(Mo5X) %*% diag(Mo5mui) %*%
Mo5X) %*% t(Mo5X) %*% Mo5mui
```

```
Mo5VDB<-(sum(Mo5Iyzero/Mo5aux2)-Mo5N)^2 / ((sum(1/Mo5aux2) - Mo5N) -
Mo5aux3)
Mo5VDB

#--------------------Observed vs Predicted Poisson-----------------#
pois1<-poi.obs.pred(len=6, model=Mo5)
pois1
plot(0:6, pois1$propPred, type="b", xlim=c(0,6), pch=2, col=2,
ylim=c(0,70),
     lwd=2, main="Observed vs Predicted Adenoma Count",
     xlab="Adenoma Count", ylab="Probability of Adenoma")
lines(pois1$propObsv~pois1$Count, type="b", pch=1, col=1, lwd=2)
legend("topright", legend=c("Predicted", "Observed"),
      lty=c(1,1), pch=c(2,1), col=c(2,1), lwd=c(2,2))


#---------------------------NB2-------------------------#
Mo6<-glm.nb(tnumaden~factor(grp) + factor(famcolon) + factor(sex) + age
 + lnBMI + offset(lnFUtimeyrs), data=ursoall3)
summary(Mo6)
confint(Mo6)
names(Mo6)
names(summary(Mo6))
Mo6alpha<-1/Mo6$theta
  Mo6alpha
Mo6$SE.theta
1/Mo6$SE.theta
LLMo6<-logLik(Mo6)
LLMo6

prMo6<-sum(residuals(Mo6, type="pearson")^2)
  prMo6
dispMo6<-prMo6/Mo6$df.residual
  dispMo6
Mo6$df.residual
Mo6$deviance
(1/Mo6$df.residual)*Mo6$deviance
BICrMo6<-Mo6$deviance-Mo6$df.residual*log(Mo6obs)
  BICrMo6
BIClMo6<--2*LLMo6 + 6*log(length(Mo6$y)) #Use this one!
  BIClMo6
mfitMo6<-modelfit(Mo6)
names(mfitMo6)
```

```
mfitMo6
#LRtest H0: alpha=0
odTest(Mo6)


#----------------------Observed vs Predicted NB2----------------#
avgnb<-nb2.obs.pred(len=6, model=Mo6)
avgnb
plot(0:6, avgnb$propPred, type="b", xlim=c(0,6), ylim=c(0,70), lwd=2,
     pch=2, col=2, main="Observed vs Predicted Adenoma Count NB2",
     xlab="Adenoma Count", ylab="Probability of Adenoma")
lines(avgnb$propObsv~avgnb$Count, type="b", pch=1, col=1, lwd=2)
legend("topright", legend=c("Predicted", "Observed"),
      lty=c(1,1), pch=c(2,1), col=c(2,1), lwd=c(2,2))
#---------------------------------------------------------------#


#Zero-Inflated Models
#####################################################################
Mo7<-zeroinfl(tnumaden~factor(grp) + factor(famcolon) + age +
factor(sex) +
lnBMI + offset(lnFUtimeyrs), data=ursoall3, dist="poisson")
summary(Mo7)
confint(Mo7)

Mo7aic<- -2*Mo7$loglik + 2*12 #k=number of predictors in both
components!
Mo7aic
Mo7aaic<- -2*Mo7a$loglik + 2*8
Mo7aaic

Mo7$loglik
Mo7$df.residual

BIClMo7<--2*Mo7$loglik + 12*log(length(Mo7$y)) #Use this one - Deviance
not needed!
  BIClMo7

prMo7<-sum(residuals(Mo7, type="pearson")^2)
  prMo7
dispMo7<-prMo7/Mo7$df.residual
  dispMo7
```

```
vuong(Mo7, Mo5) #Testing superiority of ZIP over Poisson GLM
vuong(Mo7, Mo6) #Testing superiority of ZIP over Neg Bin GLM


############    Prediction Vectors and Matrices   ############
Mo7zero<-predict(Mo7, type="zero")   #predicted prob for zero comp
Mo7count<-predict(Mo7, type="count") #predicted mean for count comp
Mo7prob<-predict(Mo7, type="prob")   #predicted (prob(Yi|Xi)) values
#############################################################

############ Isolating the vector P(Yi|Xi) and matching to x vals##
dim(Mo7prob)
is.matrix(Mo7prob)
whichcolMo7<-match(Mo7$y, min(Mo7$y):max(Mo7$y))
whichcolMo7[1:10]

Mo7predi<-numeric(length(Mo7$y))
Mo7rowi<-c(1:length(Mo7$y))
Mo7rowi[1:4]

for(k in 1:length(Mo7$y)){
  Mo7predi[k]<-Mo7prob[Mo7rowi[k], whichcolMo7[k]]
}
Mo7predi[1:5]
length(Mo7predi)

Mo7predictzero<-Mo7predi[Mo7$y==0]
Mo7predictone<-Mo7predi[Mo7$y==1]
Mo7predicttwo<-Mo7predi[Mo7$y==2]
Mo7predictthree<-Mo7predi[Mo7$y==3]
Mo7predictfour<-Mo7predi[Mo7$y==4]
Mo7predictfive<-Mo7predi[Mo7$y==5]
Mo7predictsix<-Mo7predi[Mo7$y==6]
Mo7predprops<-c(mean(Mo7predictzero), mean(Mo7predictone),
mean(Mo7predicttwo),
  mean(Mo7predictthree), mean(Mo7predictfour), mean(Mo7predictfive),
   mean(Mo7predictsix))
Mo7predprops<-Mo7predprops*100
```

```
#########################################################################
#--------------------Observed vs Predicted ZIP---------------------#
plot(0:6, Mo7predprops, type="b", xlim=c(0,6), pch=2, col=2,
ylim=c(0,70),
     lwd=2, main="Observed vs Predicted Adenoma Count (ZIP)",
     xlab="Adenoma Count", ylab="Probability of Adenoma")
lines(pois1$propObsv~pois1$Count, type="b", pch=1, col=1, lwd=2)
legend("topright", legend=c("Predicted", "Observed"),
      lty=c(1,1), pch=c(2,1), col=c(2,1), lwd=c(2,2))
#-------------------------------------------------------------#

#########################################################
Mo8<-zeroinfl(tnumaden~ factor(grp) + factor(famcolon) + factor(sex)
  + age + lnBMI + offset(lnFUtimeyrs), data=ursoall3, dist="negbin")
summary(Mo8)
confint(Mo8)
Mo8$loglik
Mo8$df.residual
Mo8$df.null

Mo8aic<--2*Mo8$loglik + 2*12 #k=number of predictors in both
components!
Mo8aic

BIClMo8<--2*Mo8$loglik + 12*log(length(Mo8$y)) #Use this one - Deviance
not needed!
  BIClMo8

vuong(Mo8, Mo6) #Testing superiority of ZINB over NB2
Mo8$SE.logtheta

prMo8<-sum(residuals(Mo8, type="pearson")^2)
  prMo8
dispMo8<-prMo8/Mo8$df.residual
  dispMo8

############    Prediction Vectors and Matrices   ############
Mo8zero<-predict(Mo8, type="zero")
Mo8count<-predict(Mo8, type="count")
Mo8prob<-predict(Mo8, type="prob")
#############################################################
```

```
############    Isolating the vector P(Yi|Xi)   ##############
dim(Mo8prob)
is.matrix(Mo8prob)
whichcolMo8<-match(Mo8$y, min(Mo8$y):max(Mo8$y))
whichcolMo8[1:10]

Mo8predi<-numeric(length(Mo8$y))
Mo8rowi<-c(1:length(Mo8$y))
Mo8rowi[1:4]

for(k in 1:length(Mo8$y)){
  Mo8predi[k]<-Mo8prob[Mo8rowi[k], whichcolMo8[k]]
}
Mo8predictzero<-Mo8predi[Mo8$y==0]
Mo8predictone<-Mo8predi[Mo8$y==1]
Mo8predicttwo<-Mo8predi[Mo8$y==2]
Mo8predictthree<-Mo8predi[Mo8$y==3]
Mo8predictfour<-Mo8predi[Mo8$y==4]
Mo8predictfive<-Mo8predi[Mo8$y==5]
Mo8predictsix<-Mo8predi[Mo8$y==6]
Mo8predprops<-c(mean(Mo8predictzero), mean(Mo8predictone),
mean(Mo8predicttwo),
  mean(Mo8predictthree), mean(Mo8predictfour), mean(Mo8predictfive),
   mean(Mo8predictsix))
Mo8predprops<-Mo8predprops*100
Mo8predprops
#########################################################################
#--------------------Observed vs Predicted ZINB-----------------------
----#
plot(0:6, Mo8predprops, type="b", xlim=c(0,6), pch=2, col=2,
ylim=c(0,70),
     lwd=2, main="Observed vs Predicted Adenoma Count (ZINB)",
     xlab="Adenoma Count", ylab="Probability of Adenoma")
lines(pois1$propObsv~pois1$Count, type="b", pch=1, col=1, lwd=2)
legend("topright", legend=c("Predicted", "Observed"),
      lty=c(1,1), pch=c(2,1), col=c(2,1), lwd=c(2,2))
#-----------------------------------------------------------------#

#-------------LRT ZIP nested in ZINB HO:alpha=0-----------------#
LRTMo78<- -2*(Mo7$loglik - Mo8$loglik)
LRTMo78
dchisq(1, LRTMo78)/2
```

REFERENCES

1. Schatzkin A, Lanza E, Freedman LS, et al. The polyp prevention trial I: rationale, design, recruitment, and baseline participant characteristics. *Cancer Epidemiol Biomarkers Prev* 1996;5:375-83.

2. Baron JA, Cole B, Sandler RS, et al. A randomized trial of aspirin to prevent colorectal adenomas. *New Engl J Med* 2003;348:891-899.

3. Alberts DS, Martinez ME, Hess LM et al. Phase III trial of ursodeoxycholic acid to prevent colorectal adenoma recurrence. *J Natl Cancer Inst* 2005;97(11):846-853.

4. Hsu CH. A weighted zero-inflated Poisson model for estimation of recurrence of adenomas. *Stat Methods Med Res* 2007;16:155-166.

5. Hsu CH, Taylor JMG, Long Q, et al. Analysis of colorectal adenoma recurrence data subject to informative censoring. *Cancer Epidemiol Biomarkers Prev* 2009;18:712-717.

6. Hilbe J. *Negative Binomial Regression.* 2011. New York: Cambridge University Press.

7. Breslow NE. Extra-Poisson variation in log-linear models. *Appl Statist* 1984;33(1):38-44.

8. Campbell MJ, Machin D, D'Arcangues C. Coping with extra Poisson variability in the analysis of factors influencing vaginal ring expulsions. *Stat Med* 1991;10:241-254.

9. Bohning D. A note on a test for Poisson overdispersion. *Biometrika* 1994;81(2):418-419.

10. Hinde J and Demetrio CGB. Overdispersion: models and estimation. *Comput Stat Data Anal* 1998;27:151-170.

11. Poortema K. Modeling overdispersion of counts. *Stat Neerl* 1999;53:5-20.

12. Cameron AC and Trivedi PK. *Regression analysis of count data.* 1998. New York:Cambridge University Press.

13. Johnson NL, Kemp AW, Kotz S. *Univariate Discrete Distributions*. 2005. New Jersey: Wiley and Sons.

14. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proc fifth Berkeley Symposium on Mathematical Statistics and Probability* 1967;1:221-233.

15. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980;48:817-838.

16. Lambert D. Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics* 1992;34:1-14.

17. Greene W. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *Working paper* 1994. Department of Economics, Stern School of Business, New York University.

18. Dean C and Lawless JF. Tests for detecting overdispersion in Poisson regression models. *J Am stat Assoc.* 1989;84(406):467-72

19. Van Den Broek J. A score test for zero inflation. *Biometrics*. 1995;51(2):738-43.

20. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 1989;57:307-333.

21. Clarke KA. Testing nonnested models of international relations: revaluating realism. *Am Journal of Political Science* 2001;45(3):724-44.

22. Akaike H. A new look at the statistical model identification. *IEEE Transaction on Automatic Control* 1974;19(6):716-723.

23. Martinez ME, Jacobs ET, Ashbeck EL, et al. Meat intake, preparation methods, mutagens and colorectal adenoma recurrence. *Carcinogenesis* 2007;28:2019-2027.

24. Wertheim BC, Martinez ME, Ashbeck EL, et al. Physical activity as a determinant of fecal bile acid levels. *Cancer Epidemiol Biomarkers Prev* 2009;18:1591-1598.

25. Wertheim BC, Smith JW, Fang C, et al. Risk modification of colorectal adenoma by CYP7A1 polymorphisms and the role of bile acid metabolism in carcinogenesis. *Cancer Prev Res(Phila)* 2012;5:197-204.

26. Zeileis A, Kleiber C, and Jackman S. Regression models for count data in R. *J Statistical Software* 2008;27(8).

27. Agresti A. *Categorical Data Analysis*. 2002. New York: Wiley-Interscience.

28. Cassella and Berger. *Statistical Inference*. 2001. Pacific Grove: Duxbury Press.

29. Mullahy, J. 1986. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*. V33: 341–365.

30. Rex DK, Cutler CS, Lemel T, et al. Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. *Gastroenterology* 1997;112:24-28.

31. Thompson PA, Wertheim BC, Roe DJ, et al. Gender modifies the effect of ursodeoxycholic acid in a randomized controlled trial in colorectal adenoma patients. *Cancer Prev Res (Phila)* 2009;2:1023-30.